Pattern Maximum Likelihood Estimation of Finite-State Discrete-Time Markov Chains

Shashank Vatedka
Dept. of Electrical Communication Engineering
Indian Institute of Science
Bengaluru, India
shashank@ece.iisc.ernet.in

Pascal O. Vontobel
Dept. of Information Engineering
The Chinese University of Hong Kong
Shatin, NT, Hong Kong
pascal.vontobel@ieee.org

***Abstract*—We study the problem of estimating the pattern maximum likelihood (PML) distribution for time-homogeneous discrete-time Markov chains (DTMCs). The PML problem for memoryless sources has been well studied in the literature and we propose an extension of the same for DTMCs. For memoryless sources, Acharya et al. have shown that plug-in estimators obtained from the PML estimate yield good estimates for symmetric functionals of the distribution. We show that this holds for the PML estimate of DTMCs as well. Finally, we express the PML estimate for DTMCs as the double minimization of a certain free energy function and discuss some mean-field approximations to approximate the PML estimate efficiently.**

I. Introduction

Consider the following setup where we observe samples from a time-homogeneous discrete-time Markov chain (DTMC) over a finite state space/alphabet $\mathcal{S}$. We assume that the DTMC is irreducible, i.e., the chain can go from any state to any other state in a finite number of steps. We know that $|\mathcal{S}| = k$, but we have no knowledge about the initial distribution or the transition kernel. By observing n samples (random walk of length n), we want to estimate the structure of the transition kernel, i.e., find the transition kernel up to a relabeling of the states. In other words, we want to find the state transition diagram, but we are not interested in knowing the exact correspondence between the states of the DTMC and the nodes in the transition diagram. This is of interest if we wish to estimate *symmetric* properties of the transition kernel, i.e., those properties of the kernel that are invariant to a relabeling of the states. The entropy rate of the DTMC is a typical example of a symmetric property, and so are the eigenvalues of the transition probability matrix. As a special case, this approach may be used to estimate several parameters of a graph (such as the eigenvalues of the Laplacian matrix, the degree distribution, and so on) or the structure of the graph from a random walk on the graph.

This work is inspired by the well-studied problem of finding the pattern/profile maximum likelihood estimate for memoryless sources [1], which we briefly describe here. Suppose that we have an i.i.d. sequence of random variables which are distributed according to a probability mass function (pmf) p over a set $\mathcal{S}$. Suppose that we observe a length-n sequence $\mathbf{x}$. The *pattern* associated with $\mathbf{x}$, $\boldsymbol{\psi}(\mathbf{x}) \triangleq \psi_1\psi_2\ldots\psi_n$, is the string obtained by replacing each symbol in $\mathbf{x}$ by its order of appearance. Let us define this formally. For every x which appears in $\mathbf{x}$, we define the *index* of x to be 1 more than the number of unique symbols that were observed before the first occurrence of x in $\mathbf{x}$. For example, if $\mathbf{x} = success$, then the index of s is 1, the index of u is 2, that of c is 3, and e has an index of 4. The pattern of $\mathbf{x}$ is the length-n string of positive integers obtained by replacing each symbol of $\mathbf{x}$ by its index. For example, if $\mathbf{x} = success$, then $\boldsymbol{\psi}(\mathbf{x}) = 1233411$. The pattern is invariant to a relabeling of the alphabet. For instance, the strings *miss* and *puff* both have the pattern 1233.

The pattern maximum likelihood (PML) estimate of a memoryless source is that distribution (pmf) on $\mathcal{S}$ which maximizes the probability of occurrence of the pattern $\boldsymbol{\psi}(\mathbf{x})$. This also yields the maximum likelihood estimate of the probability multiset [1]. The study of the PML estimate was inspired by the problem of universal compression of i.i.d. sources over unknown alphabets in [2]. It was shown that even an optimal universal compression scheme for i.i.d. sequences of unknown sources over large alphabets could result in large redundancy, whereas one can obtain low redundancy if we only compress the pattern of the observed sequence. This led to the search for efficient schemes to obtain the PML estimate [3], [4], [5]. It was recently shown in [6] that using a plug-in estimator obtained from the PML estimate yields an order-optimal sample complexity for estimating certain symmetric properties of distributions.

Inspired by the results on PML for memoryless sources, we study the PML problem for the simplest sources with memory: DTMCs. The results in [2] were extended in [7] to find the redundancy of optimal universal compression schemes for DTMCs over unknown alphabets. The authors of [7] showed that unlike the memoryless case, the redundancy for compressing the pattern of a DTMC can be unbounded for large alphabets. However, they showed that for hidden Markov models where the number of states is not too large, one can achieve low redundancy. There has also been recent work on automata and graph compression [8]. Also relevant is the work

This work was mainly done when the first author was an intern at the Dept. of Information Engineering, The Chinese University of Hong Kong, where he was supported by CUHK Direct Grant 4055039 and the Institute of Network Coding. The work of the first author was also supported in part by a Tata Consultancy Services research fellowship.

in [9], which studies compression of unlabeled graphs, or what they call graph structures. The authors give a compression scheme which is shown to be asymptotically optimal for Erdős-Rényi graphs. There is also a recent approach by [10] to estimate and compress graphs using random walks. We address the following question in the spirit of [9]: Can we infer the structure of the graph[1] from a random walk starting from a randomly chosen initial state? This is a special case of the more general problem of finding the structure of a finite-state DTMC from a length-n random walk that we study here.

