

Maximum Entropy Models: Convergence Rates and Application in Dynamic System Monitoring*

Gang Liang and Bin Yu
 University of California at Berkeley
 E-mail: {liang,binyu}@stat.Berkeley.EDU
 Nina Taft
 Intel Research Berkeley
 E-mail: {nina.taft@intel.com}

Abstract—The principle of maximum entropy (ME) is widely used as a statistical inference tool in many fields such as computer vision, econometrics, and natural language processing. In practice, iterative scaling type of algorithms are applied for fitting maximum entropy models. In this paper, we first investigate the convergence rates of the generalized iterative scaling (GIS) and improved iterative scaling (IIS) algorithms under the parameterization of log-linear models. Second, we study the problem of dynamic system monitoring with partial active measurements by applying the maximum entropy principle at each time point. Based on a linear time-variant dynamic system denoted by hidden Markov models, an information-based randomization measurement scheme is derived to select the most informative hidden states to measure. Such a dynamic monitoring setup is validated on a problem of origin-destination (OD) traffic matrix inference in network tomography.

Index Terms—dynamic systems, hidden Markov models, improved iterative scaling, generalized iterative scaling, maximum entropy models, network tomography, origin-destination traffic matrix.

I. INTRODUCTION

The principle of maximum entropy (ME) has a long history in physics, and it was first proposed as a general inference procedure by Jaynes [6]. This principle can also be derived from some reasonable axioms if we wish the final solution to obey [13]. One appealing property of the ME model is its flexibility to incorporate additional pragmatic features. For parameter estimation, iterative scaling type of algorithms have been developed for solving maximum entropy problems with linear constraints. These algorithms are all simple to implement and practically fast to converge.

In this paper, we first investigate the convergence rate of several iterative scaling procedures, namely generalized iterative scaling (GIS) [4] and improved iterative

scaling (IIS) [11]. We show that iterative scaling algorithms converge in exponential rates. The convergence rate calculation is not only of theoretic interest, but also of practical value for fitting ME models. Large ME models, such as ones in natural language processing [1], [12], may involve features of order of several thousands. The convergence rate of iterative scaling algorithms can be slow for these large models due to the small exponents in their exponential rates. Our convergence result might help better understand iterative scaling algorithms as well as boost the search for more efficient algorithms.

Second, we study the problem of dynamic system monitoring by applying the principle of maximum entropy. In practice, a large number of dynamic systems can be denoted by hidden Markov models (HMM), where observed variables are linear aggregations of internal hidden states. It is common that the number of observed variables is much less than that of hidden states, so the problem becomes an ill-posed linear inversion problem, i.e., the system is unidentifiable if no other constraint is introduced. There is a rich literature on ill-posed linear inversion problems with regularized approaches being most commonly used [5], [15]. In this paper, a different approach from a partial measurement perspective is advocated. Considering a discrete time dynamic system, we contend that the original ill-posed problem will become much well-posed even if only a tiny fraction of hidden states can be measured directly at each time point. Here, we assume that all hidden states are observable, but measurement costs or some other technical difficulties make a full measurement approach infeasible. This assumption is true for many real-domain applications. Under a scenario of a linear time-variant system we will speculate later, we show that iterative scaling algorithms approximately converge to the minimum mean square error (MSE) estimate given the true parameter, which gives a rationale to the

*It is a revised version of the original submission.

use of the maximum entropy principle in our dynamic systems. In order to gain as much information as possible through direct measurements, the key is to find the most informative hidden states to measure. In this paper, we only consider an extreme case of measuring just one hidden state at each time point, and a randomization-based measurement selection rule is proposed to select a hidden state to measure. Our approach is validated on an origin-destination traffic matrix estimation problem in network tomography [3], and yields good results showing the potential of our approach in tracking internal states of dynamic systems.

