Non-Asymptotic Analysis for Relational Learning with One Network

Peng He
Department of Automation
Tsinghua University

Changshui Zhang
Department of Automation
Tsinghua University

Abstract

This theoretical paper is concerned with a rigorous non-asymptotic analysis of relational learning applied to a single network. Under suitable and intuitive conditions on features and clique dependencies over the network, we present the first probably approximately correct (PAC) bound for maximum likelihood estimation (MLE). To our best knowledge, this is the first sample complexity result of this problem. We propose a novel combinational approach to analyze complex dependencies of relational data, which is crucial to our non-asymptotic analysis. The consistency of MLE under our conditions is also proved as the consequence of our sample complexity bound. Finally, our combinational method for analyzing dependent data can be easily generalized to treat other generalized maximum likelihood estimators for relational learning.

1 Introduction

In recent years, there has been an explosion of interest in statistical relational learning (SRL), with successful applications including social networks, link analysis, citation analysis and web mining(Neville and Jensen, 2007)(Sutton and Mccalium, 2006)(Xiang and Neville 2011). In many real world applications, relational data are drawn from a single network (e.g., Facebook). In this scenario, the graph size grows with increasing number of samples, and tasks of learning and predicting are performed in a single network. In the community of statistical relational learning, the dependencies among instances can be often modeled by Markov networks such as relational Markov networks (Taskar et al., 2002) and Markov logic networks (Richardson and Domingos, 2006). Other undirected probabilistic graphical models used in SRL include relational dependency networks (Neville and Jensen, 2007), exponential random graph (p^*) models (Robins et al., 2007), and CRFs (Sutton and Mccalium, 2006).

Although there have been a number of major technical developments, few theoretical issues of relational learning in a single network have been addressed. To our knowledge, (Xiang and Neville 2011) is the only paper addressing the asymptotic analysis of this problem. However, the weak dependence assumption proposed in (Xiang and Neville 2011) for studying the asymptotic behavior of the estimators is hard to check. Moreover, the sample complexity of relational learning in single-network domains has not been explored. In additional, asymptotic properties have been also investigated on linear-chain conditional random fields (Sinn and Chen, 2013)(Sinn and Poupart, 2011).

To our knowledge, there are some recent on-asymptotic results of learning probabilistic graphical models which often assume independent and identically distributed (i.i.d.) samples. For instance, the PAC bound has been obtained for learning parameters of high-treewidth discrete models (Abbeel et al., 2006). For learning parameters of general Markov Random Fields (MRFs) (c.f. Koller and Friedman, 2009) and Conditional Random Fields (CRFs) (Lafferty et al., 2001), the sample complexity of MLE, maximum composite likelihood estimation (MCLE) (Lindsay, 1988) were analyzed in (Bradley and Guestrin, 2012). More recently, the model selection consistency of M-estimators with geometrically decomposable penalties was investigated in (Lee et al., 2013). Nevertheless, the nature of relational learning for single network domains involves a single graph with increasing size and complex dependencies across data over the graph. Therefore, methods for obtaining sample complexity results from i.i.d. instances cannot be directly applied in our problem.

In this paper, we perform non-asymptotic analysis in

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

the framework of Markov networks for relational data due to their widely applications in relational learning and desirable analytic properties. We propose suitable assumptions on features and an intuitive dependence assumption, i.e., finite distance dependence. Under these conditions, we present the first strong sample guarantees for relational learning via MLE within the probably approximately correct (PAC) learning framework (Valiant 1984). To the best of our knowledge, this is the first PAC bound in relational learning with one network. Compared with recent sample complexity bounds of learning MRFs and CRFs (Bradley and Guestrin, 2012) that assume independence and identically distributed (i.i.d.) samples, our non-asymptotic analysis throws light on the intrinsic difficulty of learning with dependent data. Furthermore, we propose a combinational method to characterize complex dependencies of data instances. Our findings from combinational analysis of dependent cliques illustrate that it is the combinational structure related with dependent cliques has a high impact on samples required for MLE. We end with one consequence of our non-asymptotic analysis, i.e., maximum likelihood estimators will be consistent under our conditions.

2 Problem Formulation

2.1 Basic Setup

We consider relational learning on Markov networks. Let $G = (V, E)$ be a factor graph over sets of random variables $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the set of input observed attributes and $\mathcal{Y}$ is the set of output labels to be predicted. Given assignments to $\mathcal{X} \times \mathcal{Y}$ by (x, y), then the probability density has the form

$$p(y|x) = \frac{1}{Z(x)} \prod_{T \in \mathcal{T}} \prod_{c \in C_T} \Psi(x_c, y_c; \theta_T),$$

where $\mathcal{T}$ is the set of clique templates, C_T is the set of cliques corresponding to the template T and $Z(x)$ denotes the partition function. Each potential function Ψ is defined on a clique c which consists of a small subgraph of G. x_c and y_c are assignments of x and y over the clique c, while the clique parameters θ_T are tied for the same template. In this paper, we consider only one template to keep the notation simple, thus dropping the subscript T. We also assume each potential function has the log-linear form