In this article, we define the PML problem for finite-state DTMCs. We will assume that the state space/alphabet is known. We introduce the notion of a label-invariant representation of the transition kernel (what we call a canonical transition matrix), and show that the PML estimate is actually the ML estimate of the canonical transition matrix. Using the results of [6], we show that plug-in estimators obtained from the PML estimate can efficiently estimate symmetric properties of the transition kernel. We then pose the PML computation problem as the double minimization of a certain free energy function and discuss some mean-field approximations for the same.

II. The PML Problem for DTMCs

Suppose that we observe a length-n sequence $\mathbf{x} = x_1x_2\dots x_n$ of an irreducible time-homogeneous DTMC on $\mathcal{S}$. We only know $k \triangleq |\mathcal{S}|$. More generally, we can assume that we know an upper bound, k, for $|\mathcal{S}|$. Without loss of generality, we can assume that $\mathcal{S} = [k] \triangleq \{1, \dots, k\}$. Unless mentioned otherwise, we will always assume that the initial distribution is uniform over $\mathcal{S}$, i.e., $\Pr[X_1 = x] = \frac{1}{k}$ for all $x \in \mathcal{S}$. If the transition probability kernel is $p_{xy} \triangleq \Pr[X_{t+1} = y | X_t = x]$, then we see that the probability of observing $\mathbf{x}$ is

$$\Pr[\mathbf{X} = \mathbf{x}] = \frac{1}{k} \prod_{t=1}^{n-1} p_{x_t x_{t+1}}. \qquad (1)$$

Note that the probability in (1) can be rewritten as

$$\Pr[\mathbf{X} = \mathbf{x}] = \frac{1}{k} \prod_{(x,y)\in\mathcal{S}\times\mathcal{S}} p_{xy}^{|\{t\in[n-1]\,:\,x_t=x, x_{t+1}=y\}|}. \qquad (2)$$

We define $\boldsymbol{\psi}(\mathbf{x})$ to be the pattern of $\mathbf{x}$. Moreover, we define the $k \times k$ *multiplicity matrix* $M = (\mu_{ij})_{i,j}$ associated with $\boldsymbol{\psi}$ as follows: $\mu_{ij} \triangleq |\{t : \psi_t = i, \psi_{t+1} = j\}|$. In other words, μ_{ij} counts the number of times the transition (i, j) is present in the pattern $\boldsymbol{\psi}$.

The *pattern probability* is obtained by averaging (2) over all possible relabelings of the alphabet, i.e.,

$$\mathbb{P}(\boldsymbol{\psi}|p) \triangleq \frac{1}{k} \sum_{\sigma} \prod_{i=1}^{k} \prod_{j=1}^{k} p_{\sigma(i)\sigma(j)}^{\mu_{ij}}, \qquad (3)$$

where the above summation is over all possible permutations of $[k]$. In comparison, the pattern probability for an i.i.d. sequence with pmf p is given [4] by

$$\mathbb{P}_{\text{i.i.d.}}(\boldsymbol{\psi}|\mathsf{p}) \triangleq \sum_{\sigma} \prod_{j} \mathsf{p}_{\sigma(j)}^{\mu_j}.$$

We are ultimately interested in a label-invariant estimate of the transition probability matrix. To ensure that such an estimate is unique, we define the following notion of a *canonical transition probability matrix*. Given two $k \times k$ matrices A and B, we say that[2] $A > B$ if A is lexicographically greater than B, i.e., $(A)_{11} > (B)_{11}$, or $(A)_{11} = (B)_{11}$ and $(A)_{12} > (B)_{12}$, or $(A)_{1l} = (B)_{1l}$ for all $l \leq m$ and $(A)_{1,m+1} > (B)_{1,m+1}$, and so on. For example, if

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}, \ B = \begin{bmatrix} 0 & 9 \\ 21 & 10 \end{bmatrix} \ \text{and} \ C = \begin{bmatrix} 0 & 9 \\ 21 & 3 \end{bmatrix},$$

then, $A > B > C$.

Given a matrix M, we define

$$\mathcal{T}(M) \triangleq \{T'MT : T \text{ is a permutation matrix}\},$$

where T' denotes the transpose of T. A $k \times k$ stochastic matrix M is said to be *canonical* if $M = \max \mathcal{T}(M)$.[3] Suppose

$$M = \begin{bmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{bmatrix}.$$

Then,

$$\mathcal{T}(M) = \left\{ M_1 = \begin{bmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{bmatrix}, M_2 = \begin{bmatrix} 0.7 & 0.3 \\ 0.8 & 0.2 \end{bmatrix} \right\}.$$

In the above example, M_2 is canonical, whereas M_1 is not. In other words, M is canonical if no permutation matrix T yields a $T'MT$ which is lexicographically greater than M.

A careful observation reveals that if the entries of M are all distinct (which would be the case with probability 1 if all entries are i.i.d. according to some continuous distribution), then the computation of $\max \mathcal{T}(M)$ can be performed using $O(k)$ search operations (each over sets of size $O(k)$) and $O(k)$ permutation operations (of the states). However, for an arbitrary $k \times k$ matrix, there is no known polynomial-time algorithm to compute this maximum.

