Efficient Localized Inference for Large Graphical Models

Jinglin Chen1, Jian Peng1, Qiang Liu2

1 University of Illinois at Urbana-Champaign
2 University of Texas at Austin
jinglinc@illinois.edu, jianpeng@illinois.edu, lqiang@cs.utexas.edu

Abstract

We propose a new localized inference algorithm for answering marginalization queries in large graphical models with the correlation decay property. Given a query variable and a large graphical model, we define a much smaller model in a local region around the query variable in the target model so that the marginal distribution of the query variable can be accurately approximated. We introduce two approximation error bounds based on the Dobrushin’s comparison theorem and apply our bounds to derive a greedy expansion algorithm that efficiently guides the selection of neighbor nodes for localized inference. We verify our theoretical bounds on various datasets and demonstrate that our localized inference algorithm can provide fast and accurate approximation for large graphical models.

1 Introduction

Probabilistic graphical models such as Bayesian networks, Markov random fields, and conditional random fields are powerful tools for modeling complex dependencies over a large number of random variables ([Koller and Friedman, 2009; Wainwright et al., 2008]). Graphs are used to represent joint probability distributions, where nodes denote random variables, and edges represent dependency relationships between different nodes. With the specification of a graphical model, a fundamental problem is to calculate the marginal distributions of variables of interest. This problem is closely related to computing the partition function, or the normalization constant of a graphical model, which is known to be intractable and #P-complete. As a result, developing efficient approximation inference algorithms becomes a pressing need. The most popular algorithms include deterministic variational approximation inference algorithms becomes a pressing need.

In this work, we develop a new localized inference method for very large graphical models. Our approach leverages the Dobrushin’s comparison theorem that casts explicit bounds based on the correlation decay property in the graphs, in order to restrict the inference to a smaller local region that is sufficient for the inference of marginal distribution of the query variable. The use of the Dobrushin’s comparison theorem allows us to explicitly bound the truncation error which guides the selection of localized region from the original large graph. Extensive experiments demonstrate both the effectiveness of our theoretical bounds and the accuracy of our inference algorithm on a variety of datasets.

Related Work

Approximate inference algorithms of graphical models have been extensively studied in the past decades (see [Koller and Friedman, 2009; Wainwright et al., 2008; Dechter, 2013]). Query-specific inference, including [Chechetka and Guestrin, 2010] which proposed a focused belief propagation for query specific inference, and [Wick and McCallum, 2011; Shi et al., 2015] which study query-aware sampling algorithms, have recently been introduced for large graphical models. Compared with these methods, our work is theoretically motivated by the Dobrushin’s comparison theorem and enables us to efficiently construct the localized region in a principled and practically efficient manner.

2 Background on Graphical Models

Graphical models provide a flexible framework for representing relationships between random variables [Heine-
mann and Globerson, 2014]. In graph $G$, we use $X = (X_1, X_2, \cdots, X_n)$ to denote a finite collection of $n$ random variables and we use $x = (x_1, x_2, \cdots, x_n)$ to refer to an assignment. Suppose $E$ is a set of edges and $\theta$ is a set of functions with $\theta_{ij}(x_i, x_j)$ for edge $(i, j) \in E$ and $\theta_i(x_i)$ for node $i \in \{1, 2, \cdots, n\}$. We use $\mu(x; \theta)$ to represent the joint distribution of the graphical model $(G, E, \theta)$ as following.

$$\mu(x; \theta) = \frac{1}{Z_{\mu}(\theta)} \exp \left( \sum_{(i, j) \in E} \theta_{ij}(x_i, x_j) + \sum_i \theta_i(x_i) \right),$$

where $Z_{\mu}(\theta)$ is the normalization constant (also called the partition function).

