ESAIM: PS
September 2005, Vol. 9, p. 241–253
DOI: 10.1051/ps:2005012

ESAIM: Probability and Statistics

ESTIMATION OF PARAMETERS IN A NETWORK RELIABILITY MODEL WITH SPATIAL DEPENDENCE*

Ian Hepburn Dinwoodie[1]

Abstract. An iterative method based on a fixed-point property is proposed for finding maximum likelihood estimators for parameters in a model of network reliability with spatial dependence. The method is shown to converge at a geometric rate under natural conditions on data.

Mathematics Subject Classification. 62B05, 62F10.

Received October 3, 2003. Revised May 10, 2005.

Introduction

A model for network reliability with spatial dependence was formulated in [7] that generalized the Bernoulli model of [3]. The problem is to infer internal link failure probabilities from aggregate failure counts. An approximate maximum likelihood estimator (mle) was proposed based on a one-step relaxation method. In this paper, we describe an iterative scheme to find the numerical values of the mle based on a fixed point property.

Recent work on multicast network tomography has developed methodology for networks more general than trees [2], for missing data [8], for sample size [9], and for efficient probing [15], always under the assumption of independent link failures. There is also work on using unicast pairs for the same purpose [13], again assuming independent losses on different links. An attempt to deal with dependence among components in a network is in [10], where pseudo-likelihoods replace likelihoods for simplicity and approximate methods are used. A wide and useful survey of internet tomography is [4], which concludes by saying that relaxing assumptions of stationarity and independence is of great interest.

The focus here is on a specific multicast model with dependent link failures, for the simple tree topology. The model for dependence adds a single interaction parameter θ which corresponds to the exponential of inverse temperature in an interaction potential over all pairs of links. The extra parameter complicates the identifiability and estimation because the recursive method of [3] is no longer is applicable. The foundations were put in place in [7], but an efficient procedure for exact maximum likelihood was not given. Rather an efficient approximate method was given that modified the estimates for the Bernoulli model. Here we give a new iterative method for maximizing the likelihood, and we also explain the method in the context of the traditional Bernoulli model with independent link failures.

Let us recall the problem and introduce some notation. A tree with vertices V and edge set E has root node $0 \in V$ and "leaf" nodes $R \subset V$ (R stands for receivers). from the root node 0 towards the receiver nodes R,

Keywords and phrases. Curie-Weiss, EM-algorithm, iterative proportional scaling, maximum likelihood, network tomography.

* *This work was supported by NSF grant DMS-0200888 and SAMSI under grant DMS-0112069.*

[1] ISDS, Box 90251, Duke University, Durham NC 27708, USA; ihd@stat.duke.edu

and it copies itself at each vertex onto each subsequent edge on its trip towards the receiver nodes (this is the meaning of "multicast"). The probe is lost on an edge on route to the leaf nodes with a probability that depends on the edge. An observation is a vector $\mathbf{y} \in \{0,1\}^R$, where component i indicates whether the multicast signal was lost on the trip from 0 to the leaf node $i \in R$ ($y_i = 1$ means it was lost). The observation $\mathbf{y}$ is the image under a many-to-one linear map A of a hidden outcome $\mathbf{x}$ indicating success or failure on each edge.

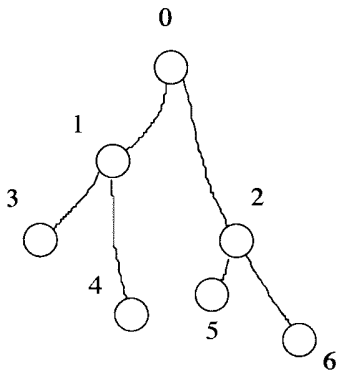

FIGURE 1. Multicast tree.

This experiment is repeated independently and identically $n \geq 1$ times, and observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are a random sample of iid vectors at the receiver nodes with components in $\{0,1\}$. The goal is to estimate internal reliability parameters on edges from the incomplete receiver data. This can be done with at least two probability models, the original Bernoulli model and an interaction model with an extra parameter θ that encourages or discourages multiple losses.

1. Bernoulli multicast model

In this section we describe in detail the Bernoulli multicast model of [3], for which a recursive estimation algorithm exists, but which also can solved with the proposed method. The section will serve to fix notation in a setting less complicated than the full interaction model.

Let $\mathcal{T}$ be the tree with vertices V numbered $0, 1, \ldots, c$. Let the parent of a node i be denoted $f(i)$, and let descendants of node i (the set of nodes whose path back to 0 goes through i, but not including i) be denoted $d(i)$. The siblings of i would be $f^{-1} \circ f(i)$. The assumption that all parent vertices (V_P) have at least two child nodes means that for each $i \in V_P$ it holds that $f^{-1} \circ f(i) - \{i\} \neq \emptyset$.

The parent nodes will be denoted $V_P := V - R$, and V_0 will be the collection of non-root nodes. All vertices in V_P will be assumed to have at least two child nodes. On the tree (1) above for example, $V_0 = \{1,2,3,4,5,6\}$, $c = 6, V_P = \{0,1,2\}, R = \{3,4,5,6\}$, and $f(3) = f(4) = 1$.

Basic hidden outcomes are vectors of counts $\mathbf{x} = (x_i)_{i \in V_0} \in \{0,1\}^{V_0}$, where x_i specifies how many probes were lost on the edge $\{f(i), i\}$ (the edges are labelled by the outer vertex). The multicast data $\mathbf{y}$ can be written as a many-to-one function of basic hidden outcomes $\mathbf{x}$ with the help of a routing matrix A. The matrix A will have $d = |R|$ rows, one for each leaf node, and $c = |V_0|$ columns indexed by edges. The row for leaf node i will have "1" in column j if j is on the path from 0 to leaf node i. For the binary tree in Figure 1, the matrix is given by

$$A = \begin{array}{c} \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}. \tag{1.1}$$

The experiment is repeated $n \geq 1$ times. Then the total observed loss vector $\mathbf{y}_k$ (for experiment k out of n) at leaf nodes is given by

$$\mathbf{y}_k = A\mathbf{x}_k.$$

It is convenient to have a $|V| \times |V_0|$ routing matrix B for all nodes in V, not just the leaf nodes. The row for vertex i would have "1" in each column for vertices on the path from 0 to i. $(B\mathbf{x})_i$ for a node $i \in V_0$ would give the total number of messages lost along the path from 0 out to i. The row for vertex 0 in B is identically 0.

