TECHNICAL RESEARCH REPORT

Statistical Parameter Learning for Belief Networks with Fixed Structure

by Hongjun Li

CSHCN T.R. 99-32
(ISR T.R. 99-59)

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Statistical parameter learning for Belief networks with fixed structure *

Hongjun Li
Institute for Systems Research and
Center for Satellite and Hybrid Communication Networks
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742

December 2, 1998

Abstract

In this report, we address the problem of parameter learning for belief networks with fixed structure based on empirical observations. Both complete and incomplete (data) observations are included. Given complete data, we describe the simple problem of single parameter learning for intuition and then expand to belief networks under appropriate system decomposition. If the observations are incomplete, we first estimate the "missing" observations and treat them as though they are "real" observations, based on which the parameter learning can be executed as in complete data case. We derive a uniform algorithm based on this idea for incomplete data case and present the convergence and optimality properties. Such algorithm is suitable trivially under complete observations.

*This work was supported by the Center for Satellite and Hybrid Communication Networks, under NASA cooperative agreement NCC3-528.

1 Introduction

As discussed in [1][13], both the network structure and the associated CPTs can be provided by human experts as the prior information. In many applications, however, such information is not available. In addition, different experts may treat the systems in various ways and thus give different and sometimes conflicting assessments. In such cases, the network structure and corresponding CPTs can be estimated using empirical data and we refer to this process as *learning* [6][10][11][12] . Even if such prior information does exist, it is still desirable to validate and improve the model using data.

Learning belief networks consists of both structural learning (deriving the dependency structure G) and parametric learning (estimating P). The structure of the network may be *known* or *unknown*, and the variables in the network may be *observable* or *hidden*. Usually, human experts would rather give for a problem domain the dependence relationships between random variables than the corresponding numerical values, especially when not all of the variables are observable. Therefore we can say that finding the topology of the network is often the relatively easy part and thus the *known structure, hidden variable* learning problem is of great importance.

In this report, we address the problem of parameter learning under fixed structure. Both complete and incomplete (data) observations are included. Given complete data, we describe the simple problem of single parameter learning for intuition and then expand to belief networks under appropriate system decomposition. If the observations are incomplete, we first estimate the "missing" observations and treat them as though they are "real" observations, based on which the parameter learning can be executed as in complete data case. We derive a uniform algorithm based on this idea for incomplete data case and present the convergence and optimality properties. Such algorithm is suitable trivially under complete observations.

Before we proceed to the learning problems, we give the following definitions which are used throughout.

Definition 1.1 *A belief network is a Directed Acyclic Graph (DAG) $G = (X, E, P)$in which: The nodes X represent variables of interest (propositions); The set of directed links or arrows E represent the causal influence among the variables and the parents of a node are all those nodes with arrows pointing to it; The strength of an influence is represented by conditional probabilities attached to each cluster of parent-child nodes in the network.*

We let $X = [X_1, \ldots, X_n]$, with n as the number of nodes in the graph. Each random variable X_i assumes discrete values from finite alphabet A_i, whose cardinality is $|A_i|$. We use capital letters to denote random variables and lower case letters to denote values. For example, $X_i = x_i$ means random variable X_i assumes the value x_i. If we know that X_i assumes its j^{th} value from A_i, we write $X_i = x_i^j$.

Definition 1.2 *An observation is an instantiation $x = [x_1, \ldots, x_n]$ of X, with $x \in R^n$ and assume values in $A \stackrel{\triangle}{=} A_1 \times, \ldots, \times A_n$.*

Definition 1.3 *If every node in X is instantiated (observed), we call x a complete (full) observation, or a complete data set. Each such an x is called a configuration.*

Definition 1.4 *If there are some random variables in X that are not instantiated, such an observation is called an incomplete data.*

Definition 1.5 *The instantiated set of nodes $S \subseteq X$ forms the evidence set, while $N = X \backslash S$ is called the non-evidential set.*

In cases of full observation, $S = X$; but in practice, mostly $S \subset X$, which means incomplete data.

Definition 1.6 *If we mark the nodes in X that belongs to S with $*$ and those in N with ?, then we form the observation schema S^+. Each instantiation of the schema is called an evidence under schema S^+.*

For example, in a 5 node belief network which takes only binary values, such a schema is $S = (*, ?, ?, *, *)$ and one possible instantiation is $(0, ?, ?, 1, 0)^T$.

Definition 1.7 *Suppose we have a batch of observations $D = [D_1, \ldots, D_L]$ with each $D_i \in R^n$ complying with schema S_i^+. If $S_i^+ \equiv S_j^+$, $\forall i, j = 1, \ldots, L$, we say D is a uniform batch of observations. If, on the other hand, S_i^+ may or may not be the same as S_j^+ for $i \neq j$, we say D is a hybrid-schema observation set.*

The rest of this report is organized as follows: We first introduce the problem of parameter learning under complete data in section 2, which includes the simple parameter learning in section 2.1 for intuition and the system decomposition mechanisms used in belief network parameter learning in section 2.2. Then based on this, the incomplete data case is studied in section 3, where the algorithm and corresponding convergence and optimality analysis is stated.

2 Parameter Learning under complete data

In this section, we begin with the simple one-parameter learning case, then we move to multi-variable parameter learning in belief networks.

2.1 Simple parameter learning

Imagine we have a (not necessarily fair) coin, and we conduct an experiment whereby we flip the coin in the air, and it comes to land as either head or tail. We assume that different tosses are

independent and that, in each toss, the probability of the coin landing heads is some underlying unknown real number θ. Our goal here is to estimate θ based on the outcomes of the experiment.

Define the *likelihood* function as $P_\theta(D) = \theta^h(1-\theta)^t$, which is the probability with which we get a particular data set D with h heads and t tails given that the probability θ has a certain value. It is straightforward to verify that the value of θ which maximizes $P_\theta(D)$ is $\frac{h}{h+t}$. This is called *maximum likelihood (ML)* estimate for θ.

In Bayesian analysis, however, we put a distribution over anything about which we have uncertainty. In this case, since we are uncertain about θ, we define a prior distribution $P(\theta)$. Then we have a joint distribution of both the tosses and the parameter θ:

$$\begin{aligned} P(x_1, \cdots, x_n, \theta) &= P(x_1, \cdots, x_n|\theta)P(\theta) \\ &= P(\theta)\theta^h(1-\theta)^t, \end{aligned} \tag{1}$$

where $P(x_1, \cdots, x_n|\theta) = P_\theta(D)$ is just the likelihood function. Now as we see more data, the *posteriori* distribution over the parameter changes. In particular, using Bayes rule, we have

$$P(\theta|D) = \frac{P(\theta)P_\theta(D)}{P(D)} = \frac{P(\theta)P_\theta(D)}{\int_0^1 P(\theta)P_\theta(D)d\theta}. \tag{2}$$

An appropriate such distribution for a parameter is the *Beta* distribution. A Beta function distribution is parameterized by two numbers α_h, α_t. Intuitively, these correspond to the number of imaginary heads and tails that have been seen. The Beta function function with these parameters has the following form,

$$P(\theta) = Beta(\theta|\alpha_h, \alpha_t) \propto \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1} \tag{3}$$

and some of the Beta distributions are shown in figure 1.

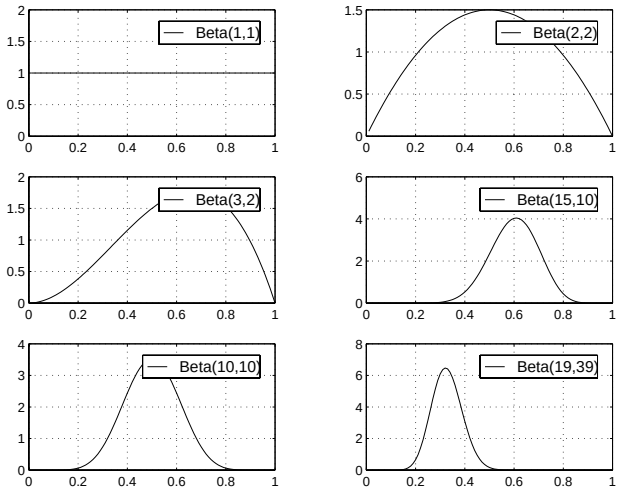

Figure 1: An Illustration of some Beta Distributions

Beta distributions have properties that make them particularly useful for parameter estimation.

- *First*, for $Beta(\theta|\alpha_h, \alpha_t)$, the probability of next coin toss coming out head is:

$$\begin{aligned} P(X = h|\theta) &= \int_0^1 P(\theta)P(X = h|\theta)d\theta \\ &= \int_0^1 P(\theta)\theta d\theta \\ &= \frac{\alpha_h}{\alpha_h + \alpha_t}. \end{aligned} \tag{4}$$

  which is just the result of ML estimate. This supports our intuition that the Beta distribution corresponds to having seen α_h heads and α_t tails (imaginary or real).

- *Second* , as we get more data, specifically we get a data set D with h heads and t tails, we have

$$P(\theta|D) = Beta(\theta|\alpha_h + h, \alpha_t + t). \tag{5}$$

  This property that the *posteriori* distribution belongs to the same family as the *prior* distribution is called *conjugacy* under the observation data. Such priors are called *conjugate priors.* Using conjugate priors will allow us focus on the hyper-parameters α_h, α_t and the observations only, without having to worry about the form of *posteriori* distribution and the usually non-trivial integration steps.

Extending multinomial (for example m-ary) case, such a conjugate prior is Dirichlet distribution:

$$P(\theta) = Dirichlet(\theta|\alpha_1, \ldots, \alpha_m), \tag{6}$$

and based on a multinomial sampling distribution $P_\theta(D)$, the *posteriori* distribution is

$$P(\theta|D) = Dirichlet(\theta|\alpha_1 + N_1, \ldots, \alpha_m + N_m), \tag{7}$$

where $N_i, 1 \leq i \leq m$ is the number of times that the outcome is i, given the data set D.

Staring from a rather uniform ("flat") distribution, the Beta distribution will become more and more focused around some certain value. The use of an entire distribution rather than a single number to model the parameter θ has the advantage that if reflects not only our current estimate for the value of θ, but also our degree of confidence about that.

Note that: 1. ML estimate is the expectation of Bayesian estimate; 2. ML estimate approaches to Bayesian estimate when sample size becomes unboundedly large (which means prior belief becomes less and less important as the data accumulate); 3. Given ML estimate, it is straightforward to obtain the hyper-parameters for Dirichlet distribution. So in this report, we focus on ML estimate.