$$\Psi(x_c, y_c; \theta) = \exp\{< \theta, \phi(x_c, y_c) >\}$$

for some real-valued parameter vector θ, and for a feature vector or sufficient statistics ϕ of each clique c. Let C_G be the set of cliques in the graph G, $n = |C_G|$ and $m = |\{v \in V | v \in c\}|$, then the Markov network model can be written as

$$p_\theta(y|x) = \frac{1}{Z_n(\theta; x)} \exp\left\{< \theta, \sum_{c \in C_G} \phi(x_c, y_c) >\right\},$$

where $Z_n(\theta; x)$ is the partition function

$$Z_n(\theta; x) = \int_{\mathcal{Y}} \exp\left\{< \theta, \sum_{c \in C_G} \phi(x_c, y_c) >\right\} dy.$$

Let $\theta^* \in \mathbb{R}^r$ be unknown true parameter, MLE seeks θ^* by maximizing the normalized log-likelihood function:

$$\ell^{(n)}(\theta; x) = \frac{1}{n} \sum_{c \in C_G} < \theta, \phi(x_c, y_c) > - \frac{1}{n} \log Z_n(\theta; x).$$

We denote $\ell^{(n)}(\theta)$ as a shorthand for $\ell^{(n)}(\theta; x)$ and $\hat{\theta}_n = \text{argmax}_\theta \ell^{(n)}(\theta)$ as the estimated parameter. Standard results on exponential families (M. J. Wainwright and M. I. Jordan, 2008) give

$$\nabla \ell^{(n)}(\theta) = \frac{1}{n} \sum_{c \in C_G} \phi(x_c, y_c) - \mathbb{E}_\theta[\phi|x], \tag{1}$$

$$\nabla^2 \ell^{(n)}(\theta) = -\text{var}_\theta[\phi|x]. \tag{2}$$

Table 1: Notations used in this paper

$G = (V, E)$: Markov network
$v, v' \in V$: node (vertex)
$c, c' \in C_G$: template potential cliques
n: number of cliques ($\|C_G\|$)
m: number of nodes in one clique
d: maximal degree of nodes
λ: distances between nodes and cliques
x_c, y_c: instantiation of $x(y)$ over the clique c
θ^*: the true parameter to be estimated
$\hat{\theta}_n$: the maximum likelihood estimate
r: the dimension of parameters
$\phi(x_c, y_c)$: features over the clique c
ϕ_{max}: the maximal magnitude of feature elements
$\ell^{(n)}(\theta)$: the normalized log-likelihood function
$\mathbb{E}_\theta$: expectation taken with respect to p_θ

2.2 Definitions and Assumptions

Before our formal analysis, we outline the conditions used in this paper. The MLE asymptotic and non-asymptotic behaviors heavily rely on the dependence of features over vertices and cliques. And it is reasonable that the dependence of features is highly related with distances of nodes. Hence, we first define distances of vertices and cliques, then propose an assumption of feature dependence, which we call the finite distance dependence.

Definition 2.1 (Distances of Nodes and Cliques)**.** *Let $\psi(v, v')$ be the minimal length of the path connecting two nodes v and v' in the graph, then we define the distance of two cliques as $\psi(c, c') = max\{\psi(v, v')|v \in c, v' \in c'\}$.*

Assumption 2.1 (Bounded Features)**.** *The magnitude of any feature vector element is upper bounded: $\phi_{max} = \max_{j,x_c,y_c} |\phi_j(x_c, y_c)|$. In addition, we further assume node degrees remain bounded as n grows, i.e., $deg(v) \leqslant d < \infty$ for any v and n.*

Assumption 2.2 (Finite Distance Dependence)**.** *There exists a constant $\lambda^* > 0$ such that the correlation coefficient $\rho_j(c, c') = 0$ for any feature vector element ϕ_j between cliques c and c', provided that $\psi(c, c') > \lambda^*$.*

Assumption 2.3 (Minimal Curvature)**.** *Assume the minimum eigenvalue of feature covariance matrix $\Lambda_{min}(var_{\theta^*}[\phi|x]) \geqslant C_{min} > 0$.*

Our assumption of finite distance dependence captures the intuition that cliques are independent if they are far enough. The assumption of finite distance dependence is similar to the irrepresentable condition for model selection consistency of lasso (Wainwright 2009)(Zhao and Yu, 2006) and the incoherence condition (Ravikumar et al., 2010), which require nonneighbors of a node are not overly dependent on neighbors of the node. Notice that graphs satisfying the Assumption 2.2 do not require that correlations of features over nodes decay as distance increase. No matter how strong the dependence of data have, they can still satisfy the condition of finite space dependence as long as dependence data lies in a bounded space.

In the paper (Xiang and Neville 2011), the asymptotic analysis of relational data with one network is performed under the weak dependence condition, where the total covariance of various cliques in the network is finite. It is easy to see that features satisfying finite distance dependence assumption will also make weak dependence assumption in (Xiang and Neville 2011) hold. However, our finite distance dependence assumption is easier checked in some real-world applications, where the constant λ^* can be estimated from theory and practice. Moreover, one can obtain the sample complexity of his learning problem using our result if the finite distance dependence holds, which is often very useful for applications. In addition, it is hard to get the sample complexity using weak dependence assumption in (Xiang and Neville 2011).