Now let β_i be the probability that a probe from vertex $f(i)$ will fail to cross edge $\{f(i), i\}$ to reach vertex $i \in V_0$. We will use an odds ratio parametrization:

$$\beta_i = \frac{\lambda_i}{1+\lambda_i},\ \lambda_i \geq 0.$$

For the Bernoulli model it is assumed that a probe fails to cross edge $\{f(i), i\}$ with probability $\lambda_i/(1+\lambda_i)$, and edges and probes all behave independently given failure count data on parent nodes (we generalize this below). The distribution μ_λ on $S_0 = \{\mathbf{x} = (x_i) \in Z_+^{V_0} : x_i \in \{0,1\}\}$ is

$$\begin{aligned}
\mu_\lambda(\mathbf{x}) &= \prod_{i \in V_0} \binom{1-(B\mathbf{x})_{f(i)}}{x_i} \frac{\lambda_i^{x_i}}{(1+\lambda_i)^{1-(B\mathbf{x})_{f(i)}}} \\
&= \lambda^{\mathbf{x}} \prod_{i \in V_0} \binom{1-(B\mathbf{x})_{f(i)}}{x_i} \frac{1}{(1+\lambda_i)^{1-(B\mathbf{x})_{f(i)}}} \\
&= \lambda^{\mathbf{x}} \left[\prod_{i \in V_0} \frac{\binom{1-(B\mathbf{x})_{f(i)}}{x_i}}{(1+\lambda_i)} \right] \prod_{i \in V_0} (1+\lambda_i)^{(B\mathbf{x})_{f(i)}} \\
&= \lambda^{\mathbf{x}} \left[\prod_{i \in V_0} \frac{\binom{1-(B\mathbf{x})_{f(i)}}{x_i}}{(1+\lambda_i)} \right] \prod_{i \in V_0 - R} \prod_{j \in d(i)} (1+\lambda_j)^{x_i}. \qquad (1.2)
\end{aligned}$$

For $i \in V_0$, let $p_i = \prod_{j \in d(i)} (1+\lambda_j)$. Then (1.2) can be written

$$\begin{aligned}
\mu_\lambda(\{\mathbf{y}\}) &= \sum_{\{\mathbf{x} \in Z_+^c : A\mathbf{x} = \mathbf{y}\}} h(\mathbf{x}) \frac{\lambda^{\mathbf{x}} \prod_{i \in V_0} p_i(\lambda)^{x_i}}{z_\lambda} \\
&= \sum_{\{\mathbf{x} \in Z_+^c : A\mathbf{x} = \mathbf{y}\}} h(\mathbf{x}) \frac{\lambda^{\mathbf{x}} \mathbf{p}(\lambda)^{\mathbf{x}}}{z_\lambda}
\end{aligned}$$

where

$$\begin{aligned}
h(\mathbf{x}) &= \prod_{i \in V_0} \binom{1-(B\mathbf{x})_{f(i)}}{x_i} \\
z_\lambda &= \prod_{i \in V_0} (1+\lambda_i) \qquad (1.3)
\end{aligned}$$

and the notation $\mathbf{p}(\lambda)^{\mathbf{x}}$ is the usual representation of $\prod_{i=1}^c p_i(\lambda)^{x_i}$. The vector of parameters $(\lambda_i : i \in V_0)$ is identifiable, meaning that two different vectors give rise to two different distributions $\mu_\lambda(\{\mathbf{y}\})$ on the observed (incomplete) data $\mathbf{y}$. (In the Bernoulli model, identifiability holds in fact even if the root vertex 0 has only one child, and the failure probabilities are allowed to be zero.) This gives consistent maximum likelihood estimates (the estimates converge to the true parameter values) as the sample size n increases.

Now we describe some new variables that simplify the likelihood function. Consider the one-to-one reparametrization $\gamma_i = \lambda_i p_i(\lambda)$, for $i \in V_0$, from the set of positive reals in R^{V_0} to itself. Then

$$\mu_{\lambda(\gamma)}(\{\mathbf{y}\}) = \sum_{\{\mathbf{x} \in Z_+^c : A\mathbf{x} = \mathbf{y}\}} h(\mathbf{x}) \frac{\gamma^{\mathbf{x}}}{z_{\lambda(\gamma)}}. \tag{1.4}$$

For each $\mathbf{y} \in Z_+^R$, let $V^{\mathbf{y}} \subset V_0$ be the collection of edge labels that are closest to the root whose failure could lead to observation $\mathbf{y}$. For example, in the binary tree (1), $V^{(1,1,1,1)} = \{1,2\}$, $V^{(1,1,0,1)} = \{1,6\}$. If one defines a partial order on V_0 by $w \leq_{\mathcal{T}} v$ if and only if w is on the path from the root 0 to v, then $V^{\mathbf{y}}$ is the collection of minimal elements with respect to this order of the union of sets of edges in $\pi^{-1}(\mathbf{y})$ where $\pi : 2^{V_0} \to \{0,1\}^R$ is defined by

$$\pi(A) = I_{\cup_{w \in A}\{v \in R \,:\, v \geq_{\mathcal{T}} w\}}.$$

Define polynomials $q_v,\ v \in V_0$ in variables $s_v,\ v \in V_0$ recursively by

$$\begin{aligned} q_r &:= s_r, r \in R \\ q_v &:= s_v + \prod_{w \in f^{-1}(v)} q_w. \end{aligned}$$

The polynomials q_v have a probabilistic meaning. For each vertex $v \in V_0$, let $\mathbf{y}^v$ be the vector in $\{0,1\}^R$ with 1 in each coordinate that is a descendent of v (including v if v is a receiver), which can be written as the indicator function $I_{\{v \cup d(v)\} \cap R}$. Then $\mu_\lambda(\{\mathbf{y}^v\}) = q_v(\gamma)/z_\lambda$, as we show below.