The paper is organized as follows: at first, we will review ME models, as well as the GIS and IIS algorithms. Then, we proceed to show the convergence rates of both algorithms. Next, we present our framework for dynamic system monitoring, and the measurement selection rule under a linear time-variant scenario. Finally, our approach is applied to an origin-destination traffic matrix estimation problem in network tomography. Proofs of two theorems can be found in the Appendix.

II. MAXIMUM ENTROPY MODELS

Maximum entropy (ME) models, also known as log-linear, Gibbs models, take the following parametric form:

$$p(x) = \frac{1}{Z_\Lambda} e^{\sum_{i=1}^I \lambda_i f_i(x)}, \quad (1)$$

where $\Lambda = (\lambda_1, \dots, \lambda_I)'$ is the unknown parameter vector and Z_Λ is a normalizing constant. The real-valued functions $f_i(x)$ are usually called features in machine learning community, and below we use $f = (f_1, \dots, f_I)'$ to denote the feature vector. Given i.i.d data $\{x_1, \dots, x_n\}$, let $\tilde{p}(x)$ be the empirical distribution. It is easy to show that the ME model (1) is the dual problem of

$$\begin{aligned} \max - \sum_x p(x) \log(p(x)) \\ \text{s.t. } E_p(f_i) = E_{\tilde{p}}(f_i) \end{aligned} \quad (2)$$

where $E_p(f_i) = \sum_x p(x) f_i(x)$ is the expectation with respect to distribution p , and $E_{\tilde{p}}$ is the expectation operator with respect to the empirical distribution.

A number of algorithms have been proposed for estimating the parameters of ME models. Among those frequently used are the generalized iterative scaling (GIS) [4] and improved iterative scaling (IIS) [11] algorithms. The GIS algorithm, proposed by Darroch and Ratcliff (1972), is a simple yet efficient procedure for estimating ME model parameters. Procedures of the GIS algorithm are described in Algorithm 1, in which the constant C is defined as $C = \max_x \sum_i f_i(x)$.

Algorithm 1 Generalized iterative scaling algorithm

```
Initialize  $\Lambda^{(0)} = (0, \dots, 0)'$ ;
for  $t = 1, \dots$ , till converge do
   $\Lambda^{(t+1)} = \Lambda^{(t)} + \frac{1}{C} \left( \log E_{\tilde{p}}(f) - \log E_{p_{\Lambda^{(t)}}}(f) \right)$ ;
end for
```

Della Pietra et al. [11] re-formulate the convergence proof of the GIS algorithm through an auxiliary function argument. In order to improve the convergence rate of the GIS algorithm, they propose an IIS algorithm as specified in Algorithm 2. Now the constant C is replaced by $f^\#(x) = \sum_i f_i(x)$, which depends on x . In doing so, a sharper lower bound for the auxiliary function is obtained; hence, a faster convergence rate is achieved. On the other hand, the updating equation in Algorithm 2 can not be solved explicitly, and a numerical line search has to be applied to find $\delta_i^{(t)}$, which may potentially exacerbate the performance of the algorithm.

Algorithm 2 Improved iterative scaling algorithm

```
Initialize  $\Lambda^{(0)} = (0, \dots, 0)'$ ;
for  $t = 1, \dots$ , till converge do
  Solve  $E_{\tilde{p}}(f_i) = \sum_x \tilde{p}(x) e^{\delta_i^{(t)} f^\#(x)} f_i(x)$  for  $\delta_i^{(t)}$ ;
   $\Lambda^{(t+1)} = \Lambda^{(t)} + \delta_i^{(t)}$ ;
end for
```

A. Convergence Rate of Iterative Scaling Algorithms

1) *Application of Ostrowski's Theorem:* Algorithm 1 shows that the GIS algorithm can be denoted as an iterative procedure $\Lambda^{(t+1)} = g(\Lambda^{(t)})$ with

$$g(\Lambda) = \Lambda + \frac{1}{C} (\log E_{\tilde{p}}(f) - \log E_p(f)),$$

where the final parameter estimation $\hat{\Lambda}$ is an attraction point of this iterative procedure, i.e., $\hat{\Lambda} = g(\hat{\Lambda})$. For the ME model in (2), the attraction point is unique because its log-likelihood function is concave. The rate of the convergence of such an iterative algorithm is governed by Ostrowski's theorem (Ostrowski, 1960).