Assumption 2.3 ensures that feature components do not become overly dependent. From the geometrical perspective, the curvature of normalized log-likelihood is supposed to be bounded away from zero from Assumption 2.3 and Eq. (2). Therefore, the function $\ell^{(n)}(\theta)$ is curved along certain directions around the true parameter θ^* in the parameter space if Assumption 2.3 holds. This guarantees the uniqueness of MLE, and it will be not possible to obtain a consistent MLE if Assumption 2.3 would not hold. The role of 2.3 is rather similar to the notion of compatibility (P. Bühlmann and Geer, 2011) and restricted strong convexity (Lee et al., 2013)(Negahban et al., 2012).

2.3 The Road Map

The intuition behind performing the non-asymptotic analysis of relational data is simple. The Hessian of the normalized log-likelihood function can be also controlled by Assumption 2.3. Suppose we have an estimate of the convergence rate of gradients of $\hat{\theta}_n$, we can obtain the sample complexity of MLE by Taylor's expansion around the true parameter θ^*. The asymptotic property such as weak consistency of estimated parameters is the offspring of the non-asymptotic analysis.

Hence, the main barrier of carrying out the above plan is estimating the gradients' convergence rates for dependent data. Our idea of overcoming this difficulty is to divide the network and tackle with them separately. Explicitly, the finite distance dependence of cliques makes it possible to partition the whole network into two groups for a given clique: the group of independent cliques and the group of other close-by cliques. Thus our problem is finally reduced to estimate the number of close-by cliques for a given clique, which can be solved by careful combinational analysis of *interactions* nodes and cliques over the network.

3 Combinational Analysis of Dependent Cliques

In this section, we analyze dependent cliques by combinational approaches. The crucial quality of our development of dependence analysis of cliques is the upper bound of the number of cliques within a certain distance λ of one given clique c. More explicitly, our destination is to estimate the upper bound of $C(c, \lambda) = \{c' \in C_G|1 \leqslant \psi(c, c') \leqslant \lambda\}$. The estimate reveals the basic relationship of parameters in the data-dependent graphical model. Therefore, the result we obtained is not only useful for our analysis of the PAC bound of MLE but also for studying other problems concerning dependent data analysis.

Before proceeding further, let us define some notations to obtain the estimate of $|C(c, \lambda)|$. Consider one vertex $v \in c$, let $V(c, v, \lambda) = \{v' \in V|\psi(v, v') \leqslant \lambda, v' \notin c\}$ and $V(c, \lambda) = \cap_{v \in c} V(c, v, \lambda)$. From the definition of

$\psi(c, c')$, it is easy to see that

$$V(c, \lambda) = \{v' \in V | v' \in c', v \notin c, c' \neq c, \psi(c, c') \leqslant \lambda\}.$$

In other words, $V(c, \lambda)$ is the set of vertices in those cliques, which distances to the given clique c are not larger than λ, excluding vertices in the clique c itself.

We approach the problem of estimating the upper bound of $C(c, \lambda) = \{c' \in C_G | 1 \leqslant \psi(c, c') \leqslant \lambda\}$ into three steps. First, we establish the upper bound of $|V(c, \lambda)|$. Then we obtain an initial upper bound of $|C(c, \lambda)|$ from some simple observations. In the third step, a final tighter upper bound is obtained by careful combinatorial analysis based on the initial result gotten in the second step. Although it is very easy to prove results in the first two steps, we still list them as the following two lemmas separately. The reason is that they are frequently used in proving the tighter bound in Theorem 3.1, which is the main result in this section.

Lemma 3.1.

$$|V(c, \lambda)| < 3d^{\lambda}.$$

Proof

For each $i \in \{1, \dots \lambda\}$, consider vertices $v' \in V(c, v, \lambda)$ with $\psi(v, v') = i$, then it is evident that the number of these vertices is upper bounded by $d(d-1)^{i-1}$. Consequently, we infer that

$$|V(c, \lambda)| \leqslant |V(c, v, \lambda)| \leqslant \sum_{i=1}^{\lambda} d(d-1)^{i-1} < 3d^{\lambda}.$$

□

Lemma 3.2.

$$|C(c, \lambda)| < 3\binom{d}{m-1} d^{\lambda}.$$

Proof

For any vertex $v' \in V(c, v, \lambda)$, it is easy to see that the maximal number of cliques covering v' is $\binom{d}{m-1}$. This fact directly leads to an obvious method to obtain one upper bound of $|C(c, \lambda)|$, i.e., counting $\binom{d}{m-1}$ cliques for each vertex in $V(c, v, \lambda)$. Hence, we arrive at the conclusion that

$$|C(c, \lambda)| \leqslant |V(c, \lambda)| \binom{d}{m-1} < 3\binom{d}{m-1} d^{\lambda}.$$