In this work, we will focus on the Ising model, an extensively studied graphical model. The Ising model is a pairwise model with binary variables $x_i \in \{-1, +1\}$. The pairwise and singleton parameters are defined as follows

$$\theta_{ij}(x_i, x_j) = \begin{pmatrix} -J_{ij} & J_{ij} \\ J_{ij} & -J_{ij} \end{pmatrix}, \quad \theta_i(x_i) = \begin{pmatrix} -h_i \\ h_i \end{pmatrix}.$$

So the distribution of an Ising model is defined as,

$$\mu(x; \theta) = \frac{1}{Z_{\mu}(\theta)} \exp \left( \sum_{(i, j) \in E} J_{ij} x_i x_j + \sum_i h_i x_i \right). \quad (1)$$

Given a graphical model, marginal inference involves calculating the normalization constant, or the marginal probabilities of small subsets of variables. These problems require summation over an exponential number of configurations and are typically #P-hard in the worst case for loopy graphical models. However, practical problems can be often easier than expected. In many applications, we are only interested in the correlation decay property has been widely studied in statistical mechanics and graphical models [Rebeschini et al., 2015].

3 Localized Inference and Correlation Decay

Given a large graphical model, marginal inference involves calculating the normalization constant, or the marginal probabilities of small subsets of variables. These problems require summation over an exponential number of configurations and are typically #P-hard in the worst case for loopy graphical models. However, practical problems can be often easier than the theoretically worst cases, and it is still possible to obtain efficient approximations by leveraging the special structures of given models. In this work, we focus on the query-specific inference, where the goal is to calculate the marginal distribution $\mu(x_i)$ of given individual variable $x_i$. For this task, it is possible to make good approximations based on a local region around $x_i$, thus significantly accelerates the inference in very large graphical models.

Theorem 1 [Föllmer, 1982] Dobrushin’s comparison theorem Let $\mu$ be a Gibbs measure on a finite product space $E = S^I$, where $I$ is an index set. For $i, j \in I$, we define

$$C_{ij} = \frac{1}{2} \sup \{ ||\mu_i(x_i|x) - \mu_i(y_i|y)|| : x_k = y_k, \forall k \neq j \},$$

where $\mu_i(x_i|x)$ is the conditional distribution of the $i$th coordinate with respect to the $\sigma$-field generated by the coordinates with index $j \neq i$, and $\| \cdot \|$ is the total variance distance. We compute

$$c = \max_{i \in I} \sum_{j \in I} C_{ij},$$

and assume $c < 1$. Let $C = (C_{ij})_{i,j \in I}$ and $D = \sum_{n=0}^{\infty} C^n = (I - C)^{-1}$, then for any probability measure $\nu$ on the same place and any function $f$, we have

$$\left| \int f d\mu - \int f d\nu \right| \leq \sum_{i \in I} (Db)_i \times \delta_i(f),$$
where \( b = (b_j) \) is the singleton perturbation coefficient of node \( x_j \):
\[
b_j = \frac{1}{2} \sup_x \| \mu_j(x_j | x) - \nu_j(x_j | x) \|, \quad (3)
\]
and \( \delta_i(f) \) is the oscillation of \( f \) in the \( i \)th coordinate, that is,
\[
\delta_i(f) = \max_{x_i, x'_i} | f(x_{-i}, x_i) - f(x_{-i}, x'_i) |.
\]

In Theorem 1, \( \mu_i(x_i | x) \) is the probability of variable \( i \) conditioned on its adjacent variables whose assignments are the same as corresponding entries in \( x \). According to the Markov property, calculating \( \mu_i(x_i | x) \) only requires information from the local star-shaped graph (the figure can be found in the full version). It is worth noting that a tighter bound can be obtained by defining \( b \) to be
\[
b_j = \frac{1}{2} \int \| \mu_j(x_j | x) - \nu_j(x_j | x) \| \nu(dx). \quad (4)
\]
Here we use the definition in (3) for lower computational complexity. The matrix \( C \) is known as the Dobrushin’s interaction matrix, and the inequality \( c = \max_{i \in I} \sum_{j \in I} C_{ij} < 1 \) is the Dobrushin condition. If this condition holds, the theorem can give us a bound between two measures, which is the result of correlation decay.