The usefulness of the coin flipping here lies in the fact that: 1. It is simple and intuitive; 2. If, by some decomposition mechanism, we can break a complex problem into multiple independent single parameter learning problems, we can use the flip coin techniques here for each of them; And, such computations would be possibly done in a distributed manner, as we will see next.

2.2 Parameter learning for a belief network

2.2.1 System decomposition

For a discrete-valued belief network $G = (V, E, P)$ with fixed structure, the joint probability distribution (JPD) can be represented as

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i|\Pi_i). \tag{8}$$

If we order the nodes in such a way that the order of a node is larger than those of its parents and smaller than those of its children (the so-called *topological ordering*), we have

$$p(x_i|x_1, ..., x_{i-1}) = p(x_i|\Pi_i), \tag{9}$$

which means given its parent set $\Pi_i \subseteq \{x_1, ..., x_{i-1}\}$, the set of variables that render x_i, each variable x_i is conditionally independent of all its other predecessors $\{x_1, ..., x_{i-1}\}\backslash\Pi_i$.

Suppose data $D = [D_1, \ldots, D_L]$ are generated independently from some underlying distribution, with each $D_i = [x_1[i], \ldots, x_n[i]]^T$. The problem here is to find the CPT parameters θ that best model the data. The parameter θ is actually a 3-dimensional matrix, with its element θ_{ijk} defined as the probability that variable X_i takes on its j^{th} possible value assignment given its parents Π_i takes on their k^{th} possible value assignment, or

$$\theta_{ijk} = P(X_i = x_i^j|\Pi_i = \pi_i^k). \tag{10}$$

We assume in this report that $\theta_{ijk} > 0$, $\forall\, i, j, k$, and $P_\theta(D)$ continuous with θ. If we define $L(\theta; D) = P_\theta(D)$ as the likelihood function given some parameter θ, the Maximum Log-Likelihood formulation for this problem is:

$$\begin{cases} \max_\theta \log L(\theta; D) \\ s.t.\ \sum_j \theta_{ijk} = 1, and \\ \theta_{ijk} \in [0, 1] \end{cases} \tag{11}$$

Moreover, we have:

$$\begin{aligned} L(\theta; D) &= \prod_{l=1}^{L} P_\theta(x_1[l], \ldots, x_n[l]) \\ &= \prod_{l=1}^{L}\prod_{i=1}^{n} P_{\theta_i}(x_i[l]|\Pi_i[l]) \\ &= \prod_{i=1}^{n}\prod_{l=1}^{L} P_{\theta_i}(x_i[l]|\Pi_i[l]) \\ &= \prod_{i=1}^{n} L_i(\theta_i; D), \end{aligned} \tag{12}$$

where $L_i(\theta_i; D) \stackrel{\triangle}{=} \prod_{l=1}^{L} P_{\theta_i}(x_i[l]|\Pi_i[l])$. The first equality comes from the fact that D consists of independent observations and the second equality follows (8). $\theta_{\mathbf{i}}$ is a 2-dimensional matrix, with rows occupied by its possible values and columns defined according to its parents' status. From (12) we get

$$\log L(\theta; D) = \sum_{i=1}^{n} \log L_i(\theta_i; D). \tag{13}$$

The above derivation gives us a decomposition of the belief networks learning problem based on independent observations. Namely, the maximum likelihood solutions of (11) are just those achieved by the sum of the solutions from the following independent estimation problems:

For each $X_i \in X$:

$$\begin{cases} \max_{\theta_i} \log L_i(\theta_i; D) \\ s.t. \ \sum_j \theta_{ijk} = 1, and \\ \theta_{ijk} \in [0, 1] \end{cases} \tag{14}$$

For each $L_i(\theta_i; D)$, we make further decomposition as follows.

Definition 2.1 *The set of nodes $S_i = (X_i, \Pi_i)$ forms the extracted schema for node X_i. X_i assumes values from A_i. Π_i takes values from $\prod_{X_j \in \Pi_i} A_j$. Let $K = |\prod_{X_j \in \Pi_i} A_j|$.*

There are all together L samples, for which we define the groups as:

Definition 2.2 *Group G_{ik} is the set of instantiations of S_i in D with $\Pi_i = \pi_i^k$. The number of elements in G_{ik} is $N_{ik} = |G_{ik}| = \sum_{l=1}^{L} I_{\{\pi_i^k|D_l\}}$, where $I_{\{\pi_i^k|D_l\}}$ is the indicator function defined as:*

$$I_{\{z|D_l\}} = \begin{cases} 1 & if \ z \ occurs \ in \ D_l \\ 0 & otherwise \end{cases} \tag{15}$$

and $\sum_k N_{ik} = L$.

Further, we can sub-divide G_{ik} according to the value of X_i:

Definition 2.3 *In G_{ik}, G_{ijk} is the set of instantiations where $X_i = x_i^j$, and $N_{ijk} = |G_{ijk}| = \sum_{l=1}^{L} I_{\{x_i^j, \pi_i^k|D_l\}}$, number of elements in group G_{ijk} with $\sum_{j=1}^{A_i} N_{ijk} = N_{ik}$.*

So the decomposition is:

$$L_i(\theta_i; D) \quad = \quad \prod_{l=1}^{L} P_{\theta_i}(x_i[l]|\Pi_i[l])$$

$$= \prod_{k=1,\{\Pi_i=\pi_i^k\}}^{K} \prod_{j=1,\{X_i=x_i^j|\Pi_i=\pi_i^k\}}^{|A_i|} \theta_{ijk}^{N_{ijk}}$$

$$= \prod_{k=1,\{\Pi_i=\pi_i^k\}}^{K} L_i^k(\theta_i; D), \tag{16}$$

or

$$\log L_i(\theta_i; D) = \sum_{k=1,\{\Pi_i=\pi_i^k\}}^{K} \log L_i^k(\theta_i; D), \tag{17}$$

where $L_i^k(\theta_i; D) = \prod_{j=1,\{X_i=x_i^j|\Pi_i=\pi_i^k\}}^{|A_i|} \theta_{ijk}^{N_{ijk}}$ is just the likelihood function of multinomial distribution under parameter θ.

Then the ML solutions of (14) are again those achieved by the sum of the solutions from the following independent estimation problems:

For each $\Pi_i = \pi_i^k$,

$$\begin{cases} \max_{\theta_i} \log L_i^k(\theta_i; D) \\ s.t. \ \sum_j \theta_{ijk} = 1, and \\ \theta_{ijk} \in [0, 1] \end{cases} \tag{18}$$

Or equivalently, by defining $r_j = \theta_{ijk}$, $y_j = N_{ijk}$, and $\tilde{L}(r; D) = \prod_{j=1}^{|A_i|} r_j^{y_j}$, we consider the maximum log-likelihood problem,

For each $\Pi_i = \pi_i^k$,

$$\begin{cases} \max_r \log \tilde{L}(r; D) \\ s.t. \ \sum_j r_j = 1, and \\ r_j \in [0, 1] \end{cases} \tag{19}$$

which is just a generalization of the simple parameter learning problem discussed in section 2.1.

The decomposition $\log L(\theta; D) = \sum_{i=1}^{n} \sum_{k=1,\{\Pi_i=\pi_i^k\}}^{K} \log L_i^k(\theta_i; D)$ first exploits the conditional independence structure embedded in belief networks and helps to reduce the problem to n independent ML estimate problems; then, each such problem is further decomposed as shown in (16). We can thus do the learning in a "local" and distributed manner.

2.2.2 Parameter estimation

In this section, we derive the ML estimates for problem (19) and provide the optimality results. For some concepts of estimation theory, we refer to [23].

Lemma 2.4 *Given complete data set D, $\log \tilde{L}(r; D)$ is negative and strictly concave in $r = [r_1, \ldots, r_{|A_i|}]$.*

Proof. Without loss of generality, $\tilde{L}(r; D)$ can be rewritten as

$$\tilde{L}(r; D) = (\prod_{j=1}^{|A_i|-1} r_j^{y_j})(1 - \sum_{j=1}^{|A_i|-1})^{y_{|A_i|}}. \tag{20}$$

Define $J = |A_i|$ for simplicity, and define function $f : R^J \to R^1$ as

$$\begin{aligned} f(r) &= \log \tilde{L}(r; D) \\ &= \sum_{j=1}^{J-1} y_j \log r_j + y_J \log(1 - \sum_{j=1}^{J-1} r_j). \end{aligned} \tag{21}$$

Obviously, $\tilde{L}(r; D) < \sum_{y \in \mathcal{Y}} \tilde{L}(r; y) = 1$, where $\mathcal{Y}$ is the set of all possible combinations of D, so $\log \tilde{L}(r; D) < 0$.

Let $r = \lambda\alpha + (1 - \lambda)\beta$, where α and β satisfy the constraints in (19) and $0 < \lambda < 1$. Since $\log(x)$ is strictly concave in x, we have

$$\sum_{j=1}^{J-1} y_j \log(\lambda\alpha_j + (1 - \lambda)\beta_j) > \lambda \sum_{j=1}^{J-1} y_j \log \alpha_j + (1 - \lambda) \sum_{j=1}^{J-1} y_j \log \beta_j. \tag{22}$$

Also note that

$$1 - \sum_{j=1}^{J-1} (\lambda\alpha_j + (1 - \lambda)\beta_j) = \lambda(1 - \sum_{j=1}^{J-1} \alpha_j) + (1 - \lambda)(1 - \sum_{j=1}^{J-1} \beta_j). \tag{23}$$

From (22) and (23), we have

$$\begin{aligned} f(\lambda\alpha + (1 - \lambda)\beta) &> \lambda \sum_{j=1}^{J-1} y_j \log \alpha_j + (1 - \lambda) \sum_{j=1}^{J-1} y_j \log \beta_j \\ &\quad + \lambda y_J \log(1 - \sum_{j=1}^{J-1} \alpha_j) + (1 - \lambda) y_J \log(1 - \sum_{j=1}^{J-1} \beta_j) \\ &= \lambda f(\alpha) + (1 - \alpha) f(\beta). \end{aligned} \tag{24}$$

$\Rightarrow$ Strictly concave. □

Lemma 2.5 *Given complete data set D, $\log L(\theta; D)$ is negative and strictly concave in $\theta = \{\theta_{ijk}\}$.*

Proof. From last section we know that

$$\begin{aligned} L(\theta; D) &= \prod_{i=1}^{n} \prod_{k=1,\{\Pi_i=\pi_i^k\}}^{K} L_i^k(\theta_i; D) \quad \Rightarrow \\ \log L(\theta; D) &= \sum_{i=1}^{n} \sum_{k=1,\{\Pi_i=\pi_i^k\}}^{K} \log L_i^k(\theta_i; D), \end{aligned} \tag{25}$$

where by lemma 2.4, each $\log L_i^k(\theta_i; D)$ is negative and strictly concave in θ_{ijk}. We conclude that the sum of those negative (and thus non-cancelling), strictly concave functions is also negative and strictly concave in θ_{ijk}, $\forall\, i, j, k$, or, concave in θ. □

Lemma 2.6 *The maximum likelihood solution for problem (19) is*

$$r_j = y_j / \sum_{j=1}^{J} y_j, \tag{26}$$

where $y_j = N_{ijk}$, as defined in definition 2.3.