□

However, this upper bound is too loose because many cliques in $C(c, \lambda)$ are counted for several times and so we call this counting method "NAIVE COUNTING". The approach of "NAIVE COUNTING" can be called the *vertex-oriented* counting strategy. In the *vertex-oriented* method, we first estimate the upper bound of numbers of vertices in $V(c, v, \lambda)$, then estimate the upper bound of numbers of cliques covering each vertex. Although the "NAIVE COUNTING" method is simple and even seems "naive", it is the foundation of our improved counting method which we call the *clique-oriented* method. From the *clique-oriented* perspective, we investigate various cases for one clique covering vertices in $V(c, v, \lambda)$. By estimating how many times one clique in $C(c, \lambda)$ are counted in "NAIVE COUNTING", we can obtain a tighter upper bound of $C(c, \lambda)$.

The next theorem (3.1) is a consequence of carrying out this "BETTER COUNTING" plan. The key idea of our proof is to divide $V(c, v, \lambda)$ into special disjoint subsets first, then investigate situations that cliques in $V(c, \lambda)$ cover on these subsets carefully. Notice that the upper bound of $|C(c, \lambda)|$ in Theorem (3.1) is independent of the clique c, which is crucial for analyzing convergence rates of MLE in the next section.

Theorem 3.1 (Dependent Cliques Bound)**.** *For any clique $c \in C_G$ and positive integer $\lambda > m$, d is the maximal degree of nodes in V and m is the number of vertices in a clique, let $C(c, \lambda) = \{c' \in C_G | 1 \leqslant \psi(c, c') \leqslant \lambda\}$, we have*

$$|C(c, \lambda)| < \frac{3}{2}\left(\frac{1}{m} + 1\right)\binom{d}{m-1} d^{\lambda}.$$

Proof

For one vertex $v \in c$, we divide vertices in $V(c, v, \lambda)$ into three disjoint subsets: V_I, V_{II} and V_{III} as follows:

$$\begin{aligned} V_I &= \{v' \in V(c, v, \lambda) | \psi(v, v') = 1\}, \\ V_{II} &= \{v' \in V(c, v, \lambda) | 2 \leqslant \psi(v, v') \leqslant \lambda - 1\}, \\ V_{III} &= \{v' \in V(c, v, \lambda) | \psi(v, v') = \lambda\}. \end{aligned}$$

Then for a clique $c' \in C(c, \lambda)$ and $m \geqslant 2$, there are at most six cases that c covers nodes in the network, as listed as follows:

$$\begin{aligned} C_1 &= \{c' \in C(c, \lambda) | c' \text{ only covers nodes in } V_I\}, \\ C_2 &= \{c' \in C(c, \lambda) | c' \text{ only covers nodes in } V_{II}\}, \\ C_3 &= \{c' \in C(c, \lambda) | c' \text{ only covers nodes in } V_{III}\}, \\ C_4 &= \{c' \in C(c, \lambda) | c' \text{ covers } v \text{ and nodes in } V_I\}, \\ C_5 &= \{c' \in C(c, \lambda) | c' \text{ covers both nodes in } V_I \text{ and } V_{II}\}, \\ C_6 &= \{c' \in C(c, \lambda) | c' \text{ covers both nodes in } V_{II} \text{ and } V_{III}\}. \end{aligned}$$

The sequence in naming $C_1, \dots, C_6$ follows sets of V_I to V_{III}. Due to $\lambda > m$, we can rule out the seventh possibility that c covers nodes in all of three subsets of V_I, V_{II} and V_{III}. Notice that each clique in $C(c, \lambda)$

belongs to one and only one set from C_1 to C_6, which leads to the fact that $|C(c,\lambda)| = \sum_{i=1}^{6} |C_i|$.

We now turn to analyze how many times for a clique $C(c,\lambda)$ is counted in "NAIVE COUNTING". To make the reasoning more clearly, it is helpful to adopt the following order of analyzing cliques.

To begin with, for each vertex $v' \in V_{II}$, due to each clique is a fully connected subgraph of the whole network (M. J. Wainwright and M. I. Jordan, 2008). Notice that there is no edge connecting v and any node in V_{II} from the definition of V_{II}. Similarly, there is no edge connecting any node in V_{II} and node which is not in $\{v\} \cup V(c,v,\lambda)$. Hence we see that each clique in C_2 is counted exactly for m times in "NAIVE COUNTING".

Next, we consider the cases of C_5 and C_6. Let $V_{I'}$ be the subset of V_I such that each node in $V_{I'}$ is covered by certain clique in C_5. And $V_{III'}$ is defined in a similar manner, i.e., the subset of V_{III} such that each node in $V_{III'}$ is covered by certain clique in C_5. Then it can be seen that each clique in $C_5 \cup C_6$ is counted exactly for m times in "NAIVE COUNTING". Thus, it is not difficult to conclude that $|C_2| + |C_5| + |C_6| \leqslant \frac{1}{m}(|V_{I'}| + |V_{II}| + |V_{III'}|)|\binom{d}{m-1}$.