In the following, we will apply Theorem 1 to undirected graphical models to derive an approximation bound of marginal distributions. We first denote \( I = \{1, 2, \cdots, n\} \) by the index set of the variables and assume that we want to query the marginal distribution of variable \( x_i \). In order to apply Theorem 1, we set \( f(x) \) to be the indicator function of the variable \( x_i \), that is, \( f(x) = f(x_1, x_2, \cdots, x_n) = \mathbb{I}[x_i = k] \). Then
\[
| \int f d\mu - \int f d\nu | \text{ becomes the absolute marginal difference } | \mu(x_i = k) - \nu(x_i = k) | \text{ between the two measures } \mu \text{ and } \nu.
\]
In addition, the oscillation of function \( f \) is thus reduced to \( \delta_i(f) = 1 \) and \( \delta_j(f) = 0 \). With these simplifications, we obtain a bound of the maximum difference between marginals of the queried node \( i \) for two measures:

**Corollary 1** Followin the assumptions in Theorem 1 and the above text, we have
\[
\max_k | \mu(x_i = k) - \nu(x_i = k) | \leq (Db)_i. \quad (4)
\]

Note that the roles of \( \mu \) and \( \nu \) in (4) are not symmetric because the Dobrushin coefficient \( D \) is solely defined based on \( \mu \) (and independent with \( \nu \)). As a result, there are two ways to use bound (4) for local inference, depending on whether we treat \( \mu \) or \( \nu \) as the original model that we want to query or the localized model that we use for approximation, respectively. We will next exploit both possibilities in the next sections. In Section 4, we take \( \mu \) as the global model (or measure) and \( \nu \) as the localized model (or measure) and derive a simple upper bound relates the approximation error to the distance between the query node and the boundary of the local region on the graph. In Section 5, we take \( \nu \) as the global model (or measure) and \( \mu \) as the localized model (or measure), we derive another upper bound that only involves the localized region, and leverage it to propose a greedy expansion algorithm to construct the localized model with guaranteed approximations.

### 4 Distance-based Upper Bound

In this section, we assume that \( \mu = \mu(x_1, x_2, \cdots, x_n) \) in Theorem 1 is defined by the original graphical model that we want to query, and \( \nu \) is a simpler and more tractable distribution that we use to approximate the marginal of \( x_i \) in \( \mu \).

For notational simplicity, we partition the node-set to two disjoint sets \( \alpha \) and \( \beta = \{1, 2, \cdots, n\} \setminus \alpha \), where \( \alpha \) is the local subgraph that contains query node \( i \in \alpha \) and \( \beta \) is the rest of the graph. We use \( \partial \alpha \) and \( \partial \beta \) to represent the set of subscripts of nodes on the boundary and in the interior of \( \alpha \). Obviously, \( \partial \alpha \subseteq \alpha \) and \( \partial \alpha \cap \partial \beta = \emptyset \). Similarly, \( \partial \beta \subseteq \beta \) and \( \partial \beta \cap \partial \alpha = \emptyset \). In addition, we use \( x_\alpha \) to denote the variables in \( \alpha \) and \( x_\beta \) to denote the variables in \( \beta \). We will first apply the following lemma to obtain our first result on the relationship between the approximation error of marginals and the radius of the local subgraph \( \alpha \).

**Lemma 1** [Rebeschini and van Handel, 2014] Assume \( I \) is a finite set and let \( m \) be a pseudo-metric on set \( I \). \( C = (C_{ij})_{i,j \in I} \) is a non-negative matrix. Suppose that
\[
\max_{i \in I} \sum_{j \in I} e^{m(i, j)} C_{ij} \leq d < 1.
\]

Then matrix \( D = \sum_{n=0}^{\infty} C^n = (I - C)^{-1} \) satisfies
\[
\max_{i \in I} \sum_{j \in I} e^{m(i, j)} D_{ij} \leq \frac{1}{1 - d}.
\]

In particular, this implies that
\[
\sum_{j \in J} D_{ij} \leq \frac{e^{-m(i, j)}}{1 - d}
\]
for every set \( J \subseteq I \), where \( m(i, j) = \max_{j \in J} m(i, j) \).