Proof. Use the f notation as above and obtain the Likelihood Equations

$$\frac{\partial f}{\partial r_j} = 0, \quad j = 1, \ldots, J. \tag{27}$$

From (21) it is straightforward to get $J-1$ independent equations:

$$\frac{y_j}{r_j} = \frac{y_J}{1 - \sum_{i=1}^{J-1} r_i}, \quad \forall j = 1, \ldots, J-1, \tag{28}$$

from which it is easy to obtain that $r_j = y_j / \sum_{j=1}^{J} y_j$, and we can check that $r_j \in [0, 1]$ and $\sum_j^J r_j = 1$.

By lemma 2.4, we know such stationary points are global maxima (because $\log \tilde{L}(r; D)$ is **strictly** concave). □

Lemma 2.7 *The ML estimates for problem (19) are minimum variance unbiased estimators (MVUE).*

Proof. Given data set D, suppose the number of occurrence of X_i's J possible values are $[y_1, \ldots, y_J]^T$, if we look at node X_i under $\Pi_i = \pi_i^k$. The multinomial distribution is

$$p_\psi(y) = \frac{(\sum_{i=1}^{J} y_i)!}{\prod_{i=1}^{J} y_i!} \prod_{i=1}^{J} \psi_i^{y_i} \tag{29}$$

Then for an underlying set of parameters ψ, we have for estimator $\hat{\psi}_j(y) = y_j / \sum_{i=1}^J y_i$ the expectation

$$\begin{aligned}
E_\psi\{\hat{\psi}_j(y)\} &= E_\psi\{y_j/N\} \qquad (N \triangleq \sum_{i=1}^J y_i) \\
&= \sum_{y_j=0}^N \frac{y_j}{N} \frac{N!}{y_j!} \psi_j^{y_j} \sum_{\substack{y_i \\ (\sum_{i\neq j} y_i) = N - y_j}} \frac{1}{\prod_{i\neq j} y_i!} \prod_{i\neq j} \psi_i^{y_i} \\
&= \sum_{y_j=0}^N \frac{y_j}{N} \frac{N!}{y_j!(N-y_j)!} \psi_j^{y_j} (1-\psi_j)^{N-y_j} \underbrace{\sum_{\substack{y_i \\ (\sum_{i\neq j} y_i) = N - y_j}} \frac{(N-y_j)!}{\prod_{i\neq j} y_i!} \prod_{i\neq j} (\frac{\psi_i}{1-\psi_j})^{y_i}}_{1} \\
&= \frac{1}{N} \sum_{y_j=0}^N y_j \binom{N}{y_j} \psi_j^{y_j} (1-\psi_j)^{N-y_j} \\
&= \frac{1}{N} N\psi_j = \psi_j. \qquad (30)
\end{aligned}$$

$\Rightarrow$ unbiased.

From (29), we have

$$\frac{\partial \log p_\psi(y)}{\partial \psi_j} = \frac{y_j}{\psi_j} - \frac{N-y_j}{1-\psi_j}. \qquad (31)$$

Then the Fisher information is,

$$\begin{aligned}
\mathcal{I}_\psi &= -E_\psi\{\frac{\partial^2 \log p_\psi(y)}{\partial^2 \psi_j}\} \\
&= E_\psi\{\frac{y_j}{\psi_j^2} + \frac{N-y_j}{(1-\psi_j)^2}\} \\
&= N/\psi_j(1-\psi_j), \qquad (32)
\end{aligned}$$

while the variance is

$$\begin{aligned}
Var_\psi\{y_j/N\} &= \frac{1}{N^2} E_\psi\{y_j^2\} - E^2\{y_j/N\} \\
&= \frac{1}{N^2}\{N\psi_j(1-\psi_j) + (N\psi_j)^2\} - \psi_j^2 \\
&= \psi_j(1-\psi_j)/N = 1/\mathcal{I}_\psi. \qquad (33)
\end{aligned}$$

$\Rightarrow$ MVUE (achieves Cramer Rao Lower Bound, CRLB). □

3 Parameter learning under incomplete data using EM algorithm

In this section, we first brief the idea of EM algorithm, followed by the derivation for belief networks, and then in section 3.3, we discuss the convergence and optimality properties.

3.1 A brief description of EM algorithm

EM algorithm is broadly applicable for computing maximum likelihood estimates from incomplete data. Each iteration consists of an expectation step followed by a maximization step, and hence the name [8][17].

Suppose we have two sample spaces $\mathcal{X}$ and $\mathcal{Y}$ with many-to-one mapping $\mathcal{X} \rightarrow \mathcal{Y}$, where $\mathbf{x}$ can not be observed directly, but instead, only through $\mathbf{y}$. Let the family of sampling densities depending on Φ for $\mathcal{X}$ and $\mathcal{Y}$ are $f_\Phi(\mathbf{x})$ and $g_\Phi(\mathbf{y})$, respectively. We call $f_\Phi(\mathbf{x})$ as the complete data specification and $g_\Phi(\mathbf{y})$ as the incomplete data specification, with the following relation:

$$g_\Phi(\mathbf{y}) = \int_{\mathcal{X}(\mathbf{y})} \mathbf{f_\Phi}(\mathbf{x})\mathbf{dx}, \tag{34}$$

where $\mathcal{X}(\mathbf{y})$ is the set of $\mathbf{x} \in \mathcal{X}$ that corresponds to $\mathbf{y}$. EM algorithm aims to find a value of Φ that maximizes $g_\Phi(\mathbf{y})$ given an observed $\mathbf{y}$, but does so by making essential use of $f_\Phi(\mathbf{x})$. When $f_\Phi(\mathbf{x})$ belongs to an exponential family, we have the EM algorithm:

Let $\Phi^{(p)}$ denotes the current value of Φ after p cycles, then for the next cycle:

E-step: Estimate the complete-data sufficient statistics $t(\mathbf{x})$ by

$$t^{(p)} = E_{\Phi^{(p)}}\{t(\mathbf{x})|\mathbf{y}\} \tag{35}$$

M-step: Determine $\Phi^{(p+1)}$ as the solution of the equations

$$E_\Phi\{t(\mathbf{x})\} = \mathbf{t^{(p)}} \tag{36}$$

3.2 EM algorithm derivation

Suppose we have a batch of observations $D = [D_1, \ldots, D_L]$ with each $D_i \in R^n$ complying with schema S_i^+. D may or may not be uniform. The objective is to find the most likely underlying parameter θ that can best model the incomplete observations, namely

$$\begin{cases} \max_\theta \log P_\theta(D) \\ s.t.\ \sum_j \theta_{ijk} = 1, and \\ \theta_{ijk} \in [0,1] \end{cases} \tag{37}$$

and we wish to do this via the maximization of an associated complete data problem. The idea is that we first estimate and "fill" the missing values based on the evidence and current guess of

the parameters, after which we treat them as the real data and apply the ML principle to do the parameter learning. The estimation of the missing values are called Expectation-step (or E-step) and the parameter learning step is called Maximization-step (or M-step). It is straightforward to see that a multinomial distribution $p_\psi(y) = \frac{(\sum_{j=1}^J y_j)!}{\prod_{j=1}^J y_j!} \prod_{j=1}^J \psi_j^{y_j}$ belongs to the exponential family and the sufficient statistics is just the set $\{y_j\}$, $j = 1, \ldots, J$.

E-step: For each sample D_l, we want to estimate the values for those nodes corresponding to "?" mark in schema S_l^+ and thus get the augmented data set $C(D_l)$. Let N_l be the set of nodes marked as "?" in schema S_l^+, then given D_l, there are all together $\prod_{m\in N_l} |A_m|$ cases in the augmented data set $C(D_l)$. For any entry $D_l^+(q) \in C(D_l)$, $\quad q = 1, \ldots, \prod_{m\in N_l} |A_m|$, the evidential nodes are "clamped" as in observation D_l, while the non-evidential nodes take the q^{th} combination from the $\prod_{m\in N_l} |A_m|$ choices. Then under the current guess $\tilde{\theta}$ of parameter θ, $P_{\tilde{\theta}}(D_l^+(q))$ can be computed using (8), and the probability of $D_l^+(q)$ given D_l is

$$P_{\tilde{\theta}}(D_l^+(q)|D_l) = \frac{P_{\tilde{\theta}}(D_l^+(q))}{\sum_q P_{\tilde{\theta}}(D_l^+(q))}. \tag{38}$$

M-step: For augmented data set $C(D_l)$ where each entry is a complete observation, we can either average over those entries or find the most-probable entry to serve as the complete data set for the ML estimator. At this time, we consider the averaging method, by which we weigh each $D_l^+(q)$ according to (38) within each D_l.

So the associated problem is: for complete data set $C(D) = [C(D_1), \ldots, C(D_L)]$,

$$\begin{cases} \max_\theta \log \bar{P}_\theta(C(D)) \\ s.t. \ \sum_j \theta_{ijk} = 1, and \\ \quad \theta_{ijk} \in [0,1] \end{cases} \tag{39}$$

where

$$\begin{aligned} \log \bar{P}_\theta(C(D)) &= \sum_{l=1}^L \log \bar{P}_\theta(C(D_l)) \\ &= \sum_{l=1}^L \sum_{q=1}^{\prod_{m\in N_l} |A_m|} P_{\tilde{\theta}}(D_l^+(q)|D_l) \log P_\theta(D_l^+(q)) \\ &= \sum_{i=1}^n \sum_{l=1}^L \sum_{q=1}^{\prod_{m\in N_l} |A_m|} P_{\tilde{\theta}}(D_l^+(q)|D_l) \log P_{\theta_i}(X_i(D_l^+(q))|\Pi_i(D_l^+(q))). \end{aligned} \tag{40}$$

Compare (40) with (13), we can take similar decompositions and by lemma 2.6 we have for node X_i under $\Pi_i = \pi_i^k$,

$$\hat{\theta}_{ijk} = \frac{\tilde{N}_{ijk}}{\tilde{N}_{ik}}, \tag{41}$$

where $\tilde{N}_{ijk}$ and $\tilde{N}_{ik}$ can be obtained as in definition 2.2 and 2.3, except that the complete data $D_l^+(q)$ is weighed by $P_{\tilde{\theta}}(D_l^+(q)|D_l)$. So

$$\hat{\theta}_{ijk} = \frac{\sum_{l=1}^{L} \sum_{q=1}^{\prod_{m \in N_l} |A_m|} I_{\{x_i^j, \pi_i^k | D_l^+(q)\}} P_{\tilde{\theta}}(D_l^+(q)|D_l)}{\sum_{l=1}^{L} \sum_{q=1}^{\prod_{m \in N_l} |A_m|} I_{\{\pi_i^k | D_l^+(q)\}} P_{\tilde{\theta}}(D_l^+(q)|D_l)}, \tag{42}$$

where $\tilde{\theta}$ is the current set of parameters and $I_{\{z|D_l^+(q)\}}$ is the indicator function defined as

$$I_{\{z|D_l^+(q)\}} = \begin{cases} 1 & \text{if } z \text{ occurs in } D_l^+(q) \\ 0 & \text{otherwise} \end{cases} \tag{43}$$

One may easily observe that such augmentation is of combinatorial complexity, and even when $A_i = \{0, 1\}, \forall i$, the entries for $C(D_l)$ would be $2^{|N_l|}$, which makes the computation in (42) intractable in most cases.