The canonical matrix corresponding to M is defined as $\max \mathcal{T}(M)$. The canonical matrix corresponding to M is essentially a label-invariant representation of M. Let $\mathcal{C}$ denote the set of all $k \times k$ canonical row-stochastic matrices. Inspired by [1], [4], we define the *pattern maximum likelihood transition probability* distribution to be

$$p_{\text{PML}}^{(\boldsymbol{\psi})} = \arg\max_{p\in\mathcal{C}} \mathbb{P}(\boldsymbol{\psi}|p). \qquad (4)$$

An interesting property is that the PML distribution is the ML estimate of the canonical transition kernel. The reader is directed to Appendix A for more details.

[1] Essentially, we want to estimate the graph up to an isomorphism/relabeling of the vertices.

[2] Throughout, $(a_{ij})_{i,j}$ denotes the $k \times k$ matrix whose (i, j)th entry is a_{ij}. Furthermore, the (i, j)th entry of a matrix A is $(A)_{ij}$.

[3] We can define a canonical matrix in a number of ways. For convenience, we have defined it using the lexicographic order. Our only objective is to ensure that the PML estimate is unique.

III. Estimation of Parameters of the Markov Chain

We say that a functional f of the transition kernel p is *symmetric* if it is invariant to a relabeling of the states. There are several symmetric functionals of DTMCs that are of practical interest. A common example is the entropy rate of an ergodic DTMC, given by

$$\mathcal{H}(p) = -\sum_{i,j} \lambda_i \cdot p_{ij} \log p_{ij},$$

where λ is the stationary distribution.

Suppose that we are interested in estimating the symmetric property/functional $f(p)$ from a random walk of length n starting from a random initial state. To do this, one could simply use the ML estimator for $f(p)$. However, if we are interested in estimating different functions, we would have to compute the ML estimate for each function. This may not always be feasible since finding the ML estimate may be computationally hard and efficient approximations may not be known. Another approach is to find the PML estimate $p_{\mathrm{PML}}^{(\boldsymbol{\psi})}$ and plug this in f, i.e., compute $f\big(p_{\mathrm{PML}}^{(\boldsymbol{\psi})}\big)$. Though suboptimal, this simple plug-in estimator approach can be used to estimate any symmetric functional of p. Moreover, we will show that such plug-in estimates obtained from the PML estimate can perform nearly as well as an optimal estimator for f.

The plug-in approach can also be used to infer properties of an unlabeled graph from a random walk starting from an arbitrary vertex. Several useful parameters of the graph such as the degree distribution, the eigenvalues of the adjacency matrix, and so on are symmetric properties. For example, the degree distribution, which we denote by $\{a_i : 1 \le i \le n\}$, is given by $a_i = |\{j : p(ij) > 0\}|$.

Another commonly used (and general) technique to estimate functionals of p is to first find the ML estimate of the transition kernel (this is called the sequence maximum likelihood or SML estimate), p_{SML}, and then find the plug-in estimate $f(p_{\mathrm{SML}})$. Here p_{SML} is defined as follows. If $\mu_{ij}^{(\mathbf{x})} = |\{l : x_l = i, x_{l+1} = j\}|$, then for all $i, j \in [k]$,

$$p_{\mathrm{SML}}(i,j) \triangleq \begin{cases} 1/k & \text{if } \mu_{ij'}^{(\mathbf{x})} = 0 \text{ for all } j' \in [k] \\ c_i \mu_{ij}^{(\mathbf{x})} & \text{otherwise.} \end{cases} \tag{5}$$

where the proportionality constants $\{c_i : 1 \le i \le k\}$ ensure that $(p_{\mathrm{SML}}(i,j))_{i,j}$ is a row-stochastic matrix.[4] Our estimate of f is then $f(p_{\mathrm{SML}})$. We call this the plug-in estimator derived from the SML estimate.

Let us now turn our attention to the plug-in estimator based on p_{PML}, i.e., $f(p_{\mathrm{PML}}^{(\boldsymbol{\psi}(\mathbf{x}))})$. The natural question to ask is the following: How suboptimal is the plug-in estimator derived from the PML estimate when compared to the best estimator for $f(p)$? We will answer this question in terms of the size of the alphabet and the number of samples. The relationship that we derive is exactly along the lines of [6] for memoryless sources.

Let $\mathcal{Z}^{(n)}$ denote the set of all length-n patterns. From [1], we have

$$|\mathcal{Z}^{(n)}| \le \min\left\{ e^{3\sqrt{n}}, \binom{n+k-1}{k-1} \right\}. \tag{6}$$

Recall that the probability of observing the pattern $\boldsymbol{\psi}$ is

$$\mathbb{P}(\boldsymbol{\psi}|p) = \frac{1}{k}\sum_{\sigma} \prod_{(i,j)} p_{\sigma(i)\sigma(j)}^{\mu_{ij}}, \tag{7}$$

where μ_{ij} is the number of times the transition (i,j) occurs in the pattern $\boldsymbol{\psi}$. Since we are interested in symmetric properties of p, the pattern $\boldsymbol{\psi}(\mathbf{x})$ is a sufficient statistic to estimate f from $\mathbf{x}$. We now state the following proposition, which follows, *mutatis mutandis* from [6, Theorem 2].