This lemma indicates that if \( C_{ij} \) decays exponentially with the distance between \( i \) and \( j \), the \( D_{ij} \), which is used in Theorem 1 and Corollary 1, also decays exponentially with the distance between \( i \) and \( j \). The condition of this correlation decay lemma is usually mild in practice. When we choose \( m \equiv 0 \), which is naturally a pseudo-metric, and use Dobrushin’s interaction matrix as \( C \), the conditions of the lemma hold once the Dobrushin condition is satisfied, because matrix \( C \) in Theorem 1 is by definition a non-negative matrix and hence \( D \) is also non-negative, and every entry in \( b \) is less than \( 1/2 \). Applying Lemma 1, we can obtain the following result.

**Theorem 2** Suppose \( \mu \) is the probability measure for a graphical model for which we want to query the marginal distribution of node \( i \). Let \( \nu \) be the another probability on the same space, whose parameters of edges on subgraph \( \alpha \) and parameters of nodes in \( \alpha \) are the same as \( \mu \). Assume the Dobrushin condition holds for \( \mu (c = \max_{i \in I} \sum_{j \in I} C_{ij} < 1) \).

Let \( d(i, \partial \alpha) \) denote the distance between node \( i \) and node-set \( \partial \alpha \) on the Markov graph \( G \) of \( \mu \). If we assume
\[
d(i, \partial \alpha) \geq \frac{\ln t}{\ln (1 - (1 - c)/tc)} \text{ for some } t > 1,
\]
then for every \( \varepsilon > 0 \), we have
\[
\max_k | \mu(x_i = k) - \nu(x_i = k) | \leq \varepsilon.
\]
This theorem characterizes the error bound when approximating the global model \( \mu \) using another model \( \nu \) that matches \( \mu \) locally in region \( \alpha \). Our result shows that in order to ensure an \( \varepsilon \) bound on the query node \( i \), the distance \( d(i, \partial \alpha) \) from the query node \( i \) to the boundary \( \partial \alpha \) should be at least linear to \( \ln(1/\varepsilon) \). In other words, the error \( \varepsilon \) decreases exponentially with \( d(i, \partial \alpha) \). The proof of Theorem 2 can be found in the Appendix C.

As a result, given \( c \) and \( \varepsilon \), we can get the minimum value of the lower bound of \( d(i, \partial \alpha) \) by optimizing \( t \). Theorem 2 gives a simple but general way to bound the local subgraph of variable \( x_i \), as we only need to check the Dobrushin condition and compute \( c \) on the whole true graphical model.

5 Localized Bound and Greedy Expansion

The bound in Theorem 2 requires computing the value \( c \) as defined in (2) for a given graphical model. However, since \( c \) is the maximum \( C_{ij} \) of the entire graph, it can be very expensive to compute when the graph is large. In this section, we explore another approach of using the bound in Corollary 1, by setting \( \nu = \nu(x_1, x_2, \ldots, x_n) \) to be the distribution of the original graphical model and \( \mu \) to be the localized model. In this way, we will derive a novel approximation approach by greedily constructing a local graph from the query variable \( i \), with guaranteed upper bounds of the approximation error between marginal distributions of \( \mu \) and \( \nu \).

To start with, we note that \( \nu \) can be decomposed to

\[
\nu(x) = \frac{\psi_\alpha(x_\alpha)\psi_\beta(x_\beta)\psi_{\alpha\beta}(x_\alpha, x_\beta)}{Z_\nu},
\]

where \( \psi_\alpha \) is the exponential of a potential function of \( x_\alpha \), \( \psi_\beta \) is the exponential of potential function of \( x_\beta \), and \( \psi_{\alpha\beta} \) is the exponential of potential function defined on \( x_{\alpha\beta} \) and \( x_{\beta\alpha} \).