However, notice that

$$P_{\tilde{\theta}}(x_i^j, \pi_i^k | D_l) = \sum_{q=1}^{\prod_{m \in N_l} |A_m|} I_{\{x_i^j, \pi_i^k | D_l^+(q)\}} P_{\tilde{\theta}}(D_l^+(q)|D_l), \tag{44}$$

and

$$P_{\tilde{\theta}}(\pi_i^k | D_l) = \sum_{q=1}^{\prod_{m \in N_l} |A_m|} I_{\{\pi_i^k | D_l^+(q)\}} P_{\tilde{\theta}}(D_l^+(q)|D_l), \tag{45}$$

we can simplify (42) as

$$\hat{\theta}_{ijk} = \frac{\sum_{l=1}^{L} P_{\tilde{\theta}}(x_i^j, \pi_i^k | D_l)}{\sum_{l=1}^{L} P_{\tilde{\theta}}(\pi_i^k | D_l)}, \tag{46}$$

where $P_{\tilde{\theta}}(x_i^j, \pi_i^k | D_l)$ and $P_{\tilde{\theta}}(\pi_i^k | D_l)$ can be calculated using standard inference algorithm given D [5][20][21]. So we don't need to do the augmentation explicitly and hence avoid the combinatorial complexity. Compare (46) with (26), we can see that we just replace the "hard" counting measures $y_j, \forall j \in J$ in (26) with the "soft" estimation $P_{\tilde{\theta}}(x_i^j, \pi_i^k | D_l)$ for $y_j, \forall j \in J$. By lemma 2.7, (46) is the MVUE estimator for the complete data augmented using $\tilde{\theta}$.

Now we got the ML estimates for problem (39). Recall that our goal is to find the ML estimates that best model the incomplete data set D, we treat the estimates $\hat{\theta}$ as the current guess of true parameter θ and do the E-step and M-step again. Repeat such process and we get the EM algorithm for discrete-valued belief network. If we define operator $H(\psi)$ as

$$H(\psi)(ijk) = \frac{\sum_{l=1}^{L} P_{\psi}(x_i^j, \pi_i^k | D_l)}{\sum_{l=1}^{L} P_{\psi}(\pi_i^k | D_l)}, \tag{47}$$

then the EM process can be summarized as the iteration

$$\tilde{\theta}^{(p+1)} = H(\tilde{\theta}^{(p)}), \tag{48}$$

for some initial $\tilde{\theta}^{(0)}$. The EM algorithm can be thought of finding the *fixed point* of operator H, and such fixed point $\tilde{\theta}^*$ is just the best set of parameters that model the data set D.

More generally, we can extend (48) to the small-step size version of iteration, which falls within the stochastic approximation framework [2][9][14][25], as shown below:

$$\tilde{\theta}^{(p+1)} = (1-\gamma_p)\tilde{\theta}^{(p)} + \gamma_p H(\tilde{\theta}^{(p)}), \tag{49}$$

where $\gamma_p \in [0,1]$. Obviously, if $\gamma_p = 0$, $\tilde{\theta}^{(p+1)} = \tilde{\theta}^{(p)}$ and we ignores the influence from data; if $\gamma_p = 1$, however, we get (48) as a special case. We assume from now on the nontrivial case where $\gamma_p > 0$.

If we define operator $M(\psi)$ as

$$M(\psi) = (1-\gamma)\psi + \gamma H(\psi), \tag{50}$$

with ψ appropriately chosen, then the EM algorithm can be summarized as the iteration

$$\tilde{\theta}^{(p+1)} = M(\tilde{\theta}^{(p)}), \tag{51}$$

and the goal here is to find the fixed point $\tilde{\theta}^*$ of operator M.

3.3 Optimality and Convergence

Lemma 3.1 *For one sample $D = [D_1]$, the algorithm $\tilde{\theta}^{(p+1)} = H(\tilde{\theta}^{(p)})$ makes $\log L(\theta; D)$ non-decrease for each iteration.*

Proof. The following proof resembles that in [17]. Let Y denote the observed nodes and Z denote the non-observed nodes. So $\log L(\theta; D) = \log L(\theta; Y = y)$, and $X = [Y, Z]$. Given those nodes in Y "clamped" as indicated in D_1, we can obtain by using $\tilde{\theta}^{(p)}$ the augmented complete data set $X = X(y)$, where $X(y)$ denotes the multiple cases of X that contains $Y = y$. Such $X(y)$ may be governed by some unknown parameter set θ, which we want to estimate using ML principle at M-step. For notation simplicity, we use Y to denote the fact that $Y = y$.

Since

$$P_\theta(X(y)|Y) = P_\theta(X(y), Y)/P_\theta(Y) = P_\theta(X(y))/P_\theta(Y), \tag{52}$$

we have

$$\log L(\theta; Y) = \log P_\theta(Y) = \log P_\theta(X(y)) - \log P_\theta(X(y)|Y). \tag{53}$$

Take expectation with respect to the conditional probability density of X given Y under current parameter set $\tilde{\theta}^{(p)}$, we get

$$\begin{aligned} E_{\tilde{\theta}^{(p)}}\{\log P_\theta(Y)\} &= \log P_\theta(Y) \\ &= E_{\tilde{\theta}^{(p)}}\{\log P_\theta(X(y))\} - E_{\tilde{\theta}^{(p)}}\{\log P_\theta(X(y)|Y)\} \\ &= Q(\theta, \tilde{\theta}^{(p)}) - T(\theta, \tilde{\theta}^{(p)}), \end{aligned} \tag{54}$$

where $Q(\theta, \tilde{\theta}^{(p)}) = E_{\tilde{\theta}^{(p)}}\{\log P_\theta(X(y))\}$, $T(\theta, \tilde{\theta}^{(p)}) = E_{\tilde{\theta}^{(p)}}\{\log P_\theta(X(y)|Y)\}$.

At M-step, as described in section 3.2, we use the ML principle to find the most probable parameter based on the complete data set $X(y)$, each entry of which is appropriately weighed according to its conditional density under $\tilde{\theta}^{(p)}$. So,

$$\begin{aligned}\tilde{\theta}^{(p+1)} &= \arg\max_{\theta} Q(\theta, \tilde{\theta}^{(p)}) \\ &= \arg\max_{\theta} \sum_{x \in X(y)} p_{\tilde{\theta}^{(p)}}(x|y) \log P_{\theta}(x). \end{aligned} \tag{55}$$

and therefore,

$$Q(\tilde{\theta}^{(p+1)}, \tilde{\theta}^{(p)}) \geq Q(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}). \tag{56}$$

Also for any θ,

$$\begin{aligned} T(\theta, \tilde{\theta}^{(p)}) - T(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}) &= E_{\tilde{\theta}^{(p)}}\{\log \frac{P_{\theta}(X|Y)}{P_{\tilde{\theta}^{(p)}}(X|Y)}\} \\ &= -D(\tilde{\theta}^{(p)}||\theta) \\ &\leq 0, \end{aligned} \tag{57}$$

where $D(\tilde{\theta}^{(p)}||\theta) \geq 0$ is the relative entropy between $P_{\tilde{\theta}^{(p)}}(X|Y)$ and $P_{\theta}(X|Y)$, see [7]. By (56) and (57), we conclude that

$$\begin{aligned} \log L(\tilde{\theta}^{(p+1)}; D) - \log L(\tilde{\theta}^{(p)}; D) &= Q(\tilde{\theta}^{(p+1)}, \tilde{\theta}^{(p)}) - Q(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}) \\ &\quad -[T(\tilde{\theta}^{(p+1)}, \tilde{\theta}^{(p)}) - T(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)})] \\ &\geq 0, \end{aligned} \tag{58}$$

or $\log L(\tilde{\theta}^{(p+1)}; D) \geq \log L(\tilde{\theta}^{(p)}; D)$. □

Lemma 3.2 *For independent uniform $D = [D_1, \ldots, D_L]$ with each $D_i \in R^n$ complying with the same schema S^+ ,the algorithm $\tilde{\theta}^{(p+1)} = H(\tilde{\theta}^{(p)})$ makes $\log L(\theta; D)$ non-decrease for each iteration.*

Proof. For uniform data D, the actual observed values of the corresponding evidential nodes may differ between different observations. At E-step, we can obtain for each D_l the augmentation and thus the complete data set $C(D)$ under $\tilde{\theta}^{(p)}$. For notation simplicity, we let $D^+ = C(D)$.

Observing that $D_1, \ldots, D_L$ are independent observations and also, each D_l^+ is obtained independently of D_m, $m \neq l$, we have

$$P_{\theta}(D_l) = P_{\theta}(D_l^+)/P_{\theta}(D_l^+|D_l), \tag{59}$$

and

$$\log P_{\theta}(D_l) = \log P_{\theta}(D_l^+) - \log P_{\theta}(D_l^+|D_l). \tag{60}$$

Take expectation with respect to the conditional probability density of D_l^+ given D_l under $\tilde{\theta}^{(p)}$, we get

$$\begin{aligned} \log P_{\theta}(D_l) &= E_{\tilde{\theta}^{(p)}}\{\log P_{\theta}(D_l^+)\} - E_{\tilde{\theta}^{(p)}}\{\log P_{\theta}(D_l^+|D_l)\} \\ &= Q_l(\theta, \tilde{\theta}^{(p)}) - T_l(\theta, \tilde{\theta}^{(p)}), \end{aligned} \tag{61}$$

where $Q_l(\theta, \tilde{\theta}^{(p)}) = E_{\tilde{\theta}^{(p)}}\{\log P_\theta(D_l^+)\}$ and $T_l(\theta, \tilde{\theta}^{(p)}) = E_{\tilde{\theta}^{(p)}}\{\log P_\theta(D_l^+|D_l)\}$.