We proceed to study cases of C_1 and C_4. It is observed that the number of nodes covered by C_1 and C_4 equals $|V_I| - |V_{I'}|$. For the clique C_1, we use the obvious estimate of $|C_1| \leqslant (|V_I| - |V_{I'}|)|\binom{d}{m-1}$. It is not difficult to verify that $|C_4| \leqslant \frac{1}{m-1}(|V_I| - |V_{I'}|)|\binom{d}{m-1}$. Therefore, we infer that $|C_1| + |C_4| \leqslant (|V_I| - |V_{I'}|)\binom{d}{m-1}$.

Finally, the number of nodes covered by C_3 is $|V_{III}| - |V_{III'}|$. On account of the fact that the clique in C_3 may be counted for 1 to m times in various networks, we use the obvious estimate: $|C_3| \leqslant (|V_{III}| - |V_{III'}|)|\binom{d}{m-1}$.

We are ready to proceed with the final step of our proof. For $m \geqslant 2$, i.e., putting estimates for $C_i (i = 1, \dots 6)$ in one piece gives

$$
\begin{aligned}
&|C(c,\lambda)| \\
=&(|C_2| + |C_5| + |C_6|) + (|C_1| + |C_4|) + |C_3| \\
\leqslant&\left[\frac{1}{m}(|V_{I'}| + |V_{II}| + |V_{III'}|) + (|V_I| - |V_{I'}|)\right. \\
&\left. + (|V_{III}| - |V_{III'}|)\right]\binom{d}{m-1} \\
=&\left[\frac{1}{m}(|V_I| + |V_{II}| + |V_{III}|) + \left(1 - \frac{1}{m}\right)(|V_I| - |V_{I'}|)\right. \\
&\left. + \left(1 - \frac{1}{m}\right)(|V_{III}| - |V_{III'}|)\right]\binom{d}{m-1} \\
\leqslant&\left[\frac{1}{m}|V(c,v,\lambda)| + \left(1 - \frac{1}{m}\right)(|V_I| + |V_{III}|)\right]\binom{d}{m-1}.
\end{aligned}
$$

We apply the fact that $|V(c,v,\lambda)| < 3d^{\lambda}$ follows from the proof of Lemma 3.1 and recall that

$$|V_I| + |V_{III}| \leqslant d + d(d-1)^{\lambda-1} < \frac{3}{2}d^{\lambda}.$$

Therefore, we arrive at the the following estimate for $m \geqslant 2$

$$|C(c,\lambda)| < \frac{3}{2}\left(\frac{1}{m} + 1\right)\binom{d}{m-1}d^{\lambda}. \tag{3}$$

In the case of $m = 1$ (singleton cliques), it is apparent that $|C(c,\lambda)| \leqslant |V(c,v,\lambda)| < 3d^{\lambda}$ for any node v.

Hence, we complete the proof for any $m \geqslant 1$. □

We observe from Theorem 3.1 that λ^* plays a crucial role in determining the upper bound of $C(c,\lambda)$, demonstrating the high impact of the dependence of clique features. We believe that the number of dependent cliques within particular distances is of fundamental importance in analyzing performances of not only MLE but also other learning methods with a single network.

4 PAC Analysis of MLE

This section presents PAC analysis of the MLE estimator and gives the corresponding sample complexity. Our first result is the convergence rates of gradients of the normalized log-likelihood function, which is the base for PAC analysis.

4.1 Convergence Rates of Gradients

Lemma 4.1. *Suppose Assumptions 2.1, 2.2 and 2.3 hold. Given n training samples, then for any $t > 0$, we have*

$$
\begin{aligned}
&\mathbb{P}\left[\|\nabla \ell^{(n)}(\theta^*)\|_2 > t\right] \\
<&\frac{r^2\phi_{max}^2}{nt^2}\left[\frac{3}{2}\left(\frac{1}{m} + 1\right)\binom{d}{m-1}d^{\lambda^*} + 1\right].
\end{aligned}
$$

Proof

For any $t > 0$ and element j of $\nabla \ell^{(n)}(\theta^*)$, Chebyshev's

inequality leads to

$$\begin{aligned}
&\mathbb{P}\left[[\nabla \ell^{(n)}(\theta^*)]_j > t\right] \\
=&\mathbb{P}\left[\frac{1}{n}\sum_{c\in C_G}\phi_j(x_c,y_c) - \mathbb{E}_\theta[\phi_j|x] > t\right] \\
\leqslant&\frac{1}{n^2t^2}D\left[\sum_{c\in C_G}\phi_j(x_c,y_c)\right] \\
=&\frac{1}{n^2t^2}\Bigg[\sum_{c\in C_G}D[\phi_j(x_c,y_c)] \\
&+\sum_{c\neq c'}cov[\phi_j(x_c,y_c),\phi_j(x'_c,y'_c)]\Bigg]. \qquad (4)
\end{aligned}$$