Proposition 1. *Let f be any symmetric functional of p. Suppose there exists an estimator $\hat{f}$ for $f(p)$ that takes as input the pattern of a length-n sequence $\mathbf{x}^{(n)}$, and has the following property: For every $\epsilon > 0$, $\delta > 0$ and transition probability distribution p, there exists a positive integer N such that*

$$\Pr\Big[|f(p) - \hat{f}(\psi(\mathbf{X}^{(n)}))| \ge \epsilon\Big] < \delta \tag{8}$$

for all $n \ge N$. Then,

$$\Pr\Big[|f(p_{\mathrm{PML}}^{(\boldsymbol{\psi}(\mathbf{X}^{(n)}))}) - f(p)| \ge 2\epsilon\Big] < \delta \cdot |\mathcal{Z}^{(n)}| \tag{9}$$

for all $n \ge N$.

Proof. See Appendix B. □

Using (6), we see that if the probability of estimation error of the optimal estimator for f decays exponentially in n, then the probability of estimation error of the plug-in estimator derived from the PML estimate also decays exponentially in n.

IV. Variational Approaches

We now discuss some techniques to efficiently approximate the PML estimate. It is known that computing the PML estimate for memoryless sources is a hard problem, as it is equivalent to maximizing the permanent over a certain class of matrices [4], [5]. Similarly, computing the PML estimate for Markov chains is also difficult. In [11], using techniques from [12], the permanent was expressed as a minimum of a certain free energy function. However, this minimization is computationally intractable, but several approximations [12] exist. Most notable is the Bethe approximation, since one can use low-complexity belief propagation algorithms to arrive at the estimate. We direct the interested reader to [12] and references therein for more details regarding the variational approach and free energy approximations and to [11] for details regarding the Bethe approximation of the permanent. Following [4], [5] for memoryless sources, we will formulate

[4] Note that the estimate so obtained may not yield a canonical TPM, and a relabeling of states may be required to ensure this. However, for the rest of this article, we will omit this technicality, but it must be understood that a relabeling is needed to make the obtained transition kernel canonical. Of course, the relabeling does not matter if we are interested in estimating symmetric properties.

(4) as a problem of double minimization of a certain free energy function. We will then study some mean-field approximations for the PML estimate.

Let us denote the class of all $k \times k$ permutation matrices by $\mathcal{K}$, and let $\mathcal{P}$ be the set of all pmfs on $\mathcal{K}$. We can rewrite the pattern probability (3) as $\mathbb{P}(\boldsymbol{\psi}|p) = Z$, where[5] Z denotes the partition function

$$Z \triangleq \sum_{\boldsymbol{\sigma} \in K} g(\boldsymbol{\sigma}) \tag{10}$$

and $g(\boldsymbol{\sigma}) \triangleq \prod_{i,l} \prod_{j,m} p_{lm}^{\mu_{ij}\sigma_{il}\sigma_{jm}}$. Let[6] $\beta \in \mathcal{P}$. The Gibbs free energy function is defined as

$$\begin{aligned} &F_{\mathrm{G}}(\,\cdot\,; p, \boldsymbol{\psi}) : \mathcal{P} \to \mathbb{R} \\ &F_{\mathrm{G}}(\beta; p, \boldsymbol{\psi}) \triangleq U_{\mathrm{G}}(\beta; p, \boldsymbol{\psi}) - H_{\mathrm{G}}(\beta), \end{aligned} \tag{11}$$

where U_{G} denotes the Gibbs average energy function and is given by

$$\begin{aligned} U_{\mathrm{G}}(\beta; p, \boldsymbol{\psi}) &= \log k - \sum_{\boldsymbol{\sigma} \in \mathcal{K}} \beta(\boldsymbol{\sigma}) \log \Big(\prod_{i,j,l,m} p_{lm}^{\mu_{ij}\sigma_{il}\sigma_{jm}} \Big) \\ &= \log k - \sum_{\boldsymbol{\sigma} \in \mathcal{K}} \sum_{i,j,l,m} \beta(\boldsymbol{\sigma}) \log \Big(p_{lm}^{\mu_{ij}\sigma_{il}\sigma_{jm}} \Big), \end{aligned}$$

and H_{G} is the Gibbs entropy function, which is defined as

$$H_{\mathrm{G}}(\beta) \triangleq - \sum_{\boldsymbol{\sigma} \in \mathcal{K}} \beta(\boldsymbol{\sigma}) \log \beta(\boldsymbol{\sigma}). \tag{12}$$

With this, the minimum of F_G over all $\beta \in \mathcal{P}$ equals $-\log Z = -\log \mathbb{P}(\boldsymbol{\psi}|p)$. Following [4], [5], the pattern maximum likelihood estimate can be written as

$$p_{\mathrm{PML}}^{(\boldsymbol{\psi})} = \arg \min_{p \in \mathcal{C}} \min_{\beta \in \mathcal{P}} F_{\mathrm{G}}(\beta; p, \boldsymbol{\psi}). \tag{13}$$

Since the above minimization is computationally hard, we will resort to approximations of the above. The Bethe approximation in particular is a popular approach, due to the fact that the sum-product algorithm (SPA) can often be used to efficiently compute the minimizer of the Bethe free energy function [12]. This was explored in detail in [4], [5] for finding the PML estimate of memoryless sources. This was aided by the success of the Bethe approximation for computing the permanent in [11], where a suitable factor graph was designed and it was shown that the SPA on this factor graph admits useful convergence and correctness guarantees. We tried a similar approach for our problem, but have so far been unable to design a good factor graph for which the SPA is computationally efficient and at the same time yields a good estimate. We found that the computational complexity grows very fast ($O(k^4)$) and the SPA algorithm becomes difficult to implement for k larger than about 15. Hence, we study mean-field approximations, which give more computationally efficient estimators.