Proposition 1.1. *For the Bernoulli model,*

$$\begin{aligned} \mu_{\lambda(\gamma)}(\{\mathbf{y}\}) &= \frac{1}{z_{\lambda(\gamma)}} \prod_{v \in V^{\mathbf{y}}} q_v(\gamma) \\ z_{\lambda(\gamma)} &= \sum_{\mathbf{y} \in \{0,1\}^R} \prod_{v \in V^{\mathbf{y}}} q_v(\gamma). \end{aligned}$$

Proof. Observe that q_v is the generating function in indeterminates $s_w, w \in V_0$ for the outcomes of edge failures that lead to data $\mathbf{y}^v$. Then $\prod_{v \in V^{\mathbf{y}}} q_v$ is the generating function for the outcomes of edge failures that lead to $\mathbf{y} = \sum_{v \in V^{\mathbf{y}}} \mathbf{y}^v$, and can be written $\sum_{\mathbf{x}: A\mathbf{x} = \mathbf{y}} \mathbf{s}^{\mathbf{x}}$. The probability of each $\mathbf{x}$ in the sum is given by $\gamma^{\mathbf{x}}/z_\lambda(\gamma)$ from (1.4). Then $\mu(\{\mathbf{x} : A\mathbf{x} = \mathbf{y}\}) = \sum_{\mathbf{x}: A\mathbf{x} = \mathbf{y}} \gamma^{\mathbf{x}}/z_{\lambda(\gamma)} = \prod_{v \in V^{\mathbf{y}}} q_v(\gamma)/z_{\lambda(\gamma)}$. □

Define the polynomial $Z(q_1, \ldots, q_c)$ by

$$Z(\mathbf{q}) := \sum_{\mathbf{y} \in \{0,1\}^R} \prod_{v \in V^{\mathbf{y}}} q_v.$$

Proposition 1.2. *For the Bernoulli model,*

$$z_{\lambda(\gamma)} = Z(\mathbf{q}(\gamma)).$$

Proof. From the definition of Z, $Z(\mathbf{q}(\gamma)) = \sum_{\mathbf{y} \in \{0,1\}^R} \prod_{v \in V^{\mathbf{y}}} q_v(\gamma) = z_{\lambda(\gamma)} \sum_{\mathbf{y} \in \{0,1\}^R} \mu_{\lambda(\gamma)}(\{\mathbf{y}\}) = z_{\lambda(\gamma)} \cdot 1$ from Proposition 1.1. □

Let $N^v, v \in V_0$ be the number of observations in the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ such that the corresponding collections $V^{\mathbf{y}_i}$ include v:

$$N^v := \#\{i : 1 \le i \le n, v \in V^{\mathbf{y}_i}\}.$$

Now we can represent the distribution μ_λ in a simplified form:

$$\begin{aligned}\prod_{i=1}^n \mu_{\lambda(\gamma)}(\{\mathbf{y}_i\}) &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{i=1}^n \prod_{v \in V^{\mathbf{y}_i}} q_v(\gamma) \\ &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{v \in V_0} \prod_{i : v \in V^{\mathbf{y}_i}} q_v(\gamma) \\ &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{v \in V_0} q_v(\gamma)^{N^v}\end{aligned}$$

which shows that $(N^v)_{v \in V_0}$ are sufficient statistics. This leads to a simple form of the Bernoulli log-likelihood function l_B in the parameters γ_i:

$$l_B(\gamma) = \frac{1}{n} \sum_{v \in V_0} N^v \log q_v(\gamma) - \log z_{\lambda(\gamma)}.$$

An interior stationary point for l_B can be found as a positive solution to a system of polynomial equations in the variables q_v, by the chain rule for derivatives. Using the definition above for $Z(\mathbf{q})$ in terms of $q_v, v \in V_0$, consider l_B in the variables $q_1, \ldots, q_c$:

$$l_B(\mathbf{q}) = \frac{1}{n} \sum_{v \in V_0} N^v \log q_v - \log Z(\mathbf{q}). \tag{1.5}$$

Let us use the notation $\mathbf{q}^{V^{\mathbf{y}}} := \prod_{v \in V^{\mathbf{y}}} q_v$. Setting $\nabla l_B = \mathbf{0}$ (assuming an interior stationary point) leads to c polynomial equations:

$$\frac{N^v}{n} = \frac{\sum\limits_{\mathbf{y} : v \in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}}}{\sum\limits_{\mathbf{y}} \mathbf{q}^{V^{\mathbf{y}}}}, \quad v \in V_0$$

which is a polynomial system with a fixed point property in the vector $\mathbf{q}$:

$$q_v = \frac{N^v}{n} \frac{\sum\limits_{\mathbf{y}} \mathbf{q}^{V^{\mathbf{y}}}}{\sum\limits_{\mathbf{y} : v \in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}}/q_v}, \quad v \in V_0.$$

This suggests the iterative method with transformation $T : R^{V_0} \to R^{V_0}$ as below from some reasonable initial point $\mathbf{q}^0$:

$$\begin{aligned}\mathbf{q}^{k+1} &= T(\mathbf{q}^k), \quad k = 0, 1, 2, \ldots \\ T(\mathbf{q})_v &= \frac{N^v}{n} \frac{\sum\limits_{\mathbf{y}} \mathbf{q}^{V^{\mathbf{y}}}}{\sum\limits_{\mathbf{y} : v \in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}}/q_v}, \quad v \in V_0.\end{aligned}$$

Obviously this would not give positive values if $\mathbf{q} = 0$ or $N^v/n = 0$, so there will be some conditions on the initial point and the data in Section 3 for it to work.

Example 1.1. Consider the binary tree (1), which has polynomial $Z(\mathbf{q}) = (1+q_3+q_4+q_1)(1+q_5+q_6+q_2)$. To find an interior stationary point for the function $l_B(\mathbf{q},\theta)$, we solve the following system for positive q_v, θ:

$$\begin{aligned}
\frac{N^1}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_1(1+q_5+q_6+q_2)\\
\frac{N^2}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_2(1+q_3+q_4+q_1)\\
\frac{N^3}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_3(1+q_5+q_6+q_2)\\
\frac{N^4}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_4(1+q_5+q_6+q_2)\\
\frac{N^5}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_5(1+q_3+q_4+q_1)\\
\frac{N^6}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)) &= q_6(1+q_3+q_4+q_1).
\end{aligned}$$

Observe that if we set $t_v := N^v/n$, and $E_{\mathbf{q}(\gamma)}(N^v/n) := \sum_{\mathbf{y}:v\in V^{\mathbf{y}}} \mathbf{q}^{V^{\mathbf{y}}}/Z(\mathbf{q})$, then the sequence $\mathbf{q}^0, \mathbf{q}^1, \mathbf{q}^2, \ldots$ can be written

$$q_v^{k+1} = q_v^k \cdot \frac{t_v}{E_{\mathbf{q}^k}(N^v/n)}.$$

This brings out some similarities with iterative proportional scaling [5].

2. Multicast model with spatial dependence: interaction model

We have described the original model of [3], and now we generalize that model to one where interaction across edges occurs. This generalization is the simplest model with spatial dependence and other models with different types of dependence structures would also be of interest.

The new interaction model has an additional parameter $\theta > 0$ affecting the probability of multiple losses and breaking the Markov property. The new model reduces to the Bernoulli model when $\theta = 1$, corresponding in a way to a temperature $t = \infty$ in a Curie-Weiss model of interaction, where $\theta = e^{1/t}$. The range of the interaction is across all edges, and values of θ greater than 1 mean that multiple losses are more likely than they would be under the Bernoulli model.