Definition 1: Fundamental eigenvalue: for an $n \times n$ symmetric matrix A , its fundamental eigenvalue is defined as

$$\lambda_A = \max_{i=1}^n |\lambda_i|,$$

where λ_i 's are eigenvalues of A .

Criterion (Ostrowski, 1960) Assume $g(\cdot)$ is differentiable at the neighborhood of a fixed point ζ , and let λ_g be the fundamental eigenvalue of the Jacobian matrix of g at ζ . For an iterative algorithm $\zeta_{n+1} = g(\zeta_n)$, a sufficient

condition for ζ to be a point of attraction is that $\lambda_g < 1$, and a necessary condition is that $\lambda_g \leq 1$. Moreover, if ζ is a attraction point, the geometric convergence rate of the iterative algorithm is λ_g , i.e.,

$$\overline{\lim}_{n \rightarrow \infty} \frac{\|\zeta_{n+1} - \zeta\|}{\|\zeta_n - \zeta\|} = \lambda_g. \quad (3)$$

Theorem 1: For the GIS algorithm, the Jacobian matrix of its iterative function $g(\lambda)$ at $\hat{\Lambda}$ is:

$$I_n - \frac{1}{C} \text{diag} \left(\frac{1}{E_{\hat{p}}(f)} \right) \text{cov}_{p_{\hat{\Lambda}}}(f). \quad (4)$$

and the rate of convergence is determined by the largest fundamental eigenvalue of this matrix.

Similarly, applying Ostrowski's theorem to the IIS algorithm, we have the follow theorem:

Theorem 2: For the IIS algorithm, the Jacobian matrix of its iterative function $g(\lambda)$ at $\hat{\Lambda}$ is:

$$I_n - \text{diag} \left(\frac{1}{E_{p_{\hat{\Lambda}}}(ff\#)} \right) \text{cov}_{p_{\hat{\Lambda}}}(ff). \quad (5)$$

and the rate of convergence is determined by the largest fundamental eigenvalue of this matrix.

We have $E_{\hat{\Lambda}}(f) = E_{\hat{p}}(f)$ at the convergence point $\hat{\Lambda}$ because all linear constraints are satisfied. Furthermore, we may see that both GIS and IIS algorithms are scale-variant, i.e., to scale a feature by multiplying a constant will change its convergence rate.

For large ME models with thousands of features, the converge of iterative scaling may be slow when the fundamental eigenvalue is very close to 1. Many new algorithms have been developed for speeding up the parameter search, for example the fast iterative scaling (FIS) [7] recently proposed. The convergence rate calculation might lead to better understanding of iterative algorithms, and boost the development of faster algorithms.

III. DYNAMIC SYSTEM MONITORING

Dynamic system monitoring is of interest in many real-world applications, such as monitoring patients' brain blood flow continuously, highway traffic monitoring and control, or estimating network origin-destination traffic flows. Usually dynamic systems can be modeled by state-space models with internal states change over time, and our observations are functionals of hidden states (mostly linear aggregations possibly with noise). In this paper, we consider the following discrete time linear system:

$$\begin{aligned} X^{(t+1)} &= DX^{(t)} + E + \epsilon^{(t)} \\ Y^{(t+1)} &= AX^{(t+1)}, \end{aligned} \quad (6)$$