Then for each D_l under $\tilde{\theta}^{(p)}$, we have

$$\begin{aligned} \log L(\theta; D) &= \sum_l \log P_\theta(D_l) \\ &= \sum_l Q_l(\theta, \tilde{\theta}^{(p)}) - \sum_l T_l(\theta, \tilde{\theta}^{(p)}), \end{aligned} \tag{62}$$

where

$$\begin{aligned} Q_l(\theta, \tilde{\theta}^{(p)}) &= E_{\tilde{\theta}^{(p)}}\{\log P_\theta(D_l^+)\} \\ &= \sum_q P_{\tilde{\theta}^{(p)}}(D_l^+(q)|D_l) \log P_\theta(D_l^+(q)). \end{aligned} \tag{63}$$

Like in Lemma 3.1, $T_l(\theta, \tilde{\theta}^{(p)}) \leq T_l(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)})$, for any $\theta \neq \tilde{\theta}^{(p)}$, so we have

$$\sum_l T_l(\theta, \tilde{\theta}^{(p)}) \leq \sum_l T_l(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}). \tag{64}$$

At M-step, we find the ML estimate of θ based on D^+ as a whole, each entry of which is weighed appropriately. The algorithm is $\tilde{\theta}^{(p+1)} = \arg\max_\theta E_{\tilde{\theta}^{(p)}}\{\log P_\theta(D^+)\}$, where

$$\begin{aligned} E_{\tilde{\theta}^{(p)}}\{\log P_\theta(D^+)\} &= \sum_l \sum_q P_{\tilde{\theta}^{(p)}}(D_l^+(q)|D_l) \log P_\theta(D_l^+(q)) \\ &= \sum_l Q_l(\theta, \tilde{\theta}^{(p)}). \end{aligned} \tag{65}$$

Thus we can conclude that

$$\sum_l Q_l(\tilde{\theta}^{(p+1)}, \tilde{\theta}^{(p)}) \geq \sum_l Q_l(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}). \tag{66}$$

Combine (62), (64) and (66) we see that $\log L(\tilde{\theta}^{(p+1)}; D) \geq \log L(\tilde{\theta}^{(p)}; D)$. □

Lemma 3.3 *For independent nonuniform $D = [D_1, \ldots, D_L]$ with each $D_i \in R^n$ complying with schema S_i^+ ,the algorithm $\tilde{\theta}^{(p+1)} = H(\tilde{\theta}^{(p)})$ makes $\log L(\theta; D)$ non-decrease for each iteration.*

Proof. Let D_l^+ be the complete data set augmented from D_l using $\tilde{\theta}^{(p)}$ and under observation schema S_l^+. Such D_l^+ is independent of D_m, $\forall m \neq l$ and let the underlying parameter is θ, as usual. Unlike with uniform data, the augmentation here differs not only in the observed values, but also in the observation schemas. Once we finish the augmentation, however, we can average the entries in D^+ as before. So in this sense we see that the only difference between nonuniform and uniform data cases lies in the way of how the augmentation is achieved. Note that for formulae (59)–(66), we didn't use the assumption of uniform data, the proof above can be adapted here and we get $\log L(\tilde{\theta}^{(p+1)}; D) \geq \log L(\tilde{\theta}^{(p)}; D)$. □

From Lemma 3.1, 3.2 and 3.3 we can immediately have the following result.

Proposition 3.1 *Given independent (but not necessarily uniform) $D = [D_1, \ldots, D_L]$, the algorithm $\tilde{\theta}^{(p+1)} = H(\tilde{\theta}^{(p)})$ makes $\log L(\theta; D)$ non-decrease for each iteration, or $\log L(H(\tilde{\theta}^{(p)}); D) \geq \log L(\tilde{\theta}^{(p)}; D)$.*

Proposition 3.2 *Given independent (but not necessarily uniform) $D = [D_1, \ldots, D_L]$, the algorithm $\tilde{\theta}^{(p+1)} = M(\tilde{\theta}^{(p)})$ makes $L(\theta; D)$ non-decrease for each iteration.*

Proof. We rewrite (62) as

$$\log L(\theta; D) = Q(\theta, \tilde{\theta}^{(p)}) - T(\theta, \tilde{\theta}^{(p)}) \tag{67}$$

where $Q(\theta, \tilde{\theta}^{(p)}) = \sum_l Q_l(\theta, \tilde{\theta}^{(p)})$, $T(\theta, \tilde{\theta}^{(p)}) = \sum_l T_l(\theta, \tilde{\theta}^{(p)})$.

We have as usual for any $\theta \neq \tilde{\theta}^{(p)}$,

$$T(\theta, \tilde{\theta}^{(p)}) \leq T(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}). \tag{68}$$

For $Q(\theta, \tilde{\theta}^{(p)})$, we have

$$\begin{aligned} Q(\theta, \tilde{\theta}^{(p)}) &= \sum_l Q_l(\theta, \tilde{\theta}^{(p)}) \\ &= \sum_l \sum_q P_{\tilde{\theta}^{(p)}}(D_l^+(q)|D_l) \log P_\theta(D_l^+(q)). \end{aligned} \tag{69}$$

Since $D_l^+(q)$ is a complete observation, by lemma 2.5 we see that each $\log P_\theta(D_l^+(q))$, negative, is strictly concave in θ. It is easy to check that

$$\begin{aligned} g_l(z) \quad &\text{negative and strictly concave} \quad \forall\, l \qquad \Rightarrow \\ g(z) = \sum_l \xi_l\, g_l(z) \quad &\xi_l > 0,\, \forall l \quad \text{also negative and strictly concave.} \end{aligned}$$

We conclude that $Q(\theta, \tilde{\theta}^{(p)})$ is such a linear combination and thus negative and strictly concave. So for operator M with $\gamma_p \in (0, 1]$,

$$\begin{aligned} Q(M(\tilde{\theta}^{(p)}), \tilde{\theta}^{(p)}) &= Q((1-\gamma_p)\tilde{\theta}^{(p)} + \gamma_p H(\tilde{\theta}^{(p)}), \tilde{\theta}^{(p)}) \\ &\geq (1-\gamma_p) Q(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}) + \gamma_p Q(H(\tilde{\theta}^{(p)}), \tilde{\theta}^{(p)}) \quad \text{by concavity} \\ &\geq (1-\gamma_p) Q(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}) + \gamma_p Q(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}) \quad \text{by proposition 3.1} \\ &= Q(\tilde{\theta}^{(p)}, \tilde{\theta}^{(p)}). \end{aligned} \tag{70}$$

Combine (68) and (70) we see that $\log L(M(\tilde{\theta}^{(p)}); D) \geq \log L(\tilde{\theta}^{(p)}; D)$ □

Proposition 3.3 *The algorithm $\tilde{\theta}^{(p+1)} = M(\tilde{\theta}^{(p)})$ will make $\log L(\theta; D)$ converge to $\log L(\theta^*; D)$, where θ^* is the the fixed point of operator M, and further, θ^* is the global maxima.*

Proof. From proposition 3.2 we see that $\log L(\theta; D)$ is non-decreasing under operator M; also from lemma 2.5, we know that $\log L(\theta; D) < 0$, bounded above. So the sequence $\{\log L(\theta; D)\}$ under M converges to the limit, say $\log L^*(\theta; D)$. For continuous (and thus measurable) function $\log L(\theta; D)$, let θ^* be the set of parameters corresponding to $\log L^*(\theta; D)$, or $\log L(\theta^*; D) = \log L^*(\theta; D)$. From $\log L(\theta^{(p+1)}; D) - \log L(\theta^{(p)}; D) \longrightarrow 0$ and $\log L(\theta^{(p+1)}; D) \geq \log L(\theta^{(p)}; D)$, we see that $\nabla \log L(\theta^{(p)}; D) \longrightarrow 0$ and so θ^* is a stationary point.

Now we turn to prove that θ^* is the fixed point of operator M.

For the constrained optimization problem (37), we exploit the Lagrange Multiplier method [16] and define the Lagrangian as

$$L = \log P_\theta(D) - \lambda(\sum_{j'} \theta_{ij'k} - 1), \tag{71}$$

which implies that $\partial L/\partial \theta_{ijk} = \partial \log P_\theta(D)/\partial \theta_{ijk} - \lambda$.

The gradient $\partial \log P_\theta(D)/\partial \theta_{ijk}$ can be computed locally by using information that is available in the normal course of belief network calculations, as shown below [4][27].

$$\begin{aligned} \frac{\partial \log P_\theta(D)}{\partial \theta_{ijk}} &= \frac{\partial \log \prod_{l=1}^{L} P_\theta(D_l)}{\partial \theta_{ijk}} \\ &= \sum_{l=1}^{L} \frac{\partial \log P_\theta(D_l)}{\partial \theta_{ijk}} \\ &= \sum_{l=1}^{L} \frac{\partial P_\theta(D_l)/\partial \theta_{ijk}}{P_\theta(D_l)}. \end{aligned} \tag{72}$$

In order to get an expression in terms of information local to the parameter θ_{ijk}, we introduce X_i and Π_i by averaging over their possible values:

$$\begin{aligned} & \frac{\partial P_\theta(D_l)/\partial \theta_{ijk}}{P_\theta(D_l)} \\ = & \frac{\frac{\partial}{\partial \theta_{ijk}} \left(\sum_{j',k'} P_\theta(D_l | x_i^{j'}, \pi_i^{k'}) P_\theta(x_i^{j'}, \pi_i^{k'}) \right)}{P_\theta(D_l)} \\ = & \frac{\frac{\partial}{\partial \theta_{ijk}} \left(\sum_{j',k'} P_\theta(D_l | x_i^{j'}, \pi_i^{k'}) P_\theta(x_i^{j'} | \pi_i^{k'}) P_\theta(\pi_i^{k'}) \right)}{P_\theta(D_l)}. \end{aligned} \tag{73}$$