For $\mathbf{x} \in \{0,1\}^{V_0}$, let $|\mathbf{x}| := \sum_{i\in V_0} x_i$ be the number of 1's in $\mathbf{x}$. Then the notation $[|\mathbf{x}|-1]_+$ will give $|\mathbf{x}|-1$ if there are two or more 1's in $\mathbf{x}$, otherwise it will vanish. The new law $\nu_{\gamma,\theta}$ in parameters $\gamma_i > 0, i = 1,\ldots,c, \theta >$ 00, is specified by

$$\begin{aligned}
\nu_{\gamma,\theta}(\mathbf{x}) &:= h(\mathbf{x})\,\frac{\gamma^{\mathbf{x}}\theta^{[|\mathbf{x}|-1]_+}}{w_{\gamma,\theta}}\\
w_{\gamma,\theta} &= \frac{\theta - 1 + z_{\lambda(\theta\gamma)}}{\theta}
\end{aligned} \tag{2.1}$$

where the formula for the normalizing constant $w_{\gamma,\theta}$ relates to z from the Bernoulli model as follows. Consider the one-to-one reparametrization from positive γ to positive λ with inverse given by

$$\gamma_i = \lambda_i p_i(\lambda), \quad i = 1,\ldots,c.$$

Then $\lambda(\theta\gamma)$ is the vector $(\lambda_1(\theta\gamma),\ldots,\lambda_c(\theta\gamma))$ that comes from finding the λ corresponding to $(\theta\gamma_1, \theta\gamma_2, \ldots, \theta\gamma_c)$. From the Bernoulli model, we know that

$$\prod_{i=1}^{c}(1+\lambda_i(\gamma\theta)) = \sum_{\mathbf{x}\in\{0,1\}^{V_0}} h(\mathbf{x})\gamma^{\mathbf{x}}\theta^{|\mathbf{x}|}$$

and separating the case $\mathbf{x} = \mathbf{0}$ gives the formula. In terms of the odds-ratio parameters λ_i, the law ν can be written

$$\nu_{\gamma(\lambda),\theta}(\mathbf{x}) := h(\mathbf{x}) \frac{\lambda^{\mathbf{x}} \mathbf{p}(\lambda)^{\mathbf{x}} \theta^{[|\mathbf{x}|-1]_+}}{w_{\gamma(\lambda),\theta}}. \tag{2.2}$$

The parameter pair (γ, θ) is called identifiable for positive γ and positive θ if $\nu_{\gamma,\theta}(\{\mathbf{x} \in \{0,1\}^{V_0} : A\mathbf{x} = \mathbf{y}\}) = \nu_{\gamma',\theta'}(\{\mathbf{x} \in \{0,1\}^{V_0} : A\mathbf{x} = \mathbf{y}\})$ for all $\mathbf{y} \in \{0,1\}^R$ implies that $\gamma = \gamma'$ and $\theta = \theta'$, where γ and γ' are assumed to be positive in each coordinate and θ and θ' are assumed to be nonnegative real numbers. If the parameters are identifiable, then different parameters will lead to different statistical patterns and consistent estimation is possible under repeated, independent experiments. Otherwise, different parameter values may be statistically indistinguishable based on repeated experimental outcomes.

In [7] it was proved that the parameters (λ, θ) are identifiable if the tree $\mathcal{T}$ has the property that all parent nodes have at least two children.

3. Estimation and inference

In this section, we propose a numerical method for finding maximum likelihood estimates of the unknown parameter values. With the "incomplete" data $\mathbf{y}_i = A\mathbf{x}_i$ as a many-to-one function of outcomes $\mathbf{x}_i$, it seems that the EM-algorithm [6] is appropriate. The EM-algorithm is used in [2, 8, 9, 13]. The paper [9] has some interesting approximations on the speed of convergence of the EM-algorithm, which is governed by the largest eigenvalue of a transformation matrix. The EM-algorithm is complicated when applied directly and in theory is only a local optimizer. Below we present an iterative method that has some similarities with the EM-algorithm, in that it is a fixed-point argument with geometric convergence. However, it does not seem to be included in the description of the EM-algorithm, rather it is a method for the "M-step" as defined in equation (2.3) p. 4, of [6]. The method is essentially a numerical way to compute a Legendre transform in a special case, and resembles iterative proportional scaling [5] in some ways. The general theorems of [5] cannot be applied however for convergence, because the variables in the procedure are not probabilities.

Local convergence is established in Theorem 3.2, which shows the stability of the iterative procedure. The convergence to a global maximum is established under conditions in Theorem 3.3. This is analogous to the results in [14] that strengthen the convergence conclusions of the EM iteration.

The basic idea can be illustrated most simply with the example of finding the value of the odds ratio parameter λ in n Bernoulli trials, say with x successes. The objective function is $\lambda^x/(1+\lambda)^n$, and the stationary point $\hat{\lambda}$ satisfies the fixed point equation $\lambda = T(\lambda) := (x/n)(1+\lambda)$. Then with $\lambda^0 = 1$ one gets a sequence of converging approximations $1, 2x/n, x/n + 2(x/n)^2, \ldots \to x/(n-x)$, as long as $x < n$.

Let the observations for an iid sample be $\mathbf{y}_1 = A\mathbf{x}_1, \ldots, \mathbf{y}_n = A\mathbf{x}_n$. Observe that the relationship between the Bernoulli model μ_λ and the interaction model $\nu_{\lambda,\theta}$ implies the following formula:

$$\nu_{\gamma,\theta}(\{\mathbf{y}\}) = \mu_{\lambda(\theta\gamma)}(\{\mathbf{y}\}) \left[\frac{\theta I_{\mathbf{0}}(\mathbf{y}) + I_{\neq \mathbf{0}}(\mathbf{y})}{\theta - 1 + z_{\lambda(\theta\gamma)}}\right] z_{\lambda(\theta\gamma)}$$

where $\theta\gamma := (\theta\gamma_1, \ldots, \theta\gamma_c)$.