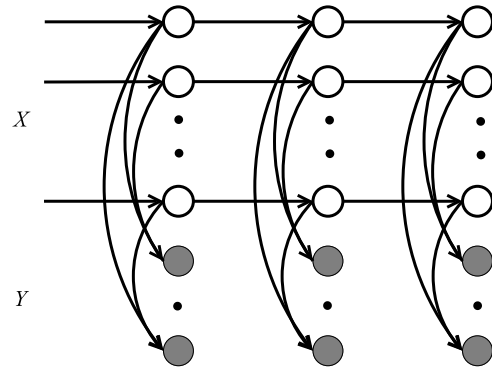


Fig. 1. A graphical illustration of a hidden Markov model.

where $X^{(t)} \in \mathcal{R}^J$ denotes the system state at time t , and $Y^{(t)} \in \mathcal{R}^I$ is the observed random variable vector with usually $J \ll I$. In (6), the evolution of the hidden state $X^{(t)}$ is governed by an auto-regression. The error term $\epsilon^{(t)}$ is assumed to be independent multivariate Gaussian variable, which is dependent on the current system state accounting for the fact that large internal states usually have large variations in most dynamic systems. In the above equation, both A, D are constant matrices with A being known, and E is an unknown constant vector. The matrix A is usually sparse, i.e., with many entries being zeros. If D is further assumed to be diagonal, then all components of state variable $X^{(t)}$ are independent of each other.

The formalism of graphical models provides a unifying framework to describe such a hidden Markov dynamic system. Fig. 1 illustrates the HMM specified in (6). At each time point, each arrow between $X^{(t)}$ and $Y^{(t)}$ corresponds to a nonzero entry in matrix A . In the plot, each hidden state is evolving independently, which is a consequence of assuming D being a diagonal matrix. It is also worth noting that this dynamic system is time-variant because we assume the innovation term $\epsilon^{(t)}$ depends on the current hidden states. In the following derivation, we consider a simplified version of (6):

$$\begin{aligned} X^{(t+1)} &= X^{(t)} + \epsilon^{(t)} \\ Y^{(t+1)} &= AX^{(t+1)}, \end{aligned} \quad (7)$$

with

$$\epsilon^{(t)} \sim N(0, \phi \text{diag}(X^{(t)})), \quad (8)$$

where ϕ is an unknown parameter. Both models (6) and (7) are similar but with the auto-regression term being removed in (7) for easy exposition in later sections.

As we assume $I \ll J$, i.e., the dimension of observed variable $Y^{(t)}$ is much less than that of the hidden state $X^{(t)}$, this is a typical ill-posed linear inversion problem,

and some constraints have to be introduced to ensure the identifiability of the model. There is a rich literature devoted to this topic from regularization points of views [5], [15]. In this paper, we approach the problem from the perspective of maximum entropy but with partial measurements at each time interval. Here, we assume that all hidden states are observable, but measurement costs or some other technical difficulties make a full measurement approach infeasible; this assumption is true for many real-domain applications. For the discrete time dynamic system (6), we contend that the original ill-posed problem will become much well-posed even if only a tiny fraction of hidden states are directly measured at each time point (we may choose a different set of hidden states to measure at different time points). In order to gain more information, the key is to find the most informative hidden states to measure. In this paper, we only consider an extreme case of selecting only one hidden state to measure at each time point.

Before we start the measurement design, let us consider the following lemma:

Lemma 1: For the dynamic system (7), if we assume the mean vector μ (i.e., $X^{(t-1)}$) is known, then the minimum mean square error (MSE) estimate of $X^{(t)}$ given $Y^{(t)}$ is

$$E(X^{(t)}|Y^{(t)}) = \mu + \Sigma A'(A\Sigma A')^{-1}(Y^{(t)} - A\mu), \quad (9)$$

which is independent of parameter ϕ . This conditional expectation is also exactly the solution to the weighted least-square estimate with square root weights:

$$\min \sum_i \left(\frac{X_i^{(t)} - \mu_i}{\sqrt{\mu_i}} \right)^2 \text{ subject to } AX^{(t)} = Y^{(t)}.$$

As pointed out in [16], we have

$$D(X^{(t)}/N||\mu/N) \approx \frac{1}{N} \sum_i \left(\frac{X_i^{(t)} - \mu_i}{\sqrt{\mu_i}} \right)^2,$$

where N is the total sum of hidden states.