A. The Traditional Mean-Field (MF) Approximation

Variational approaches typically approximate the Gibbs free energy function by either relaxing the global function, and/or restricting the class of pmfs (which is β in this case) over which we find the minimum [12]. The traditional MF approximation is obtained by restricting β to be a product distribution [12]. We call the resulting PML estimate the traditional MF PML estimate (TMFPML), and we will show that this reduces to the SML estimate defined in (5). We will also introduce another MF approximation, which we call the mean-field PML estimate (MFPML) in the next subsection.

We begin the derivation of TMFPML by rewriting the partition function/pattern probability in (10) in a suitable fashion:

$$\mathbb{P}(\boldsymbol{\psi}|p) = \frac{1}{k} \sum_{\boldsymbol{\sigma} \in \{0,1\}^{k \times k}} 1_{\mathcal{K}}(\boldsymbol{\sigma}) \prod_{i,j,l,m} p_{lm}^{\mu_{ij}\sigma_{il}\sigma_{jm}}, \tag{14}$$

where $1_{\mathcal{K}}(\boldsymbol{\sigma})$ is a function that takes the value 1 if $\boldsymbol{\sigma} \in \mathcal{K}$ and 0 otherwise. The mean-field free energy function is obtained by replacing β in (11) by a product distribution. Let $\boldsymbol{\beta} = \{(\beta_{il}(0), \beta_{il}(1)) : \beta_{il}(0), \beta_{il}(1) \geq 0,\ \beta_{il}(0) + \beta_{il}(1) = 1,\ \forall 1 \leq i, l \leq k\}$. Then, $\beta(\boldsymbol{\sigma}) = \prod_{ij} \beta_{ij}(\sigma_{ij})$, and so

$$\begin{aligned} &F_{\mathrm{TMF}}(\boldsymbol{\beta}; p, \boldsymbol{\psi}) \\ &= - \sum_{\boldsymbol{\sigma} \in \{0,1\}^{k \times k}} \Big(\big(\prod_{i,l} \beta_{il}(\sigma_{il}) \big) \log \big(1_{\mathcal{K}}(\boldsymbol{\sigma}) \prod_{i,j,l,m} p_{lm}^{\mu_{ij}\sigma_{il}\sigma_{jm}} \big) \Big) \\ &\qquad + \sum_{i,l} \sum_{\sigma_{il}=0}^{1} \beta_{il}(\sigma_{il}) \log \beta_{il}(\sigma_{il}) + \log k. \end{aligned} \tag{15}$$

The double minimization of (15) over $\boldsymbol{\beta}$ and p gives us the traditional mean-field PML estimate, i.e.,

$$p_{\mathrm{TMFPML}}^{(\boldsymbol{\psi})} = \arg \min_{p \in \mathcal{C}} \min_{\boldsymbol{\beta}} F_{\mathrm{TMF}}(\boldsymbol{\beta}; p, \boldsymbol{\psi}). \tag{16}$$

Lemma 1. *For every $\boldsymbol{\psi}$, we have $p_{\mathrm{TMFPML}}^{(\boldsymbol{\psi})} = \max \mathcal{T}(p_{\mathrm{SML}}^{(\boldsymbol{\psi})})$.*

Proof. See Appendix C. □

Since the TMFPML approximation does not offer anything more than the SML estimate, we study a variant of the mean-field approximation inspired by the work of Chertkov and Yedidia [13]. We will see that, empirically, this gives a better estimate of several parameters of interest than the SML estimate.

B. MF Approximation Inspired by Chertkov and Yedidia

We rewrite the Gibbs free energy function (11) as follows to separate transitions on self loops from other transitions

[5] Note that here we have used $\boldsymbol{\sigma} = (\sigma_{ij})_{i,j}$ to denote a permutation matrix.

[6] Although β is often used to denote the inverse temperature of some statistical model, here β represents a probability distribution (or beliefs).