Let $N_{\mathbf{0}}$ be the number of times the vector $\mathbf{0}$ appears in the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$. The objective function for maximum likelihood estimation is the log-likelihood function l in (γ, θ) given by

$$\begin{aligned} l(\gamma,\theta) &= \frac{1}{n}\sum_{i=1}^{n} \log \nu_{\gamma,\theta}(\{\mathbf{y}_i\}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log(\mu_{\lambda(\theta\gamma)}(\{\mathbf{y}_i\})) + \frac{N_{\mathbf{0}}}{n}\log(\theta) + \log\left(\frac{z_{\lambda(\theta\gamma)}}{\theta - 1 + z_{\lambda(\theta\gamma)}}\right). \end{aligned}$$

It follows that $l(\gamma,\theta) = l_0(\theta\gamma,\theta)$, where l_0 is defined by

$$l_0(\gamma',\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \mu_{\lambda(\gamma')}(\{\mathbf{y}_i\}) + \frac{N_{\mathbf{0}}}{n}\log(\theta) + \log\left(\frac{z_{\lambda(\gamma')}}{\theta - 1 + z_{\lambda(\gamma')}}\right). \tag{3.1}$$

The objective function l_0 in (3.1) can be simplified. For each $\mathbf{y} \in Z_+^R$, let $V^{\mathbf{y}} \subset V_0$ be the collection of edge labels that are closest to the root whose failure could lead to observation $\mathbf{y}$. For example, in the binary tree (1), $V^{(1,1,1,1)} = \{1,2\},\ V^{(1,1,0,1)} = \{1,6\}$. Define polynomials $q_v,\ v \in V_0$ in variables $s_v,\ v \in V_0$ recursively by

$$\begin{aligned} q_r &:= s_r, r \in R \\ q_v &:= s_v + \prod_{w \in f^{-1}(v)} q_w. \end{aligned} \tag{3.2}$$

By the independence in the Bernoulli model,

$$\mu_{\lambda(\gamma)}(\{\mathbf{y}\}) = \sum_{A\mathbf{x}=\mathbf{y}} h(\mathbf{x})\frac{\gamma^{\mathbf{x}}}{z_{\lambda(\gamma)}} = \frac{1}{z_{\lambda(\gamma)}}\prod_{v \in V^{\mathbf{y}}} q_v(\gamma),$$

which leads to the formula for the normalizing constant $z_{\lambda(\gamma)}$ in terms of the variables q_v:

$$z_{\lambda(\gamma)} = \sum_{\mathbf{y}\in\{0,1\}^R} \prod_{v \in V^{\mathbf{y}}} q_v(\gamma).$$

Let $N^v, v \in V_0$ be the number of observations in the sample $\mathbf{y}_1,\ldots,\mathbf{y}_n$ such that the corresponding collections $V^{\mathbf{y}_i}$ include v:

$$N^v := \#\{i : 1 \le i \le n, v \in V^{\mathbf{y}_i}\}.$$

Now we can represent the distribution μ_λ in a simplified form:

$$\begin{aligned} \prod_{i=1}^{n} \mu_{\lambda(\gamma)}(\{\mathbf{y}_i\}) &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{i=1}^{n} \prod_{v \in V^{\mathbf{y}_i}} q_v(\gamma) \\ &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{v \in V_0} \prod_{i : v \in V^{\mathbf{y}_i}} q_v(\gamma) \\ &= \frac{1}{z^n_{\lambda(\gamma)}} \prod_{v \in V_0} q_v(\gamma)^{N^v}. \end{aligned}$$

This leads to a simpler form of the objective function l_0:

$$l_0(\gamma,\theta) = \frac{1}{n}\sum_{v \in V_0} N^v \log q_v(\gamma) + \frac{N_{\mathbf{0}}}{n}\log(\theta) - \log(\theta - 1 + z_{\lambda(\gamma)}). \tag{3.3}$$

The procedure to maximize l over (γ,θ) is to maximize l_0 over γ',θ and transform back:

$$\begin{aligned} (\hat\gamma',\hat\theta) &:= \texttt{arg max}_{\gamma'>0,\theta>0}\ l_0(\gamma',\theta) \\ (\hat\gamma,\hat\theta) &:= (\hat\gamma'/\hat\theta,\hat\theta). \end{aligned} \tag{3.4}$$

Define the polynomial $Z(q_1,\ldots,q_c)$ by

$$Z(\mathbf{q}) := \sum_{\mathbf{y}\in\{0,1\}^R} \mathbf{q}^{V^{\mathbf{y}}} \tag{3.5}$$

with the notation $\mathbf{q}^{V_{\mathbf{y}}} := \prod_{v\in V_{\mathbf{y}}} q_v$. An interior stationary point for $l_0(\gamma,\theta)$ can be found as a positive solution to a system of polynomial equations in the variables q_v, θ, by the chain rule for derivatives. Using the definition above for $Z(\mathbf{q})$ in terms of $q_v, v \in V_0$, consider l_0 in the variables $q_1, \ldots, q_c$:

$$l_0(\mathbf{q},\theta) = \frac{1}{n}\sum_{v\in V_0} N^v \log q_v + \frac{N_{\mathbf{0}}}{n}\log(\theta) - \log(\theta - 1 + Z(\mathbf{q})). \tag{3.6}$$

Setting $\nabla l_0 = \mathbf{0}$ leads to $c+1$ polynomial equations satisfied by the mle $(\hat{\mathbf{q}}, \hat{\theta})$:

$$\begin{aligned}\frac{N^v}{n}(Z(\mathbf{q}) + \theta - 1) &= q_v\Big(\sum_{\mathbf{y}: v\in V_{\mathbf{y}}} \mathbf{q}^{V_{\mathbf{y}}}/q_v\Big)\\ \frac{N_{\mathbf{0}}}{n}(Z(\mathbf{q}) + \theta - 1) &= \theta.\end{aligned} \tag{3.7}$$

If we define a transformation T on $\mathbf{R}\times\mathbf{R}^c$ by

$$\begin{aligned}T(\mathbf{q},\theta)_v &= \frac{N^v}{n}\frac{Z(\mathbf{q})+\theta-1}{\sum_{\mathbf{y}: v\in V_{\mathbf{y}}} \mathbf{q}^{V_{\mathbf{y}}}/q_v}, \quad 1 \le v \le c,\\ T(\mathbf{q},\theta)_{c+1} &= \frac{N_{\mathbf{0}}}{n}(Z(\mathbf{q})+\theta-1)\end{aligned} \tag{3.8}$$

then the desired optimal interior values $(\hat{\theta}, \hat{\mathbf{q}})$ satisfy the equation

$$(\hat{\mathbf{q}}, \hat{\theta}) = T(\hat{\mathbf{q}}, \hat{\theta}).$$

This leads to the iterative method with T starting with initial values $(\theta^0, \mathbf{q}^0)$:

$$(\mathbf{q}^{k+1}, \theta^{k+1}) = T(\mathbf{q}^k, \theta^k), \quad k = 0, 1, 2, \ldots$$