Lemma 1 justifies the use of maximum entropy principle for the dynamic system specified in (7). It says that the maximum entropy estimate (minimum Kullback-Leibler divergence estimate) also approximately gives the minimum mean square error (MSE) estimate when $\mu = X^{(t-1)}$ is known. In practice, parameter μ is unknown hence replaced by the previous hidden state estimate $\hat{X}^{(t-1)}$. Similar results hold for (6) if we assume that

$$\text{Var}(\epsilon^{(t)}) \propto DX^{(t)} + E.$$

This lemma also suggests the partial measurement approach in part. The iterative scaling algorithms are reasonable to apply because it approximately finds the MSE estimate if good estimation of system state X at previous stage is obtained. But the estimate may finally drift away if the iterative procedure is repeatedly applied. In order to capture the dynamics of the underlying system, some new information has to be introduced, and this is achieved through a partial measurement of hidden states in this paper.

In order to realize an online dynamic system monitoring scheme, a fast algorithm is critical. In this paper, we use iterative scaling algorithms. As we have shown in the previous sections, iterative scaling algorithms converge in exponential rates, which are fast enough in practice for most applications. Furthermore, in dynamic system monitoring scheme proposed in this paper, we use the estimation $\hat{X}^{(t-1)}$ obtained from the previous step as the initial value for the iterative scaling algorithms at the current time point t . If the dynamic system is smooth enough (which is true for most systems), then the starting value is already in the neighborhood of the estimate at current time point. It further speeds up the convergence of iterative scaling algorithms.

A. Measurement design

The key in our approach to the problem of dynamic system monitoring is to design a good information-theoretic based scheme in choosing the most informative hidden state to measure. We may also view the dynamic system (7) as a Gaussian process, where the standard deviation (SD) of a normal random variable corresponds to its entropy. Intuitively, the entropy implies uncertainty, so we should always choose a hidden state which has the largest conditional entropy based on the previous step estimation to drive the uncertainty of the whole system down. Here, we briefly describe our algorithm for measurement selection. At time t , we need to determine which hidden state to measure at next time point $t+1$ based on the previous one-step prediction result because of the Markovian property. Assuming that the state of the dynamic system $X^{(t-1)}$ at previous stage $t-1$ is known, we have ($\mu = X^{(t-1)}$)

$$X^{(t)} \sim N(\mu, \phi\mu), \text{ and } Y^{(t)} = AX^{(t)}, \quad (10)$$

then the conditional distribution of $X^{(t)}$ given $Y^{(t)}$ is

$$\text{Var}(X^{(t)}|Y^{(t)}) = \Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma \quad (11)$$

If we choose the hidden state with the largest conditional standard deviation, i.e., the largest entropy, to measure, such a greedy approach is one-step optimal

when $X^{(t-1)}$ is known. First, the hidden state $X^{(t-1)}$ will never be known to us in practice. Furthermore, this greedy approach will lead to a selection scheme in which measurements will mostly focus on a small number of hidden states with large conditional standard deviation. On the long run, it will not be able to adapt to the underlying dynamic system because of the ill-posedness of the problem. In order to remedy this problem, we propose a randomized selection scheme such that a hidden state is chosen to be measured with probability proportional to its conditional standard deviation. The use of conditional SD instead of other quantities such as variance is due to a simple fact:

given two independent mean zero normal random variables u, v , then the ratio of $P(|u| > |v|)$ to $P(|u| < |v|)$ is proportional to their SDs.

Such an approach is similar to a minimax decision rule, but note that the argument is only approximately true because given $Y^{(t)}$, components of $X^{(t)}$ are no longer independent.