We want to approximate \( \nu \) with a simpler model \( \mu \) in which the nodes in \( \alpha \) and \( \beta \) are disconnected, so that the inference over \( i \in \alpha \) can be performed locally within \( \alpha \), irrelevant to the nodes in \( \beta \). Formally, we want to approximate \( \nu \) by

\[
\mu(x) = \frac{\psi_\alpha(x_\alpha)\psi_\beta(x_\beta)\tilde{\psi}_{\alpha\beta}(x_{\alpha\beta})\tilde{\psi}_{\beta\alpha}(x_{\beta\alpha})}{Z_\mu},
\]

where the factors \( \tilde{\psi}_{\alpha\beta} \) and \( \tilde{\psi}_{\beta\alpha} \) are with approximations \( \tilde{\psi}_{\alpha\beta} \) and \( \tilde{\psi}_{\beta\alpha} \). Therefore, the marginal distributions of \( x_\alpha \) and \( x_\beta \) get decoupled in \( \mu \), that is,

\[
\mu(x) = \mu(x_\alpha)\mu(x_\beta).
\]

This decomposition thus allows us to approximately calculate marginal \( \mu(x_i) \) efficiently within subgraph \( \alpha \). The challenges here are 1) how to construct the factors \( \tilde{\psi}_{\alpha\beta} \) and \( \tilde{\psi}_{\beta\alpha} \) in \( \mu \) to closely approximate \( \nu \), 2) how to decide the subgraph region and 3) how to bound the approximation error. We consider two methods for constructing \( \tilde{\psi} \) in this work:

1. **[Dropping out]** Simply remove the \( \tilde{\psi}_{\alpha\beta} \) in \( \nu \). To do so, we set

\[
\tilde{\psi}_{\alpha\beta} = 1.
\]

This corresponds to directly remove all the edges between \( \partial \alpha \) and \( \partial \beta \), which is also referred as the “dropping out” method in our experiments.

2. **[Mean field]** Find \( \tilde{\psi}_{\alpha\beta} \) to closely approximate \( \tilde{\psi}_{\alpha\beta} \) by performing a mean field approximation, that is, we solve the following optimization problem:

\[
\min_{\tilde{\psi}_{\alpha\beta}} KL(\tilde{\psi}_{\alpha\beta} \| \psi_{\alpha\beta}),
\]

where the \( KL(\cdot||\cdot) \) refers to the KL divergence of the corresponding normalized distributions. To apply the mean field approximation and reduce complexity, we further assume that the nodes are independent in \( \psi_{\alpha\beta} \) and \( \psi_{\beta\alpha} \). By using the optimized approximation \( \tilde{\psi}_{\alpha\beta} \), we will be able to compensate the error of marginal of \( x_i \), which is introduced by simply removing the edges between \( \partial \alpha \) and \( \partial \beta \), as mentioned above.

Note that the potentials \( \tilde{\psi}_{\beta\alpha} \) and \( \tilde{\psi}_{\alpha\beta} \) in \( \mu \) do not influence the calculation of \( \mu(x_i) \), for \( i \in \alpha \). Therefore, we remove all the edges in \( \beta \). The marginal of node \( i \in \alpha \) will not change.

By applying Corollary 1, we can now obtain an error bound which, remarkably, only involves the local region \( \alpha \). The details can be found in Appendix B.

**Corollary 2** Assume \( \mu(x) = \mu(x_\alpha)\mu(x_\beta) \), and the conditions in Theorem 1 holds, we have

\[
\max_k |\mu(x_i = k) - \nu(x_i = k)| \leq \sum_{j \in \partial \alpha} D_{ij} b_j,
\]

where \( b_j \) is defined in Eq 3, and \( D \) is defined by \( D_{\alpha\beta} = (I - C_{\alpha\beta})^{-1} \); here \( C \) is defined in Theorem 1.

Note that the upper bound in (8) only involves the local region \( \alpha \) and hence can be computed efficiently using mean field or belief propagation within the subgraph on \( \alpha \). The proof of Corollary 2 and the details on how to calculate \( C \) and \( D \) for Ising models in practice can be found in the full version.