Example 3.1. Consider the binary tree (1). To find an interior stationary point for the function $l_0(\mathbf{q},\theta)$, we solve the following system for positive q_v, θ:

$$\begin{aligned}
\frac{N^1}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1) &= q_1(1+q_5+q_6+q_2)\\
\frac{N^2}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1) &= q_2(1+q_3+q_4+q_1)\\
\frac{N^3}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1) &= q_3(1+q_5+q_6+q_2)\\
\frac{N^4}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1) &= q_4(1+q_5+q_6+q_2)\\
\frac{N^5}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1) &= q_5(1+q_3+q_4+q_1)\\
\frac{N^6}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1) &= q_6(1+q_3+q_4+q_1)\\
\frac{N_{\mathbf{0}}}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1) &= \theta.
\end{aligned}$$

The transformation T on $(\mathbf{q}, \theta)$ is

$$\begin{aligned}
T(\mathbf{q},\theta)_1 &= \frac{N^1}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_5+q_6+q_2)\\
T(\mathbf{q},\theta)_2 &= \frac{N^2}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_3+q_4+q_1)\\
T(\mathbf{q},\theta)_3 &= \frac{N^3}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_5+q_6+q_2)\\
T(\mathbf{q},\theta)_4 &= \frac{N^4}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_5+q_6+q_2)\\
T(\mathbf{q},\theta)_5 &= \frac{N^5}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_3+q_4+q_1)\\
T(\mathbf{q},\theta)_6 &= \frac{N^6}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1)/(1+q_3+q_4+q_1)\\
T(\mathbf{q},\theta)_7 &= \frac{N_{\mathbf{0}}}{n}((1+q_3+q_4+q_1)(1+q_5+q_6+q_2)+\theta-1).
\end{aligned}$$

The solution must be transformed to $\hat{\gamma}'$ using (3.2), then again transformed to $\hat{\gamma}$ using (3.4).

To prove theoretical results about the optimization procedure, introduce new parameters $\phi := (\phi_0, \phi_1, \phi_2, \ldots, \phi_c) = (\log(\theta), \log(\mathbf{q}))$ and statistics $\mathbf{t} := (t_0 = N_{\mathbf{0}}/n, t_1 = N^1/n, \ldots, t_c = N^c/n)$. This simplifies the objective function l_0 to a standard concave form:

$$\begin{aligned}
l_0(\phi) &= \phi \cdot \mathbf{t} - \log(\zeta_\phi)\\
\zeta_\phi &:= \sum_{\mathbf{y}\in\{0,1\}^R} e^{\phi\cdot\mathbf{a}_\mathbf{y}} = Z(\mathbf{q}) + \theta - 1
\end{aligned}$$

where $\mathbf{a}_\mathbf{y} \in \{0,1\}^{\{0,1,\ldots,c\}}$ is a vector defined in terms of the standard basis vectors $\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_{c+1}$ by

$$\begin{aligned}
\mathbf{a}_\mathbf{y} &= \sum_{i\in V^\mathbf{y}} \mathbf{e}_i \quad \texttt{if } \mathbf{y} \neq \mathbf{0}\\
\mathbf{a}_\mathbf{0} &= \mathbf{e}_0 = (1,0,0,\ldots,0).
\end{aligned}$$

In the tree of Figure 1, $\zeta(\phi) = (1 + e^{\phi_1} + e^{\phi_3} + e^{\phi_4}) \cdot (1 + e^{\phi_2} + e^{\phi_5} + e^{\phi_6}) + e^{\phi_0} - 1$, and $\mathbf{a}_{(1,1,0,1)} = (0,1,0,0,0,0,1)$, $\mathbf{a}_{(0,0,0,0)} = (1,0,0,0,0,0,0)$.

Let $C \subset \mathbf{R}^{c+1}$ be the closed, convex hull of the vectors $\mathbf{a}_\mathbf{y}, \mathbf{y} \in \{0,1\}^R$. Then we have the linear equation $\mathbf{t} = A\mathbf{n}/n$ if A is the matrix with 2^R columns equal to the collection $\mathbf{a}_\mathbf{y}, \mathbf{y} \in \{0,1\}^R$, and $\mathbf{n}$ is a vector of length 2^R that counts the number of each of the 2^R types of outcomes $\mathbf{y}$ in the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$. So if the sample includes at least one of each type of vector $\mathbf{y}$, then the data of sufficient statistics $\mathbf{t}$ will be in the interior of C, as a full convex combination of the defining vectors of C.

The first result uses standard convexity theory for uniqueness.

Theorem 3.1. *If the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ includes at least one of each of the 2^R possibilities (or more generally, suppose $\mathbf{t} \in \texttt{int}\, C$) then there is a unique positive solution $(\hat{\mathbf{q}}, \hat{\theta})$ to the system* (3.7) *and a unique positive fixed point for T in* (3.8).

Proof. This follows from Theorem 9.13 of [1] with parametrization $\phi = (\log(\theta), \log(\mathbf{q}))$. □

The following result is the basic theorem on stability of the fixed point map.

Theorem 3.2. *Suppose the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ includes at least one of each of the 2^R possibilities. If $(\mathbf{q}^0, \theta^0)$ is sufficiently close to the optimal values $(\hat{\mathbf{q}}, \hat{\theta})$, if $\hat{\theta}$ is sufficiently close to 1, and if the observed failure counts are sufficiently small, then the sequence $T^k(\mathbf{q}^0, \theta^0)$ converges at a geometric rate to the positive fixed point $(\hat{\mathbf{q}}, \hat{\theta})$.*

Proof. Let $\bar{T}(\phi) := \log(T(e^\phi))$ be the transformation T in the ϕ variables, which with (3.8) becomes

$$\begin{aligned}\bar{T}(\phi)_v &= \log(N^v/n) - \log\left(\frac{\sum_{\mathbf{y}: v \in V^{\mathbf{y}}} \mathbf{q}^{V^y}/q_v}{\zeta_\phi}\right) \\ &= \phi_v + \log(t_v) - \log \partial_{\phi_v} \log(\zeta_\phi) \\ &= \phi_v - \log\left(\frac{\partial_{\phi_v} \log(\zeta_\phi)}{\partial_{\phi_v} \log(\zeta_{\hat{\phi}})}\right),\end{aligned}$$

using the optimality condition $\partial_{\phi_v} \log(\zeta_{\hat{\phi}}) = t_v$ for interior $\mathbf{t}$ in the last line. The above, together with the formula for the last variable $\log(\theta) = \phi_0$ gives the final representation for the transformation T in terms of the canonical parameters ϕ:

$$\bar{T}(\phi) = \phi + \log \mathbf{t} - \log(\nabla \log \zeta_\phi) = \phi - \log\left(\frac{\nabla \log \zeta_\phi}{\nabla \log \zeta_{\hat{\phi}}}\right).$$

It is enough to show that the eigenvalues of the derivative $D_{\hat{\phi}}\bar{T}$ are strictly less than one in absolute value.