In summary, the proposed dynamic monitoring algorithm with randomization measurement scheme is given as below:

Algorithm 3 Dynamic Monitoring: Randomization

```

Initialize hidden state  $\hat{X}^{(0)}$  with  $\mathbf{1}$ ;
for each time interval  $t$  do
  1. Use a randomized scheme to determine the
     most informative hidden state to measure;
  2. Replace  $\hat{X}^{(t-1)}$  with new observations
     accordingly;
  3. Obtain  $\hat{X}^{(t)}$  by applying iterative scaling on
      $\hat{X}^{(t-1)}$  to match  $Y^{(t)}$  with observed components
     being fixed.
end for

```

In the above algorithm, $\hat{X}^{(0)}$ is set to be 1 component-wise, which implies that the algorithm starts from a maximum entropy estimation.

IV. EXAMPLE: OD TRAFFIC ESTIMATION

Origin-Destination (OD) traffic matrices quantify the volume of traffic flows between all possible origin-destination (OD) pairs of a network, which are important inputs for a variety of network traffic engineering tasks, including capacity planning, load balancing, and traffic provisioning.

Two type of approach exist to obtain traffic matrices: direct and indirect. Direct approaches for OD matrix inference are via some router softwares, such as Netflow

supported by Cisco routers, to monitor traffic flows directly. But in practice, estimating full traffic matrices through direct measurements is prohibitively expensive on large operational IP networks; hence, it is of interest to estimate traffic matrices through network link counts or some other readily available information, which can be obtained through SNMP (Simple Network Management Protocol).

For indirect approaches, the network link traffic is linear aggregations of OD traffic. Because the number of links are much less than that of OD pairs, the OD traffic matrix inference problem is an ill-posed linear inversion problem. Many efforts have been devoted to the OD traffic inference problem recently. Vardi [14] proposes a Poisson model assuming i.i.d. Poisson distributions for the OD traffic byte counts on a general network topology. With real network data, Cao et al. [2] revise the Poisson assumption to propose a Gaussian model. Medina et al. [9] propose a logit choice model in order to incorporate additional information into OD traffic modelling. Zhang et al. [16] assume a source-destination independent model, which is equivalent to an gravity model widely used in transportation system, and a minimum mutual information (MMI) method with regularization is used to match internal link counts while shrinking to the gravity model estimation. In practice, the time-varying nature of the network traffic is evident and is partly treated in Cao et al. [2] by a local likelihood approach.

A. Model and Data

In this paper, model (7) is used to describe the network OD traffic, which is a slight modification to the Gaussian traffic model proposed in Cao et al. [2]. The difference is that the Gaussian model is now embeded into a dynamic system framework, which reflects and treats the time course of the network traffic in a more natural way. Now in (7), the matrix A is a known $I \times J$ routing matrix determined by the network topology and the routing table at each internal node. For most networks, elements of A only take 0-1 entries. $X^{(t)}$ is the traffic count vector between all pairs of network nodes at time interval t , and $Y^{(t)}$ is our observed link traffic counts, which is a linear aggregation of OD traffic $X^{(t)}$.

Our validation data come from the Sprint European PoP (Points of Presence) network. In this network, each node is a PoP, and each internal link corresponds to aggregated links connecting two given PoPs. After aggregation, this PoP network has 13 nodes and 18 links. The data we use are collected from Netflow measurements, routing tables and other routing configuration files, i.e.,

we have the “ground truth” for this network. Due to the limitations of SNMP, link traffic data obtained from SNMP polling may be lost in transition or be incorrect. In order to avoid the inconsistency in data collection, a consistent set of traffic, topology and link measurement data are derived from the flow level measurements $X^{(t)}$ instead.

B. Experiment Results

In this paper, we use two methods to assess and compare the performance of estimation results. The first method is to draw the cumulative distribution function (CDF) of the absolute values of relative errors, i.e., the absolute error relative to its OD traffic count. The weight probability of each relative error is proportional to its OD traffic count. Usually large relative errors occur at small traffic matrix elements; with a weight on relative errors, the CDF plot will automatically focus on large OD traffic. The second method is to compute the average relative errors for large OD traffic flows. The percentage of total traffic we choose is 80%, i.e., we find a cutoff point such that all OD pairs with traffic larger than the cutoff point accounts for 80% of total traffic, then we compute average relative errors for these OD traffic elements.

