

Maximum Entropy Models: Convergence Rates and Applications in Dynamic System Monitoring

Gang Liang and Bin Yu¹
 University of California at Berkeley
 {liang,binyu}@stat.berkeley.edu

Nina Taft
 Intel Research Berkeley
 nina.taft@intel.com

Abstract — We investigate the convergence rates of generalized iterative scaling (GIS) and improved iterative scaling (IIS) algorithms for fitting maximum entropy (ME) models. We also study a particular linear dynamic system monitoring with partial active measurements. An information-theoretic based measurement scheme is derived to select informative hidden states, which is validated on a problem of origin-destination matrix estimation for internet traffic.

I. INTRODUCTION

Iterative scaling algorithms are widely used for fitting ME models with linear constraints. The convergence rate calculation is not only of theoretic interest, but also of practical value for fitting ME models. In this talk, we derive the convergence rates for two popular algorithms: GIS [2] and IIS [3].

An ME model is ideal for dynamic system monitoring due to its flexibilities and ease to fit, but its performance may degenerate over time. We consider a scenario in which active measurements can be made to improve the estimate performance, but the number of measurements is limited due to practical constraints. Based on a particular time-varying hidden Markov model (HMM), we propose an information-theoretic based scheme in choosing informative states to measure, and contend that the ME estimate can be greatly improved even with only a tiny fraction of selected hidden states being measured.

II. CONVERGENCE RATE OF ITERATIVE SCALING

Consider the maximum entropy model:

$$p(x) \propto \exp\left(\sum_i \lambda_i f_i(x)\right), \quad i = 1, \dots, I, \quad (1)$$

where $\Lambda = (\lambda_1, \dots, \lambda_I)'$ is an unknown parameter, and $f = (f_1, \dots, f_I)'$ is the feature vector. Both GIS and IIS are iterative algorithms for fitting (1): they initialize $\Lambda^{(0)} = (0, \dots, 0)'$, then keep updating the estimate by $\Lambda^{(t+1)} = \Lambda^{(t)} + \delta^{(t)}$ till it converges. Let \tilde{p} be the empirical distribution, $C = \max_x \sum_i f_i(x)$, and $f^\#(x) = \sum_i f_i(x)$. For GIS, $\delta^{(t)} = C^{-1} \left(\log E_{\tilde{p}}(f) - \log E_{p_{\Lambda^{(t)}}}(f) \right)$, while for IIS we have to solve $E_{\tilde{p}}(f_i) = \sum_i p_{\Lambda^{(t)}}^{(i)}(x) f_i(x) \exp(\delta_i^{(t)} f^\#(x))$ to get $\delta^{(t)}$.

Theorem 1 *The Jacobian matrices of GIS and IIS algorithms are respectively*

$$I_n - C^{-1} \text{diag} \left(E_{\tilde{p}}(f)^{-1} \right) \text{cov}_{p_{\Lambda}}(f), \quad (2)$$

and

$$I_n - \text{diag} \left(E_{p_{\Lambda}}(f f^\#)^{-1} \right) \text{cov}_{p_{\Lambda}}(f). \quad (3)$$

Their convergence rates are determined by the largest fundamental eigenvalues of these matrices.

¹This work was partially supported by NSF grant FD01-12731 and ARO grant DAAD19-01-1-0643.

III. DYNAMIC SYSTEM MONITORING

Consider a particular time-varying HMM dynamic system:

$$X^{(t+1)} = X^{(t)} + \epsilon^{(t)}, \quad Y^{(t+1)} = AX^{(t+1)}, \quad (4)$$

where $X^{(t)}$ ($\in \mathcal{R}^J$) is the hidden state and $Y^{(t)}$ ($\in \mathcal{R}^I$) is observation at time t with usually $J \ll I$, and A is a known matrix. The error term $\epsilon^{(t)} \sim N(0, \phi \text{diag}(X^{(t)}))$, accounting for that large internal states tend to have large variations. Our goal is to estimate $X^{(t)}$ through $Y^{(t)}$. Usually the ME principle can be used to estimate $X^{(t)}$, and the following theorem further justifies the use of the ME principle.

Theorem 2 *If $X^{(t-1)}$ is known, the ME estimate of $X^{(t)}$ is approximately the minimum square error estimate.*

But in practice, $X^{(t-1)}$ is unknown (replaced by its estimate), and more important the problem itself is ill-posed due to $J \ll I$, so the performance of ME estimate may degenerate over time. Here, assuming all hidden states are observable, we approach the problem from an active partial measurement perspective: the key is to pick the most informative hidden states to measure to improve the estimation performance.

Intuitively, the entropy (exactly the standard deviation (SD) here) implies uncertainty, so hidden states with large conditional entropies are desirable to measure to reduce the system uncertainty. But such a greedy approach tends to select only a small number of hidden states with large SD. In order to remedy this problem, a randomization scheme is proposed such that a hidden state is chosen with a probability proportional to its conditional SD, which is similar to a minimax decision rule. We call our method “InfoRand”, i.e., information-based randomization.

The InfoRand approach is applied to a problem of origin-destination (OD) matrix estimation, which is a key problem in network engineering. The model (4) we use is a modification of the Gaussian traffic model in [1] with the time course also being considered. We have a real dataset from Sprint European network, in which all OD traffic is observable, but it is an expensive operation. Here, we report the relative error rates of the InfoRand approach: 5.27%. Two other cutting-edge methods (namely pseudo-IPF and gravity-MMI) yield an average relative error around 30%. The partial measurement approach is a very promising method to pursue.

REFERENCES

- [1] J. Cao, D. Davis, S. Vander Wiel, and B. Yu. Time-varying network tomography: router link data. *Journal of American Statistics Association*, 95(452):1063–1075, 2000.
- [2] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.
- [3] 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.