Observe that the important property of this expression is that θ_{ijk} appears only in one term in the summation: the term for $j' = j, k' = k$. For this term, $P_\theta(x_i^{j'} | \pi_i^{k'})$ is just θ_{ijk}, so we have

$$\begin{aligned} \frac{\partial P_\theta(D_l)/\partial \theta_{ijk}}{P_\theta(D_l)} &= \frac{P_\theta(D_l | x_i^j, \pi_i^k) P_\theta(\pi_i^k)}{P_\theta(D_l)} \\ &= \frac{P_\theta(x_i^j, \pi_i^k | D_l) P_\theta(D_l) P_\theta(\pi_i^k)}{P_\theta(x_i^j, \pi_i^k) P_\theta(D_l)} \\ &= \frac{P_\theta(x_i^j, \pi_i^k | D_l)}{P_\theta(x_i^j | \pi_i^k)} = \frac{P_\theta(x_i^j, \pi_i^k | D_l)}{\theta_{ijk}}. \end{aligned} \tag{74}$$

Since θ^* is the stationary point, we have $[\partial \log P_\theta(D)/\partial\theta_{ijk} - \lambda]_{\theta=\theta^*} = 0$, or

$$\sum_l^L \frac{P_{\theta^*}(x_i^j, \pi_i^k|D_l)}{\theta^*_{ijk}} - \lambda = 0, \tag{75}$$

and by which

$$\begin{aligned} \lambda\theta^*_{ijk} &= \sum_l^L P_{\theta^*}(x_i^j, \pi_i^k|D_l) \quad \Rightarrow \\ \lambda \sum_j \theta^*_{ijk} &= \sum_l^L \sum_j P_{\theta^*}(x_i^j, \pi_i^k|D_l) \quad \Rightarrow \\ \lambda &= \sum_{l=1}^L P_{\theta^*}(\pi_i^k|D_l). \end{aligned} \tag{76}$$

so we have $\theta^*_{ijk} = P_{\theta^*}(x_i^j, \pi_i^k|D_l)/\lambda = P_{\theta^*}(x_i^j, \pi_i^k|D_l)/P_{\theta^*}(\pi_i^k|D_l)$, or θ^* is the fixed point for operator H. It is straightforward to check that $\theta^* = M(\theta^*)$.

Since $\log L(\theta; D)$ is strictly concave, θ^* is the global maxima. □

Based on the above lemmas and propositions, we sum up and get the following proposition:

Proposition 3.4 *For any given data set D and let sequence $\{\gamma_p\} \in (0, 1]$, $\forall p$, the small step size EM algorithm according to operator M converges to the global maxima of the likelihood function $L(\theta; D)$, starting from any guess of initial parameter set $\theta^{(0)}$.*

Note that as long as sequence $\{\gamma_p\} \in (0, 1]$, $\forall p$ the proposition holds; the choice of step size, however, only affects the rate of convergence, as discussed in section 3.4. For Bayesian analysis, $\theta^{\text{MAP}}_{ijk} = \frac{\theta^*_{ijk}+\alpha_{ijk}}{\theta^*_{ik}+\alpha_{ik}}$, where α_{ijk} and α_{ik} are the hyper-parameters for Dirichlet distribution.

Proposition 3.5 *After convergence, θ^* is the MVUE for augmented complete data $C(D)$ and henceforth.*

Proof. By lemma 2.7, we see that after every iteration, $\tilde{\theta}^{(p+1)}$ is MVUE for $C(D)$ obtained under $\tilde{\theta}^{(p)}$. θ^* is the fixed point and thus MVUE henceforth. □

But is this θ^* also the MVUE for the original incomplete data D? The following is such a proposition which is yet to be proved (or disproved).

Proposition 3.6 *The algorithm $\tilde{\theta}^{(p+1)} = M(\tilde{\theta}^{(p)})$ achieves MVUE for original problem under D.*

3.4 Convergence rate and choice of step size

For the 3-dimensional matrix θ, we define $\psi = \theta_{ik}$, which is a $J \times 1$ vector with $\psi_j = \theta_{ijk}$ as the j^{th} component. We study the convergence rate problem of matrix θ by looking at each such ψ. It can be shown that at the neighborhood of ψ^*,

$$\psi^{(p+1)} - \psi^* \approx J(\psi^*)(\psi^{(p)} - \psi^*), \tag{77}$$

where $J(\psi^*)$ is the Jacobian matrix and the rate of convergence is defined as

$$\nu = \lim_{p \to \infty} \frac{\| \psi^{(p+1)} - \psi^* \|}{\| \psi^{(p)} - \psi^* \|}. \tag{78}$$

Usually, under some regularity conditions, the rate of convergence is

$$\nu = \lambda_{\max}(J(\psi^*)) \quad \text{the largest eigenvalue of } J(\psi^*). \tag{79}$$

Let $J^H(\psi)$ and $J^M(\psi)$ denote the Jacobian matrices under operator H and M, respectively. To obtain $J^H(\psi)$ and $J^M(\psi)$ under $\psi = \psi^*$, we make the following definitions first.

Definition 3.4 *The gradient vector of* $\log L(\theta; D)$ *with respect to* ψ *is*

$$S(D; \psi) = \partial \log L(\theta; D)/\partial\psi. \tag{80}$$

We can see from (74) that

$$S(D; \psi)(j) = \sum_l P_\theta(x_i^j, \pi_i^k | D_l)/\theta_{ijk}. \tag{81}$$

Definition 3.5 *The gradient vector of* $\log L_c(\theta; D^+)$ *with respect to* ψ *is*

$$S_c(D^+; \psi) = \partial \log L_c(\theta; D^+)/\partial\psi. \tag{82}$$

Similarly we have

$$\begin{aligned} S_c(D^+; \psi)(j) &= \sum_l \sum_q P_\theta(x_i^j, \pi_i^k | D_l^+(q))/\theta_{ijk} \\ &= \sum_l \sum_q I_\theta(x_i^j, \pi_i^k | D_l^+(q))/\theta_{ijk} \\ &= \sum_l \sum_q S_c(D_l^+(q); \psi)(j), \end{aligned} \tag{83}$$

where $I_\theta(x_i^j, \pi_i^k | D_l^+(q))$ is the indicator function.

Lemma 3.6 $S(D; \psi) = E_\theta\{S_c(D^+; \psi) | D\}$

Proof.

$$\begin{aligned}\sum_l P_\theta(x_i^j, \pi_i^k | D_l) &= \sum_l \sum_q P_\theta(x_i^j, \pi_i^k, D_l^+(q)|D_l) \\ &= \sum_l \sum_q P_\theta(D_l^+(q)|D_l) P_\theta(x_i^j, \pi_i^k | D_l, D_l^+(q)) \\ &= \sum_l \sum_q P_\theta(D_l^+(q)|D_l) I_\theta(x_i^j, \pi_i^k | D_l^+(q)). \end{aligned} \quad (84)$$

$\Rightarrow S(D;\psi) = E_\theta\{S_c(D^+;\psi)|D\}$. □

Definition 3.7 *The negative of Hessian matrix under complete data is* $I(\psi;D) = -\partial^2 \log L(\theta;D)/\partial\psi\partial\psi^T$.

Definition 3.8 *The negative of Hessian matrix under incomplete data is* $I_c(\psi;D^+) = -\partial^2 \log L_c(\theta;D^+)/\partial\psi\partial\psi^T$.

Lemma 3.9 $I(\psi;D) = \mathcal{I}_{\rm c}(\psi;D) - \mathcal{I}_{\rm m}(\psi;D)$, *where* $\mathcal{I}_{\rm c}(\psi;D) \stackrel{def}{=} E_\theta\{I_c(\psi;D^+)|D\}$ *and* $\mathcal{I}_{\rm m}(\psi;D) \stackrel{def}{=} E_\theta\{\partial^2 \log P_\theta(D^+|D)/\partial\psi\partial\psi^T\}$.

Proof. This lemma is called the missing information principle, firstly originated by Orchard in 1972 [19]. It is straightforward to see that

$$\begin{aligned} I(\psi;D) &= E_\theta\{I(\psi;D)|D\} \\ &= -\sum_l \sum_q P_\theta(D_l^+(q)|D_l) \partial^2 \log L_c(\theta; D_l^+(q))/\partial\psi\partial\psi^T \\ &\quad + \sum_l \sum_q P_\theta(D_l^+(q)|D_l) \partial^2 \log P_\theta(D_l^+(q)|D_l)/\partial\psi\partial\psi^T \\ &= E_\theta\{I_c(\psi;D^+)|D\} + E_\theta\{\partial^2 \log P_\theta(D^+|D)/\partial\psi\partial\psi^T\}. \end{aligned} \quad (85)$$

$\Rightarrow I(\psi;D) = \mathcal{I}_{\rm c}(\psi;D) - \mathcal{I}_{\rm m}(\psi;D)$. □

Lemma 3.10 *For operator* H, $J^H(\psi^*) = \mathcal{I}_{\rm c}{}^{-1}(\psi^*;D)\,\mathcal{I}_{\rm m}(\psi^*;D)$.

Proof. This lemma is adapted from [8]. By proposition 3.3 we know that $S(D;\psi^*) = 0$ and at the neighborhood of ψ^*,

$$\begin{aligned} S(D;\psi^*) &\approx S(D;\psi^{(p)}) - I(\psi^{(p)};D)(\psi^* - \psi^{(p)}) \quad \Rightarrow \\ \psi^* &\approx \psi^{(p)} + I^{-1}(\psi^{(p)};D) S(D;\psi^{(p)}) \end{aligned} \quad (86)$$

At M-step, for $Q(\psi,\psi^{(p)}) = \sum_l \sum_q P_{\theta^{(p)}}(D_l^+(q)|D_l) \log P_\theta(D_l^+(q))$ we have

$$0 = [\partial Q(\psi,\psi^{(p)})/\partial\psi]_{\psi=\psi^{(p+1)}}$$

$$
\begin{aligned}
&\approx \underbrace{[\partial Q(\psi,\psi^{(p)})/\partial\psi]_{\psi=\psi^{(p)}}}_{S(D;\psi^{(p)})\text{ by lemma 3.6}} + [\partial^2 Q(\psi,\psi^{(p)})/\partial\psi\partial\psi^T]_{\psi=\psi^{(p)}}\,(\psi^{(p+1)}-\psi^{(p)}) \\
&= S(D;\psi^{(p)}) - \mathcal{I}_{\mathrm{c}}(\psi^{(\mathrm{p})};\mathrm{D})(\psi^{(\mathrm{p+1})}-\psi^{(\mathrm{p})}) \quad \Rightarrow \\
S(D;\psi^{(p)}) &\approx \mathcal{I}_{\mathrm{c}}(\psi^{(\mathrm{p})};\mathrm{D})(\psi^{(\mathrm{p+1})}-\psi^{(\mathrm{p})}). \qquad (87)
\end{aligned}
$$