Using the bound in (8), we propose a greedy algorithm to expand the local graph starting from query node \( i \) incrementally. At iteration, we add a neighboring node that yields the tightest bound using the above bound and repeat this process until the bound is tight enough or a maximum of graph size is reached. This process is summarized in Algorithm 1. After we complete the expanding phase, we can apply exact inference or on local region \( \alpha \) to calculate the marginal of the query variable \( x_i \) if the size of \( \alpha \) is small or perform approximate inference methods if the size of \( \alpha \) is medium. The actual size of \( \alpha \) can vary in different graphical models, which is mainly determined by the correlation decay property near the query variable \( x_i \) or the tightness of the upper bound in Eq (8).

**Computational Complexity** We find that both the complexity of expanding the local subgraph and the complexity of localized inference can be bounded. The details will be included in the Appendix A.

6 Experiments

We test our algorithm on both simulated and real-world datasets. The results indicate that our method provides an efficient localized inference technique.
Algorithm 1 Greedy expansion algorithm for localized inference
1: given a graphical model $\nu$ and a node $i$, approximate
marginal probability $\nu(x_i)$
2: input: $K$ = maximum number of nodes in the local sub-
graph $\alpha$ and $\delta$ = the improvement threshold
3: initialize local subgraph $\alpha = \{i\}$ and $bound_{best} = 1$
4: while $|\alpha| < K$ ($|\alpha|$ is the number of nodes in $\alpha$) do
5:     set $\beta = \{1, 2, \cdots, n\} \setminus \alpha$ and $\partial \beta$ the nodes in $\beta$ that
     connects with $\alpha$ in $\nu$.
6:     for node $k \in \partial \beta$ do
7:         add node $k$ to $\alpha$ and get a candidate local subgraph
     $\alpha_{k}^{new} = \alpha \cup \{k\}$.
8:         construct local model $\mu$ by setting $\tilde{\psi}_{\alpha} = 1$ (dropping
     out, Eq (6)) or estimating it using mean field as in
     Eq (7).
9:         calculate the bound $bound_k$ in (8) where the $bound_k$
     refers to $\sum_{j \in \partial \alpha} D_{ij} b_j$
10:     end for
11:     if $\min_{k \in \partial \beta} bound_k < bound_{best} - \delta$ then
12:         update $bound_{best} = \min_{k \in \partial \beta} bound_k$
13:     if $\alpha = \alpha \cup \{\arg \min_{k \in \partial \beta} bound_k\}$
14:     end if
15: end while

6.1 2D Ising Grid
In this section, we perform experiments on 2D-grid Ising
models and regard the localized probability as $\nu$ and regard
the true probability as $\mu$. The graph is a $10 \times 10$ lattice and
the coordinate of query node is $(5, 5)$. The parameters in the Ising
model is generated by drawing $h_i$ uniformly from $[-I_1, I_1]$ for
all nodes $i$ and $J_{ij}$ uniformly from $[-I_2, I_2]$ for all edges
$(i, j)$. Here $I_1$ and $I_2$ control the locality and hardness of
this Ising model.

Checking Dobrushin’s condition We start with numeri-
cally checking the Dobrushin condition $\epsilon = \max_{ij} \sum_j C_{ij} < 1$.
We find that $\epsilon$ is smaller than one in most regions, but is
larger than one when $I_2$ is very large and $I_1$ is very small, in
which case the nodes are strongly coupled together (no cor-
relation decay) and there is no significant local information.
The hope, however, is that real problems tend to be easier be-
cause a large amount of information is available. The figure
is shown in the full version.

Comparing Different Expansion Algorithms In this part,
we compare the true approximation error $\max_k \|\mu(x_i = k) - \nu(x_i = k)\|$ to the bound $\sum_{j \in \partial \alpha} D_{ij} b_j$ given by our algorithm when we expand the local subgraph. The true er-
ror $\max_k \|\mu(x_i = k) - \nu(x_i = k)\|$ is evaluated using
the brute-force algorithm. When removing the bipartite graph,
we try both simply dropping edges and the mean field ap-
proximation. In all the experiments, we use the UGM Matlab
package \footnote{\url{http://www.cs.ubc.ca/~schmidt/Software/UGM.html}} for the mean field approximation.