The derivative $D_{\hat{\phi}}\bar{T}$ is a matrix with rows $\nabla \bar{T}(\hat{\phi})_v$ indexed by edge labels $v \in V_0$. Now

$$\begin{aligned}\partial_{\phi_w} \bar{T}(\hat{\phi})_v &= \delta_{vw} + 0 - \frac{\partial_{\phi_w} \partial_{\phi_v} \log \zeta_{\hat{\phi}}}{\partial_{\phi_v} \log \zeta_{\hat{\phi}}} \\ &= \delta_{vw} - \frac{\Sigma_{\hat{\phi}}(v, w)}{t_v},\end{aligned}$$

where $\Sigma_{\hat{\phi}}(v, w)$ is the covariance of coordinates v, w in the vectors $\{\mathbf{a}_\mathbf{y}\}$ with probabilities in the exponential family $p_\phi(\mathbf{a}_\mathbf{y}) = \frac{e^{\phi \cdot \mathbf{a}_\mathbf{y}}}{\zeta_\phi}$. Thus

$$\begin{aligned}D_{\hat{\phi}}\bar{T} &= I - A_\mathbf{t} \Sigma_{\hat{\phi}} \\ A_\mathbf{t} &:= \begin{pmatrix} 1/t_0 & 0 & 0 & \ldots & \ldots \\ 0 & 1/t_1 & 0 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1/t_c \end{pmatrix}.\end{aligned}$$

Therefore it is sufficient to show that the eigenvalues of $A_\mathbf{t}\Sigma_{\hat{\phi}}$ are in $(0, 2)$, which is the same as having the eigenvalues of $\sqrt{A_\mathbf{t}}\Sigma_{\hat{\phi}}\sqrt{A_\mathbf{t}}$ in $(0, 2)$. Now $\sqrt{A_\mathbf{t}}\Sigma_{\hat{\phi}}\sqrt{A_\mathbf{t}}$ is the covariance matrix for $\sqrt{A_\mathbf{t}} \cdot \mathbf{a}$, or in other words $\sqrt{A_\mathbf{t}}\Sigma_{\hat{\phi}}\sqrt{A_\mathbf{t}}\,(v, w) = \mathtt{Cov}_{\hat{\phi}}(\mathbf{a}(v)/\sqrt{t_v}, \mathbf{a}(w)/\sqrt{t_w})$, where $\mathbf{a}$ is a random vector of length $c + 1$ taken from the distribution p_ϕ. This matrix is positive definite as a nondegenerate covariance matrix.

Consider first the diagonal entries $\mathtt{Var}_{\hat{\phi}}(\mathbf{a}(v)/\sqrt{t_v})$. Since the vectors $\mathbf{a}_\mathbf{y}$ have entries $\mathbf{a}_\mathbf{y}(v)$ in $\{0, 1\}$, $\mathtt{Var}_{\hat{\phi}}(\mathbf{a}(v)) = E_{\hat{\phi}}(\mathbf{a}(v) - t_v)^2 \le E_{\hat{\phi}}(\mathbf{a}(v)^2) \le E_{\hat{\phi}}(\mathbf{a}(v)) = t_v$. Thus the diagonal entries are less than 1.

Consider next the off-diagonal entries $\mathtt{Cov}_{\hat{\phi}}(\mathbf{a}(v)/\sqrt{t_v}, \mathbf{a}(w)/\sqrt{t_w})$. Recall that the index "0" is special and corresponds to no loss at all in the network, unlike $v = 1, 2, \ldots, c$. If $\hat{\theta} = 1$ then $\hat{\phi}_0 = 0$ so there is an

independence property and

$$\begin{aligned}\mathtt{Cov}_{\hat{\phi}}(\mathbf{a}(v)/\sqrt{t_v}, \mathbf{a}(w)/\sqrt{t_w}) &= -\sqrt{t_v}\sqrt{t_w}, \quad \text{if } v, w \geq 1, \text{ and } v \in d(w) \text{ or } w \in d(v) \\ &= 0, \quad \text{if } v, w \geq 1, \text{ and neither } v \in d(w) \text{ nor } w \in d(v) \\ &= -\sqrt{t_0}\sqrt{t_w}, \quad \text{if } v = 0 \text{ and } w \geq 1.\end{aligned}$$

Now by the method of Gersgorin discs (p. 146, [11]), the eigenvalues of $A_{\mathbf{t}}\Sigma_{\hat{\phi}}$ will be in $(0, 2)$ if $\sum_{w=0, w\neq v}^{c+1} |-\sqrt{t_v}\sqrt{t_w}| < 1$ for each $v = 0, 1, \ldots, c$. This will follow if the values of $t_w, w = 1, 2, \ldots, c$ are sufficiently small, in particular if $\sum_{w=1}^{c} \sqrt{t_w} < 1$ and $\sqrt{t_w} < \frac{1}{2}$, $w = 1, 2, \ldots, c$.

Then by continuity, the eigenvalues of $A_{\mathbf{t}}\Sigma_{\hat{\phi}}$ will be in $(0, 2)$ if the same condition holds on $\mathbf{t}$ and if $\hat{\phi}_0$ is close to 0. □

Theorem 3.2 should hold with much weaker assumptions, but it is difficult to bound the eigenvalues when the interaction parameter $\hat{\theta}$ is not close to 1. Below, we have a final global convergence result that justifies the iteration procedure starting from any positive initial point.

Theorem 3.3. *Suppose the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ includes at least one of each of the 2^R possibilities (or more generally suppose $\mathbf{t} \in \mathtt{int}\, C$). If $(\mathbf{q}^0, \theta^0)$ is any positive initial value, let $(\mathbf{q}^k, \theta^k) = T^k(\mathbf{q}^0, \theta^0)$. If the sequence $\{\log(\theta^k)\}$ is bounded, then $(\mathbf{q}^k, \theta^k)$ converges to the optimal fixed point $(\hat{\mathbf{q}}, \hat{\theta})$ as $k \to \infty$.*

Proof. From (3.8), we have

$$\theta^{k+1} = \frac{N_{\mathbf{0}}}{n}(Z(\mathbf{q}^k) + \theta^k - 1), \ k = 0, 1, 2, \ldots$$

which implies that $\theta^{k+1} - \frac{N_{\mathbf{0}}}{n}(\theta^k - 1) = \frac{N_{\mathbf{0}}}{n} Z(\mathbf{q}^k)$. If $M \geq \theta^{k+1}$ for all $k = 0, 1, 2, \ldots$, then $M + \frac{N_{\mathbf{0}}}{n} \geq Z(\mathbf{q}^k)$. This implies that $\{\mathbf{q}^k\}$ is bounded.