$$
\begin{aligned}
&F_{\mathrm{G}}(\beta; p, \boldsymbol{\psi}) \\
&= \log k - \sum_{\substack{i,j,l,m \\ j \neq i \\ m \neq l}} \sum_{\boldsymbol{\sigma} \in \mathcal{K}} \beta(\boldsymbol{\sigma}) \log \left(p_{lm}^{\mu_{ij} \sigma_{il} \sigma_{jm}} \right) \\
&\qquad - \sum_{i,l} \sum_{\boldsymbol{\sigma} \in \mathcal{K}} \beta(\boldsymbol{\sigma}) \log \left(p_{ll}^{\mu_{ii} \sigma_{il}} \right) + \sum_{\boldsymbol{\sigma} \in \mathcal{K}} \beta(\boldsymbol{\sigma}) \log \beta(\boldsymbol{\sigma}). \\
&= \log k - \sum_{\substack{i,j,l,m \\ j \neq i \\ m \neq l}} \Big(\sum_{\substack{\boldsymbol{\sigma} \in \mathcal{K} \\ \sigma_{il} = \sigma_{jm} = 1}} \beta(\boldsymbol{\sigma}) \Big) \log p_{lm}^{\mu_{ij}} \\
&\qquad - \sum_{i,l} \Big(\sum_{\substack{\boldsymbol{\sigma} \in \mathcal{K} \\ \sigma_{il} = 1}} \beta(\boldsymbol{\sigma}) \Big) \log p_{ll}^{\mu_{ii}} + \sum_{\boldsymbol{\sigma} \in \mathcal{K}} \beta(\boldsymbol{\sigma}) \log \beta(\boldsymbol{\sigma}).
\end{aligned}
$$

Inspired by the MF approach of Chertkov and Yedidia for the problem of computing the permanent [13], we define the following approximation of the Gibbs free energy function. In the traditional MF estimate, the minimization of the free energy function, viewed as a function of the matrix $\mathbf{b} \triangleq (\beta_{ij}(1))_{i,j}$, is performed over the set of all $k \times k$ permutation matrices. The new MF approximation is obtained by minimizing the free energy function over doubly stochastic matrices. Let $\mathcal{D}$ denote the set of all $k \times k$ doubly stochastic matrices.

$$
\begin{aligned}
&F_{\mathrm{MF}}(\,\cdot\,; p, \boldsymbol{\psi}) : \mathcal{D} \to \mathbb{R} \\
&F_{\mathrm{MF}}(\mathbf{b}; p, \boldsymbol{\psi}) = - \sum_{\substack{i,j,l,m \\ j \neq i \\ m \neq l}} b_{il} b_{jm} \log p_{lm}^{\mu_{ij}} - \sum_{i,l} b_{il} \log p_{ll}^{\mu_{ii}} \\
&\qquad + \sum_{i,l} \big(b_{il} \log b_{il} + (1 - b_{il}) \log (1 - b_{il}) \big) + \log k.
\end{aligned}
\tag{17}
$$

The mean-field PML (MFPML) estimate is defined as

$$
p_{\mathrm{MFPML}}^{(\boldsymbol{\psi})} \triangleq \arg\min_{p \in \mathcal{C}} \min_{(b_{ij}) \in \mathcal{D}} F_{\mathrm{MF}}(\mathbf{b}; p, \boldsymbol{\psi}).
$$

Some remarks are in order. Substituting $b_{il} = \beta_{il}(1)$ in (15), it can be verified that apart from the domain of (b_{ij}), the difference between (15) and (17) is mainly in the term $\sum_{\boldsymbol{\sigma} \in \{0,1\}^k} \big(\prod_{i,l} b_{il}^{\sigma_{il}} (1 - b_{il})^{1 - \sigma_{il}} \log 1_{\mathcal{K}}(\boldsymbol{\sigma}) \big)$, which is absent in (17). While the traditional MF approximation always gives a lower bound for the partition function [12], the above approximation does not. Additionally, F_{MF} is in general not a convex function of $(b_{il})_{i,l}$.

C. Simulation Results

We randomly generated $k \times k$ transition probability matrices where each non-diagonal entry was chosen independently and uniformly at random, and each diagonal entry was set to a constant. We computed plug-in estimates of the entropy rate and the absolute value of the second-largest eigenvalue of the transition matrix. These plug-in estimates were computed using the SML and the MFPML estimates and compared with the values computed using the (true) transition matrix. We performed an alternate minimization of F_{MF} with respect to p and $\{b_{il}\}$ to obtain the MFPML estimate. We used Algorithm 1 for computing the minimum of F_{MF} with respect to $\{b_{il}\}$. The algorithm simply performs iterations over the fixed-point equations obtained from the Lagrangian derived from F_{MF}.

The plots of the various estimates can be seen in Figures 1, 2 and 3. We can see in Fig. 1 that the prediction of the entropy rate by the MF estimate are close to that obtained using the SML estimate for large values of n. When n is of the order of k^2 however, the MF estimate does a better job, as seen in Figs. 2 and 3. Fig. 4 shows that, on average, MFPML predicts the eigenvalues of the transition matrix better than the SML estimate when the number of observed samples is of the order of k^2. Histograms of the average error in estimation of the second-largest eigenvalue and the entropy rate are provided in Figs. 5–8.

```
Data: $\{p_{lm}^{\mu_{ij}} : 1 \leq i, j, l, m \leq k\}$
Result: $\{b_{il}^{*}\}$
Initialize beliefs $\{b_{il}^{(0)} : 1 \leq i, l \leq k\}$;
Initialize Lagrange multipliers $\{u_i^{(0)} : 1 \leq i \leq k\}$ and
$\{v_l^{(0)} : 1 \leq l \leq k\}$;
$t \leftarrow 1$;
while suitable stopping condition not reached do
    $b_{il}^{(t)} \leftarrow \left( \frac{p_{ll}^{\mu_{ii}} \prod_{\substack{j,m \\ j \neq i \\ m \neq l}} p_{lm}^{\mu_{ij} b_{jm}}}{u_i^{(t-1)} v_l^{(t-1)} + p_{ll}^{\mu_{ii}} \prod_{\substack{j,m \\ j \neq i \\ m \neq l}} p_{lm}^{\mu_{ij} b_{jm}}} \right)$
    $u_i^{(t)} \leftarrow u_i^{(t-1)} \sum_{l} b_{il}^{(t)}, \quad v_l^{(t)} \leftarrow v_l^{(t-1)} \sum_{i} b_{il}^{(t)}$
    $t \leftarrow t + 1$;
end
$b_{il}^{*} = b_{il}^{(t-1)}$;
```

Algorithm 1: Algorithm used in MFPML approximation.