First, we notice that $D[\phi_j(x_c,y_c)] \leqslant \mathbb{E}[\phi_j^2(x_c,y_c)] \leqslant \phi_{max}^2$. Dividing the sum of covariances in Eq. (4) into two parts

$$\sum_{c\neq c'}cov[\phi_j(x_c,y_c),\phi_j(x'_c,y'_c)] = \sum_{1\leqslant d(c,c')\leqslant\lambda^*} + \sum_{d(c,c')>\lambda^*}.$$

We can upper bound the first term by applying Theorem 3.1:

$$\begin{aligned}
&\sum_{1\leqslant d(c,c')\leqslant\lambda^*}cov[\phi_j(x_c,y_c),\phi_j(x'_c,y'_c)] \\
=&\sum_{c\in C_G}|C(c,\lambda)|\rho_j(c,c')\sqrt{D[\phi_j(x_c,y_c)]D[\phi_j(x'_c,y'_c)]} \\
<&\frac{3}{2}\left(\frac{1}{m}+1\right)\binom{d}{m-1}nd^{\lambda^*}\phi_{max}^2. \qquad (5)
\end{aligned}$$

Evidently,

$$\begin{aligned}
&\sum_{d(c,c')>\lambda^*}cov[\phi_j(x_c,y_c),\phi_j(x'_c,y'_c)] \\
=&\sum_{d(c,c')>\lambda^*}\rho_j(c,c')\sqrt{D[\phi_j(x_c,y_c)]D[\phi_j(x'_c,y'_c)]} = 0.
\end{aligned}$$

Substituting the Ineq. (5) into Eq. (4) and using $D[\phi_j(x_c,y_c)] \leqslant \phi_{max}^2$, we arrive at

$$\begin{aligned}
&\mathbb{P}\left[[\nabla \ell^{(n)}(\theta^*)]_j > t\right] \\
<&\frac{\phi_{max}^2}{nt^2}\left[\frac{3}{2}\left(\frac{1}{m}+1\right)\binom{d}{m-1}d^{\lambda^*}+1\right].
\end{aligned}$$

Applying a union bound over all r elements of $\nabla\ell^{(n)}(\theta^*)$ leads to

$$\begin{aligned}
&\mathbb{P}\left[\|\nabla \ell^{(n)}(\theta^*)\|_\infty > t\right] \\
<&\frac{r\phi_{max}^2}{nt^2}\left[\frac{3}{2}\left(\frac{1}{m}+1\right)\binom{d}{m-1}d^{\lambda^*}+1\right]. \qquad (6)
\end{aligned}$$

Continuing

$$\begin{aligned}
&\mathbb{P}\left[\|\nabla \ell^{(n)}(\theta^*)\|_2 > t\right] \\
\leqslant&\mathbb{P}\left[\|\nabla \ell^{(n)}(\theta^*)\|_\infty > \frac{t}{\sqrt{r}}\right] \\
<&\frac{r^2\phi_{max}^2}{nt^2}\left[\frac{3}{2}\left(\frac{1}{m}+1\right)\binom{d}{m-1}d^{\lambda^*}+1\right].
\end{aligned}$$

Hence we complete the proof. □

A notable characteristics of convergence rates of gradients is that the order of convergence rates is $\mathcal{O}(n^{-1})$ with n dependent samples. The polynomially convergence rates differ much than those in many learning exponential families with i.i.d. n samples, which often have the exponential convergence rates of n. More details of this phenomenon are discussed in Section 4.3.

4.2 Sample Complexity Bounds

We are now in a position to present our main theoretical results: the sample complexity as the PAC bound of relational learning with one network using MLE.

Theorem 4.1 (MLE Sample Complexity)**.** *Suppose Assumptions 2.1, 2.2 and 2.3 hold. For any $\varepsilon > 0$ and $0 < \delta < 1/2$, the MLE learns the parameter within L_2 error ε with probability at least $1-\delta$, provided that*

$$n > \frac{4r^2\phi_{max}^2}{\varepsilon^2\delta C_{min}^2}\left[\frac{3}{2}\left(\frac{1}{m}+1\right)\binom{d}{m-1}d^{\lambda^*}+1\right].$$

Proof

Taking Taylor's expansion of the normalized log-likelihood $\ell^{(n)}(\hat{\theta}_n)$ around θ^* gives

$$\begin{aligned}
&\ell^{(n)}(\hat{\theta}_n) \\
=&\ell^{(n)}(\theta^*) + [\nabla\ell^{(n)}(\theta^*)]^T(\hat{\theta}_n-\theta^*) \\
+&\frac{1}{2}(\hat{\theta}_n-\theta^*)^T\nabla^2\ell^{(n)}(\theta^*)(\hat{\theta}_n-\theta^*) \\
+&\frac{1}{6}\sum_{j=1}^{p}[(\hat{\theta}_n-\theta^*)]_j(\hat{\theta}_n-\theta^*)^T\left[\frac{\partial}{\partial\theta_j}\nabla^2\ell^n(\tilde{\theta}_n)\right](\hat{\theta}_n-\theta^*), \qquad (7)
\end{aligned}$$

where $\tilde{\theta}_n = \theta^* + \eta(\hat{\theta}_n-\theta^*)$ for certain $\eta\in(0,1)$.