In order to better compare the error, we also add two bas-
elines. One baseline is that we expand the local subgraph in
each step by randomly selecting a node in the boundary $\partial \beta$.
Another baseline is that we expand the local subgraph greed-
ily by choosing the node in $\partial \beta$ that has the maximum $L_2$

norm over the edge-set between such node and the subgraph
$\alpha$. The intuition is that when the the edges’ weights are large,
the node may be more related to the nodes in the subgraph.

In Figure 2, we compare our greedy expansion method
stated in Algorithm 1 to the baselines stated above. For this
experiment, we fix $I_1 = 1$ and $I_2 = 0.25$ and average on
100 random trials. We stop expanding the graph when the
local subgraph contains 16 nodes. We calculate the mean value
of the true errors and bounds in the 100 trials for a different
number of nodes in the subgraph.

From Figure 2, we can see that, when combined the drop-
ping out method for constructing $\psi$, our greedy expansion
method significantly outperforms the two baselines. We also
find that the mean field method for constructing $\tilde{\psi}$ gives about
the same true error as the dropping out method, but provides
a tighter upper bound. It is interesting to note that the true
errors of the two baseline expansion methods are sometimes
even worse than the upper bounds of our greedy expansion,
indicating the strong advantage of our method.

We further investigate how the parameters of the Ising
model may influence the results of the algorithms and the
tightness of the bound. For this purpose, we fix $I_2 = 0.25$ and vary $I_1$ in the range of $\{0, 0.5, 1, \cdots, 10\}$ in Figure 3.
For each setting, we simulate 100 times and then calculate
the mean error and bound. From Figure 3, we can find that
the bound is again relatively tight, especially when the value
of $I_1$ is large. both the bounds and the true errors decrease
as $I_1$ increases because the correlation decay is stronger and
the inference task is easier with strong local evidence on the
singleton potentials (large $I_1$).

6.2 Cora Data Set
We perform experimental evaluations on the Cora data set\footnote{https://people.cs.umass.edu/~mccallum/data.html}.
Cora consists of a large collection of machine learning pa-
ners with citation relations between the papers, in which each

Figure 2: The true errors and our upper bounds for the marginal ap-
proximation when we expand the local subgraphs to different sizes.
In this paper, we address query-specific marginal inference in large-scale graphical models using a new localized inference algorithm. We leverage the Dobrushin’s comparison theorem to derive two error bounds for localized inference, including a simple bound based on graph distance and a localized bound from which we derive an efficient greedy expansion algorithm for constructing local regions for localized inference. Our experiments have shown that our bounds are practically useful and the algorithm works efficiently on various graphical models. Future directions include theoretical investigation on tighter bounds and development of more efficient greedy expansion algorithms.

Figure 4: The accuracy of different algorithms when $I_1$ changes. Red and blue: the accuracy of the labels given by the global and the local inference evaluated w.r.t. the true labels. Green: the accuracy of the local inference evaluated w.r.t. the labels provided by the global inference.

Figure 5: Precision, Recall, and F-measure of the labels given by local inference vs. global inference when the value of $I_1$ changes as well. We also report in Figure 5 the precision, recall, and F-measure when comparing local inference with global inference, by treating label 1 as positives. Both figures show that when $I_2$ is fixed and $I_1$ increases, which means that the correlation decay is stronger, our local inference method achieves better results.