V. Acknowledgments

The authors would like to thank Navin Kashyap for useful discussions and Jayadev Acharya for sharing a draft version of [6].

Appendix A
The PML Estimate as an ML Estimate of the Canonical TPM

We now show that the the PML estimate is the canonical TPM that maximizes the probability of occurrence of the observed pattern, and that the initial distribution of the DTMC is not relevant. Let $\widetilde{P}$ be the transition matrix of the DTMC, and $P = (p_{ij})_{i,j}$ denote the canonical matrix corresponding to $\widetilde{P}$. We can obtain $\widetilde{P}$ from P by a permutation of the states. Let σ denote this permutation. Let π be a pmf on $[k]$. Here, π denotes the initial distribution, and is not necessarily a stationary measure. Given a sequence $\mathbf{x}$, let $(\mu_{ij})_{i,j}$ be the

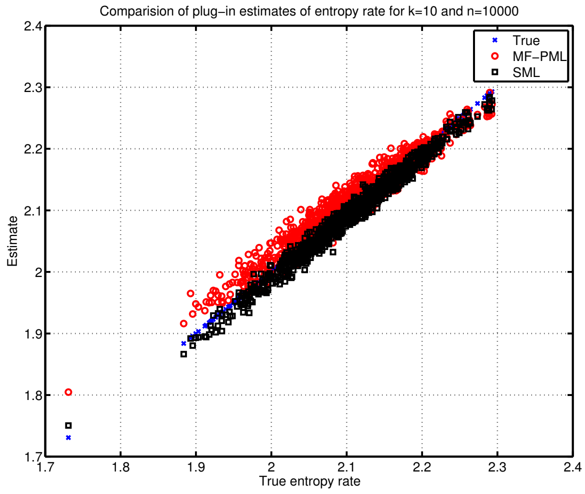

Fig. 1. Plot of entropy rate for $k = 10$ and $n = 10000$.

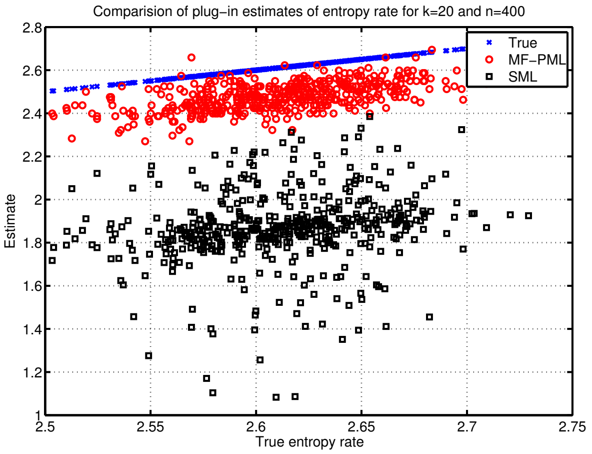

Fig. 2. Plot of entropy rate for $k = 20$ and $n = 400$.

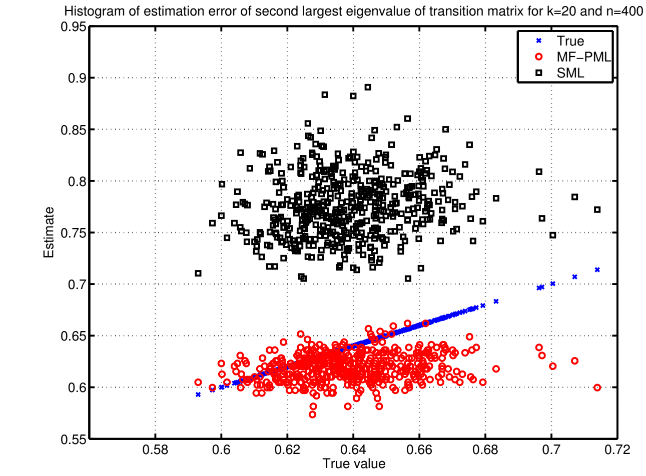

Fig. 3. Plot of second-largest eigenvalue (in absolute value) of transition matrix for $k = 20$ and $n = 400$.