From (86) and (87) we have

$$
\begin{aligned}
\psi^*-\psi^{(p)} &\approx I^{-1}(\psi^{(p)};D)\mathcal{I}_{\mathrm{c}}(\psi^{(\mathrm{p})};\mathrm{D})(\psi^{(\mathrm{p+1})}-\psi^{(\mathrm{p})}) \quad \Rightarrow \\
\psi^{(p+1)}-\psi^* &\approx [I_J - \mathcal{I}_{\mathrm{c}}^{-1}(\psi^{(\mathrm{p})};\mathrm{D})\mathrm{I}(\psi^{(\mathrm{p})};\mathrm{D})](\psi^{(\mathrm{p+1})}-\psi^*) \\
&\approx [I_J - \mathcal{I}_{\mathrm{c}}^{-1}(\psi^{*};\mathrm{D})\mathrm{I}(\psi^{*};\mathrm{D})](\psi^{(\mathrm{p+1})}-\psi^*) \\
&\approx \mathcal{I}_{\mathrm{c}}^{-1}(\psi^{*};\mathrm{D})\mathcal{I}_{\mathrm{m}}(\psi^{*};\mathrm{D})(\psi^{(\mathrm{p+1})}-\psi^*). \quad \text{by lemma 3.9} \qquad (88)
\end{aligned}
$$

$\Rightarrow J^H(\psi^*) = \mathcal{I}_{\mathrm{c}}^{-1}(\psi^*;\mathrm{D})\mathcal{I}_{\mathrm{m}}(\psi^*;\mathrm{D})$. □

Lemma 3.11 $\mathcal{I}_{\mathrm{m}}(\psi^*;\mathrm{D}) = \sum_{\mathrm{l}}\sum_{\mathrm{q}} \mathrm{P}_\theta(\mathrm{D}_{\mathrm{l}}^+(\mathrm{q})|\mathrm{D}_{\mathrm{l}})\,\mathrm{S}_{\mathrm{c}}(\mathrm{D}_{\mathrm{l}}^+(\mathrm{q});\psi)\,\mathrm{S}_{\mathrm{c}}^{\mathrm{T}}(\mathrm{D}_{\mathrm{l}}^+(\mathrm{q});\psi) - \sum_{\mathrm{l}}\mathrm{S}(\mathrm{D}_{\mathrm{l}};\psi)\,\mathrm{S}^{\mathrm{T}}(\mathrm{D}_{\mathrm{l}};\psi)$

Proof. This lemma is similar to and adapted from that proposed by Louis in 1982 [15].

$$
\begin{aligned}
I(\psi;D) &= -\partial S(D;\psi)/\partial\psi \\
&= = -\sum_l \partial S(D_l;\psi)/\partial\psi = -\sum_l \partial[\frac{\partial L(\theta;D_l)/\partial\psi}{L(\theta;D_l)}]/\partial\psi \\
&= -\sum_l \partial[\frac{\sum_q \partial L_c(\theta;D_l^+(q))/\partial\psi}{L(\theta;D_l)}]/\partial\psi \\
&= -\sum_l\sum_q \frac{\partial^2 L_c(\theta;D_l^+(q))/\partial\psi\partial\psi^T}{L(\theta;D_l)} \\
&\quad +\sum_l \left[\frac{\sum_q \partial L_c(\theta;D_l^+(q))/\partial\psi}{L(\theta;D_l)}\right]\left[\frac{\sum_q \partial L_c(\theta;D_l^+(q))/\partial\psi}{L(\theta;D_l)}\right]^T \\
&= -\sum_l\sum_q \frac{\partial^2 L_c(\theta;D_l^+(q))/\partial\psi\partial\psi^T}{L(\theta;D_l)} + \sum_l S(D_l;\psi)\,S^T(D_l;\psi) \\
&= -\sum_l\sum_q \left[\partial^2 \log L_c(\theta;D_l^+(q))/\partial\psi\partial\psi^T\right]\frac{L(\theta;D_l^+(q))}{L(\theta;D_l)} \\
&\quad -\sum_l\sum_q \left[\frac{\partial L_c(\theta;D_l^+(q))/\partial\psi}{L_c(\theta;D_l^+(q))}\right]\left[\frac{\partial L_c(\theta;D_l^+(q))/\partial\psi}{L_c(\theta;D_l^+(q))}\right]^T + \sum_l S(D_l;\psi)\,S^T(D_l;\psi) \\
&= \sum_l\sum_q I_c(\psi;D_l^+(q))\,P_\theta(D_l^+(q)|D_l) \\
&\quad -\sum_l\sum_q S_c(D_l^+(q);\psi)\,S_c^T(D_l^+(q);\psi)\,P_\theta(D_l^+(q)|D_l) + \sum_l S(D_l;\psi)\,S^T(D_l;\psi). \qquad (89)
\end{aligned}
$$

Since $\sum_l\sum_q I_c(\psi;D_l^+(q))\,P_\theta(D_l^+(q)|D_l) = I(\psi;D)$, and by lemma 3.9 we finish the proof. □

So at the neighborhood of θ^*, we can now compute the Jacobian matrix by using lemma 3.10 and 3.11, as shown below:

- Computation of $\mathcal{I}_{\rm c}^{-1}(\psi^*;{\rm D})$:

  Obviously, $I_c(\psi;D_l^+(q)) = -\partial S_c(D_l^+(q);\psi)/\partial\psi$, where $S_c(D_l^+(q);\psi)(j) = I_\theta(x_i^j,\pi_i^k|D_l^+(q))/\theta_{ijk}$. So it is easy to check that for $j \neq j'$, $\partial S_c(D_l^+(q);\psi)(j)/\partial\psi_{j'} = 0$ and so

$$\begin{aligned}
\mathcal{I}_{\rm c}(\psi;\mathcal{D}) &= \sum_l \sum_q P_\theta(D_l^+(q)|D_l)\, I_c(\psi;D_l^+(q)) \\
&= \sum_l \sum_q P_\theta(D_l^+(q)|D_l)\, {\rm diag}\{I_\theta(x_i^j,\pi_i^k|D_l^+(q))/\theta_{ijk}^2\} \\
&= {\rm diag}\left\{\sum_l \sum_q P_\theta(D_l^+(q)|D_l)\, I_\theta(x_i^j,\pi_i^k|D_l^+(q))/\theta_{ijk}^2\right\} \\
&= {\rm diag}\left\{\sum_l P_\theta(x_i^j,\pi_i^k|D_l)/\theta_{ijk}^2\right\} \\
&= {\rm diag}\{a_j\}, \qquad (90)
\end{aligned}$$

  where $a_j \stackrel{\triangle}{=} \sum_l P_\theta(x_i^j,\pi_i^k|D_l)/\theta_{ijk}^2 \quad j=1,\ldots,J.$

- Computation of $\sum_l \sum_q S_c(D_l^+(q);\psi)\, S_c^T(D_l^+(q);\psi)\, P_\theta(D_l^+(q)|D_l)$:

  From $S_c(D_l^+(q);\psi)(j) = I_\theta(x_i^j,\pi_i^k|D_l^+(q))/\theta_{ijk}$, we get $I_\theta(x_i^j,\pi_i^k|D_l^+(q))\, I_\theta(x_i^{j'},\pi_i^k|D_l^+(q)) = 0\ \forall j' \neq j$ and thus

$$S_c(D_l^+(q);\psi)\, S_c^T(D_l^+(q);\psi) = {\rm diag}\{I_\theta(x_i^j,\pi_i^k|D_l^+(q))/\theta_{ijk}^2\}, \qquad (91)$$

  and similarly we have

$$\sum_l \sum_q S_c(D_l^+(q);\psi)\, S_c^T(D_l^+(q);\psi)\, P_\theta(D_l^+(q)|D_l) = {\rm diag}\{a_j\} \quad j=1,\ldots,J. \qquad (92)$$

- Computation of $\sum_l S(D_l;\psi)\, S^T(D_l;\psi)$:

  Let $V = \sum_l S(D_l;\psi)\, S^T(D_l;\psi)$, and since $S(D_l;\psi)(j) = P_\theta(x_i^j,\pi_i^k|D_l)$, we have for matrix $V = \{v_{ms}\}$:

$$v_{ms} = \begin{cases} \frac{\sum_l P_\theta(x_i^m,\pi_i^k|D_l)\, P_\theta(x_i^s,\pi_i^k|D_l)}{\theta_{imk}\,\theta_{isk}} & \text{if } m \neq s \\ \frac{\sum_l P_\theta^2(x_i^m,\pi_i^k|D_l)}{\theta_{imk}^2} & \text{if } m = s \end{cases} \qquad (93)$$

  And then we can see that

$$\begin{aligned}
J^H(\psi^*) &= \mathcal{I}_{\rm c}^{-1}(\psi^*;{\rm D})\,\mathcal{I}_{\rm m}(\psi^*;{\rm D}) \\
&= {\rm diag}\{1/a_j\}\ [{\rm diag}\{a_j\} - V] \\
&= I_J - \bar{W}, \qquad (94)
\end{aligned}$$

where $\bar{W} = \text{diag}\{1/a_j\}\, V$. For each of its element,

$$\begin{aligned} \bar{w}_{mm} &= \frac{\theta_{imk}^2}{\sum_l P_\theta(x_i^m, \pi_i^k|D_l)} \frac{\sum_l P_\theta^2(x_i^m, \pi_i^k|D_l)}{\theta_{imk}^2} \\ &= \frac{\sum_l P_\theta^2(x_i^m, \pi_i^k|D_l)}{\sum_l P_\theta(x_i^m, \pi_i^k|D_l)} \end{aligned} \tag{95}$$

$$\begin{aligned} \bar{w}_{ms} &= \frac{\theta_{imk}^2}{\sum_l P_\theta(x_i^m, \pi_i^k|D_l)} \frac{\sum_l P_\theta(x_i^m, \pi_i^k|D_l)\, P_\theta(x_i^s, \pi_i^k|D_l)}{\theta_{imk}\, \theta_{isk}} \\ &= \frac{\theta_{imk}}{\theta_{isk} \sum_l P_\theta(x_i^m, \pi_i^k|D_l)} \sum_l P_\theta(x_i^m, \pi_i^k|D_l)\, P_\theta(x_i^s, \pi_i^k|D_l) \\ &= \frac{\sum_l P_\theta(x_i^m, \pi_i^k|D_l)\, P_\theta(x_i^s, \pi_i^k|D_l)}{\sum_l P_\theta(x_i^s, \pi_i^k|D_l)}, \end{aligned} \tag{96}$$

based on the fact that if $\theta = \theta^*$, $\theta_{imk} = \sum_l P_\theta(x_i^m, \pi_i^k|D_l) \,/\, \sum_l P_\theta(\pi_i^k|D_l)$ under operator H. If we define $w_{jl} = P_\theta(x_i^j, \pi_i^k|D_l)$ at $\theta = \theta^*$, it is easy to see that $\bar{W} = W\,\Lambda$, where $\Lambda = \text{diag}\,\{1/\sum_l w_{jl}\}$ and

$$W = \begin{pmatrix} \sum_l w_{1l}^2 & \sum_l w_{1l} w_{2l} & \dots & \sum_l w_{1l} w_{Jl} \\ \sum_l w_{1l} w_{2l} & \sum_l w_{2l}^2 & \dots & \sum_l w_{1l} w_{Jl} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_l w_{1l} w_{Jl} & \sum_l w_{1l} w_{2l} & \dots & \sum_l w_{Jl}^2 \end{pmatrix} \tag{97}$$

Obviously, $W = \sum_l w_l w_l^T$ where $w_l^T = [w_{1l}, \dots, w_{Jl}]$.