Let $t = \hat{\theta}_n - \theta^*$, then straightforward calculations of the third-order term in Eq. (7) give rise to

$$\begin{aligned}
&\frac{1}{6}\sum_{j=1}^{r}[(\hat{\theta}_n-\theta^*)]_j(\hat{\theta}_n-\theta^*)^T\left[\frac{\partial}{\partial\theta_j}\nabla^2\ell^n(\tilde{\theta}_n)\right](\hat{\theta}_n-\theta^*) \\
=&\frac{1}{6}\left\{\mathbb{E}[(\phi^Tt)^3] + 2[\mathbb{E}(\phi^Tt)]^3 - 3\mathbb{E}[\phi^Tt]\mathbb{E}[(\phi^Tt)^2]\right\},
\end{aligned}$$

where $\phi = \phi(x, y)$ and $\mathbb{E}$ is taken with respect to $p_{\tilde{\theta}_n}$. Applying some basic inequality gives

$$\frac{1}{6}\sum_{j=1}^{r}[(\hat{\theta}_n-\theta^*)]_j(\hat{\theta}_n-\theta^*)^T\left[\frac{\partial}{\partial\theta_j}\nabla^2\ell^n(\tilde{\theta}_n)\right](\hat{\theta}_n-\theta^*)\leqslant 0. \tag{8}$$

Using the fact that $\ell^{(n)}(\hat{\theta}_n) \geqslant \ell^{(n)}(\theta^*)$, we obtain the following inequality from Eq. (7) and Ineq. (8)

$$[\nabla\ell^{(n)}(\theta^*)]^T(\hat{\theta}_n-\theta^*)+\frac{1}{2}(\hat{\theta}_n-\theta^*)^T\nabla^2\ell^{(n)}(\theta^*)(\hat{\theta}_n-\theta^*)\geqslant 0.$$

Then, we have

$$\begin{aligned}
&|\nabla\ell^{(n)}(\theta^*)\|_2\|\hat{\theta}_n-\theta^*\|_2\\
\geqslant&[\nabla\ell^{(n)}(\theta^*)]^T(\hat{\theta}_n-\theta^*)\\
\geqslant&-\frac{1}{2}(\hat{\theta}_n-\theta^*)^T\nabla^2\ell^{(n)}(\theta^*)(\hat{\theta}_n-\theta^*)\\
=&\frac{1}{2}(\hat{\theta}_n-\theta^*)^T\mathrm{var}_{\theta^*}[\phi|x](\hat{\theta}_n-\theta^*)\\
\geqslant&\frac{1}{2}\Lambda_{min}(\mathrm{var}_{\theta^*}[\phi|x])\|\hat{\theta}_n-\theta^*\|_2^2\\
\geqslant&\frac{1}{2}C_{min}\|\hat{\theta}_n-\theta^*\|_2^2.
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{P}\left[\|\hat{\theta}_n-\theta^*\|_2>\varepsilon\right]\\
\leqslant&\mathbb{P}\left[\|\nabla\ell^{(n)}(\theta^*)\|_2>\frac{\varepsilon C_{min}}{2}\right]\\
\leqslant&\frac{4r^2\phi_{max}^2}{n\varepsilon^2C_{min}^2}\left[\frac{3}{2}\left(\frac{1}{m}+1\right)\binom{d}{m-1}d^{\lambda^*}+1\right].
\end{aligned}$$

Therefore, the MLE learns the parameter within L_2 error ε with probability at least $1-\delta$, provided that

$$n>\frac{4r^2\phi_{max}^2}{\varepsilon^2\delta C_{min}^2}\left[\frac{3}{2}\left(\frac{1}{m}+1\right)\binom{d}{m-1}d^{\lambda^*}+1\right].$$

Hence we complete the proof. □

4.3 Discussions on Sample Complexity Bounds

The first conclusion to be drawn from the sample complexity bound is that as λ^* becomes larger (e,g. the data dependence becomes stronger), the sample complexity of MLE will *remarkably* increase. Second, we find that larger C_{min} will make the parameter be learnt easier, which agrees with our geometric intuitions that the maximum of more curved surfaces can be estimated with smaller searching steps. In addition, it is easy to see that more samples are required for larger m, i.e., more nodes of each training clique are taken. As a consequence of this point, we expect that estimators using smaller cliques can be more statistically efficient than those using larger cliques. Finally, dependent data in our problem makes our bound with another term concerning with clique dependence, which grows rapidly as λ^* increases.

Compared with recent sample complexity results of learning MRF and CRFs, we find that our sample complexity result has the similar form with Corollary 4.2 in (Bradley and Guestrin, 2012), although parameter estimations in our and their analysis are tested within L_2 and L_1 errors, respectively. For instance, ε, ϕ_{max} and C_{min} in our result have the same form with that in (Bradley and Guestrin, 2012). However, roles of δ in sample complexities in our paper and those in (Bradley and Guestrin, 2012) are totally different. More explicitly, the term of $1/\delta$ in ours bound requires much more samples than learning MRFs and CRFs with i.i.d. data, where the form of δ in (Bradley and Guestrin, 2012) is just $\log(1/\delta)$. This intrinsic disparity of δ terms illustrates why learning with dependent data is much more difficult than learning with i.i.d. data.