For comparison purpose, we also implement a simple random choice (SRC) scheme [?] for selecting hidden states to measure at each time interval. Given a preset integer k , the SRC method is to randomly choose k OD pairs to measure at each time interval. Fig. 2 gives the CDF plot of relative errors for three different schemes, and the average relative errors are reported in Table I. From the results, we can see that the dynamic system monitoring with a small number of active measurements is successful even for an SRC scheme. Here we simply recall some numbers obtained from some other OD estimation methods: for the same dataset, another two competitive methods, pseudo-IPF [8] and gravity-MMI [16], yield similar results with an average relative error of around 30%. We can also see that the randomization measurement scheme improves the estimation result: its performance is comparable to an SRC(3) scheme, i.e., randomly choosing three OD pairs to measure each time.

V. CONCLUSIONS

In the paper, we derive the convergence rates of two iterative scaling algorithms. We also investigate the problem of dynamic system monitoring using the principle of maximum entropy with active measurements. An information-based randomization measurement scheme is derived based on a linear time-variant dynamic system.

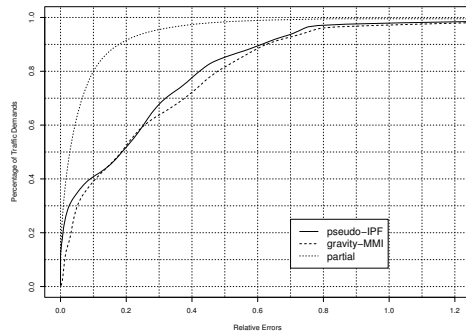


Fig. 2. The CDF plot of relative errors of three different measurement schemes.

TABLE I
AVERAGE RELATIVE ERROR RATES OF THREE METHODS:
RANDOMIZATION, SRC(1) AND SRC(3).

Algorithm	Avg Relative Error Rate
SRC(1)	7.26%
SRC(3)	5.32%
Randomization	5.27%

Our experiment demonstrates that such an approach, with a few (even only one) active measurements, makes the originally ill-posed problem well behaved. But there are still many open questions. For instance, given a precision, how many measurements will suffice for monitoring a given dynamic system. Moreover, whether the randomization is near optimal in the sense of a minimax estimation. Here, we defer all these questions to our future work.

VI. ACKNOWLEDGEMENTS

We would like to thank Sprint Advanced Laboratory, Nina Taft, Dina Antonio Nucci and Dina Papagianaki for letting us have access to their dataset. We would also like to thank Anukool Lakhina for preparing the dataset and many helpful discussions.

VII. APPENDIX

Proof of Theorem 1

Proof: For GIS algorithm, the Jacobian matrix of its iterative function $g(\lambda)$ is

$$J(\Lambda) = \mathcal{I}_n - \frac{1}{C} \frac{\partial \log E_{p_\Lambda}(f)}{\partial \Lambda}. \quad (12)$$

Consider the ij th element of the second part, we have

$$\frac{\partial \log E_{p_\Lambda} f_i}{\partial \lambda_j} = \frac{1}{E_{p_\Lambda} f_i} \frac{\partial E_{p_\Lambda} f_i}{\partial \lambda_j}$$

and

$$\frac{\partial E_{p_\Lambda} f_i}{\partial \lambda_j} = \text{cov}_{p_\Lambda}(f_i, f_j),$$

so the Jacobian matrix J can be written as

$$J(\Lambda) = \mathcal{I}_n - \frac{1}{C} \text{diag} \left(\frac{1}{E_{p_\Lambda}(f)} \right) \text{cov}_{p_\Lambda}(f), \quad (13)$$