For operator M, we have

$$\begin{aligned} J^M(\psi^*) &= (1-\gamma) I_J + \gamma J^H(\psi^*) \\ &= I_J - \gamma W\,\Lambda. \end{aligned} \tag{98}$$

To study the eigenvalues of $J^M(\psi^*)$, we first look at those of $W\,\Lambda$. It is straightforward to see that for any non-zero vector $\xi \in R^J$,

$$\begin{aligned} \lambda_{\max}(\bar{W}) &= \max_\xi \frac{(\xi, \sum_l w_l w_l^T \Lambda\, \xi)}{(\xi, \xi)} \\ &= \max_\xi \sum_l \underbrace{\frac{(\xi, w_l w_l^T \Lambda\, \xi)}{(\xi, \xi)}}_{\le \lambda_{\max}(w_l w_l^T \Lambda)} \le \sum_l \lambda_{\max}(w_l w_l^T \Lambda) \quad \text{and,} \end{aligned} \tag{99}$$

$$\begin{aligned} \lambda_{\min}(\bar{W}) &= \min_\xi \frac{(\xi, \sum_l w_l w_l^T \Lambda\, \xi)}{(\xi, \xi)} \\ &= \min_\xi \sum_l \underbrace{\frac{(\xi, w_l w_l^T \Lambda\, \xi)}{(\xi, \xi)}}_{\ge \lambda_{\min}(w_l w_l^T \Lambda)} \ge \sum_l \lambda_{\min}(w_l w_l^T \Lambda), \end{aligned} \tag{100}$$

and so we have the following lemmas.

Lemma 3.12 $\lambda_{\max}(w_l w_l^T \Lambda) = \sum_j w_{jl}^2 / \sum_l w_{jl}, \quad \lambda_{\min}(w_l w_l^T \Lambda) = 0$

Proof. If we let $\Lambda = \text{diag}\{\rho_j\}$ with $\rho_j > 0$ (this condition stands true if data sample size L is large enough), then for any non-zero $\xi \in R^J$, we have

$$\begin{aligned} \sum_l \xi^T w_l w_l^T \Lambda \xi &= \sum_l (\xi_1 w_{1l} + \ldots + \xi_J w_{Jl}) (\rho_1 \xi_1 w_{1l} + \ldots + \rho_J \xi_J w_{Jl}) \\ &\geq (\sqrt{\rho_1}\xi_1 w_{1l} + \ldots + \sqrt{\rho_J}\xi_j w_{jl})^2 \geq 0. \end{aligned} \tag{101}$$

So matrix $w_l w_l^T \Lambda$ is positive semi-definite, with all of its eigenvalues non-negative. Note also that in $w_l w_l^T \Lambda$, each row is of a constant factor of any other row, we see that

$$\det (w_l w_l^T \Lambda) = \prod_j \lambda_j (w_l w_l^T \Lambda) = 0 \tag{102}$$

and $\lambda_{\min}(w_l w_l^T \Lambda) = 0$.

Let $\Upsilon = (w_l w_l^T \Lambda)(w_l w_l^T \Lambda)^T$, and we know that $\lambda_{\max}(w_l w_l^T \Lambda) = \|w_l w_l^T \Lambda\| = [tr(\Upsilon)]^{1/2}$. It is easy to check that

$$v_{jj} = \frac{w_{jl}^2}{\sum_l w_{jl}} \left[\frac{w_{1l}^2}{\sum_l w_{1l}} + \ldots + \frac{w_{Jl}^2}{\sum_l w_{Jl}} \right] \quad \Rightarrow \tag{103}$$

$$\lambda_{\max}(w_l w_l^T \Lambda) = \left[\left(\frac{\sum_j w_{jl}^2}{\sum_l w_{jl}} \right)^2 \right]^{1/2} \tag{104}$$

$w_l w_l^T \Lambda$ positive semi-definite $\quad \Rightarrow \quad \lambda_{\max}(w_l w_l^T \Lambda) = \sum_j w_{jl}^2 / \sum_l w_{jl}$ □

Lemma 3.13 $\lambda_{\max}(J^M(\psi^*)) \leq 1, \quad \lambda_{\min}(J^M(\psi^*)) \geq 0$

Proof.

$$\begin{aligned} \lambda_{\max}(\bar{W}) &\leq \sum_l \lambda_{\max}(w_l w_l^T \Lambda) \\ &= \sum_l \left(\sum_j w_{jl}^2 / \sum_m w_{jm} \right) \\ &= \sum_j \left(\sum_l w_{jl}^2 / \sum_m w_{jm} \right) \\ &\leq \sum_j \left[(\sum_l w_{jl})^2 / \sum_m w_{jm} \right] \end{aligned}$$

$$= \sum_l \sum_j P_{\theta^*}(x_i^j, \pi_i^j | D_l)$$
$$= \sum_l P_{\theta^*}(\pi_i^j | D_l) \leq 1 \quad (105)$$

$$\lambda_{\min}(\bar{W}) \geq \sum_l \lambda_{\min}(w_l w_l^T \Lambda) = 0 \quad (106)$$

Then for operator M with $\gamma \in (0, 1]$, we have $\lambda_{\max}(J^M(\psi^*)) = 1 - \gamma\lambda_{\min}(\bar{W}) \leq 1$ and $\lambda_{\min}(J^M(\psi^*)) = 1 - \gamma\lambda_{\max}(\bar{W}) \geq 0$. □

From lemma 3.13 we see that the rate of convergence $\nu = \lambda_{\max}(J^M(\psi^*)) \leq 1$, which states also that at the neighborhood of θ^*, operator M will make the parameter sequence converge to the local maxima θ^*. To let the algorithm behaves well, we discuss the choice of step size next.

If we consider vector $\psi \in R^J$, for every component j we choose $\gamma_p(j)$ as follows [3]:

$$\gamma_p(j) > 0 \quad (107)$$
$$\sum_{p=0}^{\infty} \gamma_p(j) = \infty \quad \text{with probability 1} \quad (108)$$
$$\sum_{p=0}^{\infty} \gamma_p^2(j) < \infty \quad \text{with probability 1} \quad (109)$$

References

[1] J. S. Baras, H. Li and G. Mykoniatis, "Integrated, Distributed Fault Management for Communication Networks", Technical Report, CSHCN TR 98-10, University of Maryland, 1998

[2] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations,* Springer-Verlag, 1987

[3] D. P. Bertsekas, and J. N. Tsitsiklis, *Neuro-Dynamic Programming,* Athena Scientific, 1996

[4] J. Binder, D. Koller, S. Russell, K. Kanazawa, " Adaptive Probabilistic Networks with Hidden Variables." *Machine Learning*, in press, 1997.

[5] E. Castillo, J. M. gutierrez, and A. S. Hadi, *Expert Systems and Probabilistic Network Models,* Springer, 1997

[6] G. F. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning,* 9, 309-347, 1992

[7] T. Cover, *Elements of Information Theory ,* John Wiley and Sons, Inc., 1991

[8] A. P. Dempster, N. M. Laird and D. B. Rubin "Maximum Likelihood from Incomplete Data via the *EM* Algorithm (with discussion)" *Journal of the Royal Statistics Society B*, 39, pp. 1-38, 1977.

[9] V. Fabian, "Asymptotically Efficient Stochastic Approximation; The RM Case", *The Annals of Statistics,* vol. 1, No. 3, pp. 486-495, 1973

[10] N. Friedman, M. Goldszmidt, D. Heckerman, and S. Russell, "Challenge: Where is the Impact of Bayesian Networks in Learning?", Technical Report, 1997

[11] N. Friedman and M. Goldszmidt, "AAAI-98 Tutorial on learning Bayesian networks from Data", 1998

[12] D. Heckerman, "A tutorial on learning with Bayesian networks", Microsoft Research Technical Report MSR-TR-94-09, 1996

[13] H. Li, "An Introduction to Belief Networks", Technical Report, CHSCN, University of Maryland, 1998

[14] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation and Optimization of Random Systems,* Birkhäuser, 1992

[15] T. A. Louis, "Finding the observed information matrix when using the EM algorithm", *Journal of the Royal Statistical Society B*, 44, pp. 226-233, 1982

[16] D. G. Luenberger, *Linear and Nonlinear Programming,* Addison-Wesley, 1984

[17] G. J. McLachlan and T. Krishnan, *The EM Algorithm ad Extensions*, Wiley Interscience, 1997

[18] R. M. Neal, "Probabilistic Inference Using Markov Chain Monte Carlo Methods", Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993

[19] T. Orchard and M. A. Woodbury, "A missing information principle: theory and applications", In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, Berkeley, California: University of California Press, pp. 697-715, 1972

[20] J. Pearl, "Fusion, Propagation, and Structuring in Belief Networks", *Artificial Intelligence,* vol. 29, pp. 241-288, 1986

[21] J. Pearl, *Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann, 1988

[22] B. T. Polyak and A. B. Juditsky,"Acceleration of stochastic approximation by averaging", *SIAM J. Contr. Opt.* vol. 30, pp. 838-855, July 1992

[23] H. V. Poor, *An Introduction to Signal Detection and Estimation,* Springer-Verlag, 1994

[24] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, 4-16, 1986

[25] H. Robbins and S. Monro, "A stochastic approximation method", *The Annals of Mathematical Statistics,* vol. 22, pp. 400-407, 1951

[26] R. Y. Rubinstein, *Simulation and the Monte Carlo Method*, Wiley, 1981

[27] S. Russell and P. Norvig, *Artificial Intelligence - A Modern Approach,* Prentice-Hall, 1995