The difficulty of learning with dependent data can be seen much more clearly from distinct convergence rates of the gradients of likelihood functions. The order of convergence rates of the gradient infinity norm with n samples in our analysis is $\mathcal{O}(n^{-1})$ from the inequality (6). However, the order of convergence rates of the gradient infinity norm in (Bradley and Guestrin, 2012) is $\mathcal{O}(e^{-n})$, which can be found in Lemma 9.1 in the appendix of that paper. Notice that δ in Lemma 9.1 in (Bradley and Guestrin, 2012) corresponding to t in the inequality (6) of our paper. In summarize, the fact of the subexponential order of gradient convergence rates reveals the inherent difficulty of learning with dependent data.

4.4 Consistency of MLE

Owing to Theorem 4.1, it is evident that $\hat{\theta}_n$ will converge to θ^* in probability as the number of samples tends to infinity ($n \to \infty$). Hence we prove the (weak) consistency of the MLE estimators. In view of the consistency of estimators is of major importance both in theory and practice, we state the result in the following theorem. The proof is trivial and thus omitted.

Theorem 4.2 (MLE Consistency)**.** *Suppose Assumptions 2.1, 2.2 and 2.3 hold, then the sequence of estimators $\hat{\theta}_n$ from maximal likelihood estimation converges in probability to the true parameter θ^*.*

5 Concluding Remarks

We proved the first PAC bound for learning parameters of relational learning with one network under suitable conditions. The sample complexity we obtained reveals various roles of problem-specific constants. The

crucial issue of analyzing MLE performances is estimating convergence rates of gradients of the normalized likelihood function. Furthermore, the estimation of convergence rates of gradients depends upon dependencies of relational data. In this paper, we deal with this difficulty by careful combinational analysis of dependent cliques. We believe that the proposed combinational method and results in this paper will also be useful for solving other problems of relational learning.

Last but not the least, while we have considered only the MLE in this paper, our analysis framework can be extended to other learning methods such as maximum pseudolikelihood estimation (MPLE) and maximum composite likelihood estimation (MCLE).

Acknowledgements

We thank anonymous reviewers for helpful comments. This research is supported by 973 Program (2013CB329503)NSFC (Grant No. 91120301) and Beijing Municipal Education Commission Science and Technology Development Plan key project under grant KZ201210005007.

References

P. Abbeel, D. Koller, and A. Y. Ng. Learning factor-graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.* **7**:1743-1788, 2006.

J. K. Bradley and C. Guestrin. Sample complexity of composite likelihood. *International Conference on Artificial Intelligence and Statistics*, 2012.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: methods, theory and applications.* 2011.

J. Dedecker, P. Doukhan, G. Lang, J. R. Leon, S. Louhichi, and C. Prieur. *Weak Dependence: With Examples and Applications.* Springer, 2007.

J. D. Lee, Y. Sun, and J. Taylor. On model selection consistency of M-estimators with geometrically decomposable penalties. *Advances in Neural Information Processing Systems*, 2013.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of International Conference on Machine Learning*, 2001.

P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. *Proceedings of International Conference on Machine Learning*, 2008.

B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221-239, 1988.

S. N. Negahban, P. Ravikumar, P. Ravikumar, M. J. Wainwright and B. Yu. A Unified framework for high-dimensional analysis of M-Estimators with decomposable regularizers. *Statis. Sci.* **27**(4), 538-557, 2012.

J. Neville and D. Jensen. Relational dependency networks. *J. Mach. Learn. Res.***8**:653-692, 2007.

M. Richardson and P. Domingos. Markov logic network. *Mach. Learn.* **62**(1-2):107-136, 2006.

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p^*) models for social networks. *Social networks* **29**(2), 173-191, 2007.

M. Sinn and B. Chen. Central limit theorems for conditional Markov chains. *International Conference on Artificial Intelligence and Statistics*, 2013.

M. Sinn and P. Poupart. Asymptotic theory for linear-chain conditional random fields. *International Conference on Artificial Intelligence and Statistics* , 2011.

B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. *Proceedings of The Conference on Uncertainty in Artificial Intelligence*, 2002.

P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics* **38**(3):1287-1319, 2010.

C. Sutton and A. McCallum. *An Introduction to Conditional Random Fields for Relational Learning.* MIT Press, 2006.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge, 2000.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory* **55**(5):2183-2202, 2009.

M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference.* Now Publisher Inc., Hanover, MA, USA, 2008.

L. G. Valiant. A theory of the learnable. *Comm. ACM,* **27**(11):1134-1142, 1984.

R. Xiang and J. Neville. Relational learning with one network: an asymptotic analysis. *International Conference on Artificial Intelligence and Statistics*, 2011.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**:2541-263, 2006.