Let $(\mathbf{q}^{n_k}, \theta^{n_k})$ be any convergent subsequence to a limit $(\mathbf{q}, \theta)$. If $\mathbf{q} = 0$ ($q_v = 0$ for all $v \in V_0$), then by continuity of the last coordinate of T in θ, it follows that $\theta = \frac{N_{\mathbf{0}}}{n}\theta$. This is impossible since $0 < N_{\mathbf{0}} < n$ and θ^k is bounded above 0, by assumption. Thus some $q_v > 0$, so $Z(\mathbf{q}) > 1$. This implies that $\theta > 0$ and $q_v > 0$ for each $v \in V_0$, so the limit is positive. By uniqueness from Theorem 3.1, $(\mathbf{q}, \theta) = (\hat{\mathbf{q}}, \hat{\theta})$, the unique, positive, optimal solution.

Since any convergent subsequence converges to the unique solution, the entire original sequence must converge to the unique solution, because it is bounded. □

The convergence conclusion of Theorem 3.3 should hold under weaker conditions, but we have been unable to prove a result with no assumption of boundedness.

4. Conclusions

We have proposed an iterative method for solving a polynomial system of equations to find the maximum likelihood estimators for a problem of network reliability, and we have proven that the method has essential convergence qualities. In practice, it seems to be stable and efficient. The conditions on the data under which the method is proved to work are natural, and may possibly be further weakened. On the down side, the conditions require a sample size on the order of $R\,2^R$ (where R is the number of receiver nodes) to to get a tractable likelihood function to which convexity theory can be applied. A further area of research would be Bayesian methodology, which can make likelihood functions well-behaved with smaller sample sizes.

The primary competing method for parameter estimation in network tomography is the EM-algorithm. In theory it is not guaranteed to find the global maximum of the likelihood function unless some extra conditions are verified [14]. In practice it seems to have worked well for the Bernoulli model in several studies, including [9, 13]. For the model that we study here, it seems to be more complicated than the proposed method because the EM-algorithm involves a series of expectations and maximizations, rather than repeated evaluation of a

rational function. It is hard to make clear and useful comparisons because most studies have been done with relatively small simulated examples. In fact the reparametrization to ϕ variables in an exponential family makes a quasi-Newton method possible, but a quasi-Newton method will be relatively complex because of complicated components like the normalizing constant ζ_ϕ.

It seems valuable as a general strategy to formulate optimal solutions to dependent network problems as real solutions of polynomial systems. This opens the way to using new methods from computational real algebraic geometry. This may be the key to handling more realistic and complex network dependencies. For example, one may try new methods of semi-definite programming for solving polynomial systems, as described in Parrilo and Sturmfels [12] and implemented in SOS Tools. On the other hand, it seems unlikely that existing polynomial homotopy methods would be efficient. These numerical system solvers start with solutions to an easy "nearby" system (say the Bernouli model) and repeatedly adjust them to get all solutions to the desired polynomial system. Since homotopy methods find all complex solutions rather than just nonnegative real solutions, the work involved may be unreasonable for large networks where there will be thousands of complex solutions. A comparative numerical study would be interesting.

The failure model we have studied is worthwhile because it generalizes the Bernoulli model to allow for dependence across the network, and it also has a method of exact solution. Other more refined or sophisticated models may ultimately be more practical. For example, a model with more local dependence may seem more suitable. Local dependence is difficult to model, first because interactions may occur in "neighborhoods" that have little to do with geography or physical distance, and second because the mathematical problem of getting identifiable parameters is quite difficult the way the data is collected. This is certainly an area for interesting work that would be a subtle blend of theory and practice. We would hope that some of the techniques and ideas from the present paper would carry over to any useful model with spatial dependence.

References

[1] O. Barndorff-Nielsen, *Information and Exponential Families.* Wiley, New York (1978).

[2] T. Bu, N. Duffield, F. Lo Presti and D. Towsley, Network tomography on general topologies. *Proc. ACM Sigmetrics 2002*, Marina Del Ray, June 15–19 (2002).

[3] R. Cáceres, N.G. Duffield, J. Horowitz, D. Towsley and T. Bu, Multicast-based inference of network internal characteristics: accuracy of packet loss estimation. *IEEE Trans. Inform. Theory* **45** (2000) 2462–2480.

[4] M. Coates, A.O. Hero, R. Nowak and B. Yu, Internet tomography. *IEEE Signal Processing Magazine* **19** (2002) 47–65.

[5] J.N. Darroch and D. Ratcliff, Generalized iterative scaling for log-linear models. *Ann. Math. Stat.* **43** (1972) 1470–1480.

[6] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39** (1997) 1–38.

[7] I.H. Dinwoodie and E. Mosteig, Statistical inference for network reliability with spatial dependence. *SIAM J. Discrete Math.* **16** (2003) 663–674.

[8] N. Duffield, J. Horowitz, D. Towsley, W. Wei and T. Friedman, Multicast-based loss inference with missing data. *IEEE J. Selected Areas Communications* **20** (2002) 700–713.

[9] C. Ji and A. Elwalid, Measurement-based network monitoring and inference: scalability and missing information. *IEEE J. Selected Areas Communications* **20** (2002) 714–725.

[10] G. Liang and B. Yu, Maximum pseudo-likelihood estimation in network tomography. *IEEE Trans. Signal Process.* **51** (2003) 2043–2053.

[11] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities.* Allyn and Bacon, Boston (1964).

[12] P. Parrilo and B. Sturmfels, Minimizing polynomial functions. `http://xyz.lanl.gov/abs/math.OC/0103170` (2002).

[13] Y. Tsang, M. Coates and R. Nowak, Passive network tomography using EM algorithms. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, Utah* **3** (May 2001) 1469–1472.

[14] C.F. Jeff Wu, On the convergence of the EM algorithm. *Ann. Statist.* **11** (1983) 95–103.

[15] B. Xi, G. Michailidis and V.N. Nair, *Estimating network internal losses using a new class of probing experiments.* University of Michigan Department of Statistics Technical Report 397 (2003).