At the attraction point $\hat{\Lambda}$, $E_{p_{\hat{\Lambda}}}(f) = E_{\bar{p}}(f)$, and

$$J|_{\Lambda=\hat{\Lambda}} = I_n - \frac{1}{C} \text{diag} \left(\frac{1}{E_{\bar{p}}(f)} \right) \text{cov}_{p_{\hat{\Lambda}}}(f). \quad (14)$$

Proof of Theorem 2

Proof: For the IIS algorithm, the updating vector δ satisfies a functional constraint $u(\Lambda, \delta) = 0$, where

$$u_i(\Lambda, \delta) = E_{\bar{p}}(f_i) - \sum_x p_\Lambda(x) e^{\delta_i f^\#(x)} f_i(x). \quad (15)$$

It implies that

$$\frac{\partial u(\Lambda, \delta)}{\partial \Lambda} + \frac{\partial u(\Lambda, \delta)}{\partial \delta} \frac{\partial \delta}{\partial \Lambda} = 0,$$

i.e.,

$$\frac{\partial \delta}{\partial \Lambda} = - \left(\frac{\partial u(\Lambda, \delta)}{\partial \delta} \right)^{-1} \frac{\partial u(\Lambda, \delta)}{\partial \Lambda} \quad (16)$$

Similarly, we can compute part by part that

$$\left. \frac{\partial u(\Lambda, \delta)}{\partial \delta} \right|_{\Lambda=\hat{\Lambda}} = \text{diag}(E_{p_{\hat{\Lambda}}}(f f^\#)) \quad (17)$$

and

$$\left. \frac{\partial u(\Lambda, \delta)}{\partial \Lambda} \right|_{\Lambda=\hat{\Lambda}} = \text{cov}_{p_{\hat{\Lambda}}}(f). \quad (18)$$

So the convergence rate of IIS algorithm is determined by its Jacobian matrix at $\Lambda = \hat{\Lambda}$:

$$J|_{\Lambda=\hat{\Lambda}} = I_n - \text{diag} \left(\frac{1}{E_{p_{\hat{\Lambda}}}(f f^\#)} \right) \text{cov}_{p_{\hat{\Lambda}}}(f). \quad (19)$$

REFERENCES

- [1] A.L. Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.
- [2] J. Cao, D. Davis, S. Vander Wiel, and B. Yu. Time-varying network tomography: router link data. *Journal of American Statistics Association*, 95:1063–1075, 2000.
- [3] M. Coates, A. Hero, R. Nowak, and B. Yu. Internet tomography. *Signal Processing Magazine*, 19(3):47–65, 2002.
- [4] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.
- [5] M. Hanke and P.C. Hansen. Regularization methods for large-scale problems. *Surveys on Mathematics for industry*, 3:253–315, 1993.
- [6] E.T. Janyes. Information theory and statistical mechanics I. *Physical Review*, 101:620–630, 1957.
- [7] R. Jin, R. Yan, J. Zhang, and Hauptman A. A fast iterative scaling for conditional exponential models. In *The twentieth international conference on machine learning*, 2003.
- [8] G. Liang and B. Yu. Maximum pseudo-likelihood estimation in network tomography. *IEEE Transactions on Signal Processing*, 51(8):2043–2053, August 2003.
- [9] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic matrix estimation: Existing techniques and new directions. In *ACM Sigcomm*, Pittsburg, USA, August 2002.
- [10] A.M. Ostrowski. *Solution of equations and systems of equations*. New York: Academic Press, 1960.
- [11] S. Della Pietra, V. Della Pietra, and J. Lafferty. Induce features of random fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.
- [12] R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228, 1996.
- [13] J.E. Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.
- [14] Y. Vardi. Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91:365–377, 1996.
- [15] G. Wahba. Spline models for observational data. In *SIAM*, Philadelphia, 1990.
- [16] Y. Zhang, M. Roughan, C. Lund, and D. Donoho. An information-theoretic approach to traffic matrix estimation. In *ACM SIGCOMM*, 2003.