7 Conclusion

In this paper, we address query-specific marginal inference in large-scale graphical models using a new localized inference algorithm. We leverage the Dobrushin’s comparison theorem to derive two error bounds for localized inference, including a simple bound based on graph distance and a localized bound from which we derive an efficient greedy expansion algorithm for constructing local regions for localized inference. Our experiments have shown that our bounds are practically useful and the algorithm works efficiently on various graphical models. Future directions include theoretical investigation on tighter bounds and development of more efficient greedy expansion algorithms.
A Computational Complexity

Here we consider the computational complexity of expanding the local subgraph and the complexity of localized inference. We always suppose that the maximum degree of the graph $G$ is $d$ and we define the maximum distance between the query node and any node in the subgraph to be the radius of the subgraph. First, given a threshold $\epsilon$, from Theorem 2, we just need a subgraph with radius

$$r = \inf_{\epsilon > 1} \left[ \frac{\ln \frac{t}{2(\ln(1-\epsilon))}}{\ln \frac{1}{1-\epsilon}} \right],$$

where we recall that $c$ is the Dobrushin coefficient $c = \max_{i \in I} \sum_{j \in I} C_{ij}$. In particular, taking $t = 2$ shows that we just need $r = \lfloor \ln(1-\epsilon) \rfloor / \ln(1-\epsilon)$, which is worth noting that $r$ decreases when $\epsilon$ becomes small and/or the accuracy threshold $\epsilon$ becomes large. Since the size of the subgraph with radius $r$ is no more than $1 + d + d^2 + \cdots + d^r = (d^{r+1} - 1)/(d - 1)$, it can be much smaller than the whole graph. As a result, the inference over the subgraph is much more efficient.

Then, we discuss the computation complexity in each expansion step (Algorithm 1, line 6-10). We need to loop over the nodes in $\partial \Gamma$. In the loop, we need to calculate the vector $b$ and the matrix $D$. The calculation for each element in $b$ requires the enumeration of different assignments in the neighborhood of such node, which is bounded because it is not related to the size of the whole graph. In the calculation of matrix $C$, we only need to update the elements related to the new node. The number of such elements is no more than $d$ and the calculation of each element is not related to the size of the whole graph. $D$ can be derived from $C$ and use historical information to calculate incrementally. The complexity is no more than computing the inverse of the whole matrix $I - C_{\partial \alpha}$. If we use mean field approximation in the greedy expansion, the computation is also cheap because the sizes of $\partial \alpha$ and $\partial \beta$ are small.

B Proof of Corollary 2

Proof 1 Note that the Dobrushin’s interaction matrix $C$ of $\mu$ is a block diagonal matrix. Since there are no edges between $\alpha$ and $\beta$, the corresponding blocks equal to zero. If the Dobrushin condition holds, $D$ would also be a block-diagonal matrix and can be calculated easily from $C$. To see this, we have

$$C = \begin{bmatrix} \alpha & \beta \\ C_{\alpha\alpha} & O \\ \alpha & C_{\beta\beta} \end{bmatrix},$$

and

$$D = \begin{bmatrix} \alpha & \beta \\ (I - C_{\alpha\alpha})^{-1} & O \\ \alpha & (I - C_{\beta\beta})^{-1} \end{bmatrix}.$$

Applying the bound in Corollary (1) gives the result.

C Proof of Theorem 2

Proof 2 Let $m(i, j) = d(i, j) \ln \frac{1 + (t-1)c}{tc}$, where $d(i, j)$ represents the distance between node $i$ and node $j$ in the graph and $t > 1$. Then

$$\max_{i \in I} \sum_{j \in I} e^{m(i, j)} C_{ij} = \frac{1 + (t-1)c}{tc} \max_{i \in I} \sum_{j \in I} e^{d(i, j)} C_{ij} \leq 1 + \frac{(t-1)c}{t} < 1.$$

Applying Lemma 1, we have

$$\max_{k = \{1, -1\}} |\mu(x_i = k) - \nu(x_i = k)| \leq e^{-m(i, \partial \alpha)} \frac{1 + (t-1)c}{tc} \frac{1 - \frac{1}{t}}{1 - 1 + \frac{(t-1)c}{t}} = e^{-d(i, \partial \alpha)} \ln \frac{1 + (t-1)c}{tc}.$$

Substituting the inequality of $d(i, \partial \alpha)$ in the condition into right-hand side yields to the result.

References