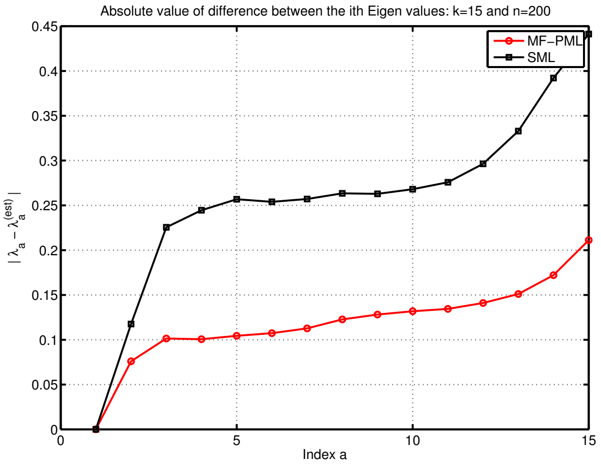

Fig. 4. Plot of $|\lambda_a^{(\text{true})} - \lambda_a^{(\text{estimate})}|$ for $1 \le a \le k$, for $k = 15$ and $n = 200$.

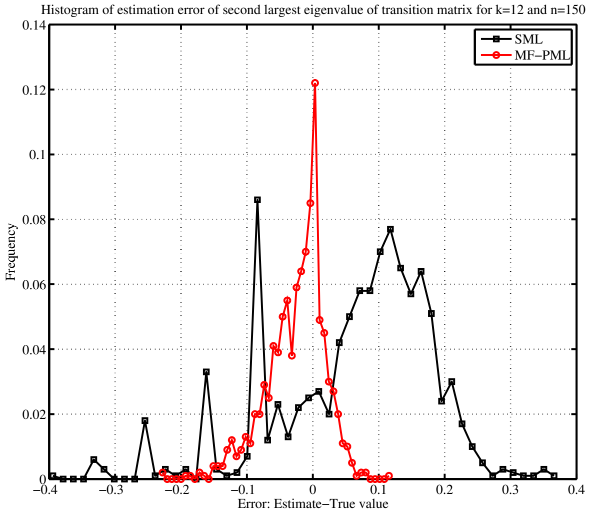

Fig. 5. Histogram of estimation error of second eigenvalue of transition matrix for $k = 12$ and $n = 150$.

multiplicity matrix corresponding to $\boldsymbol{\psi} = \boldsymbol{\psi}(\mathbf{x})$. Let $\mathcal{K}$ denote the set of all permutations of $[k]$. Then, the probability of observing $\boldsymbol{\psi}$ is given by

$$\Pr[\boldsymbol{\psi}|P, \pi] = \sum_{\sigma \in \mathcal{K}} \pi_{\sigma(1)} \prod_{(i,j)\in[k]\times[k]} p_{\sigma(i)\sigma(j)}^{\mu_{ij}}.$$

The ML estimate of P is the canonical matrix that maximizes the likelihood function,

$$L(P|\boldsymbol{\psi}) \triangleq \int_{\pi} \sum_{\sigma \in \mathcal{K}} \frac{1}{k!}\, \pi_{\sigma(1)} \prod_{(i,j)\in[k]\times[k]} p_{\sigma(i),\sigma(j)}^{\mu_{ij}} d\nu(\pi), \quad (18)$$

where we assume a uniform prior on π, i.e., $\nu(\pi)$ is the uniform distribution over the $(k-1)$-simplex. Here, we use the fact that the ML estimate is the most likely Bayes estimate assuming a uniform prior, and $\Pr[\boldsymbol{\psi}|P] = \int_{\pi} \Pr[\boldsymbol{\psi}|P, \pi]\nu(\pi)d\pi$.

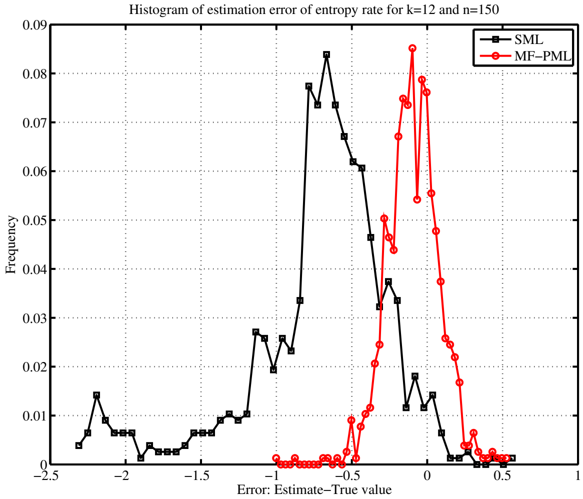

Fig. 6. Histogram of estimation error of entropy rate for $k = 12$ and $n = 150$.

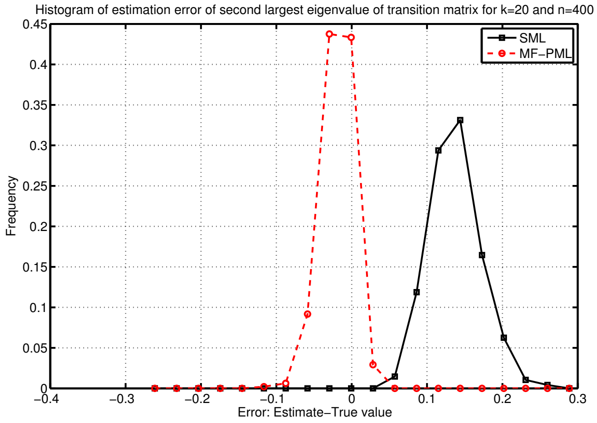

Fig. 7. Histogram of estimation error of second eigenvalue of transition matrix for $k = 20$ and $n = 400$.

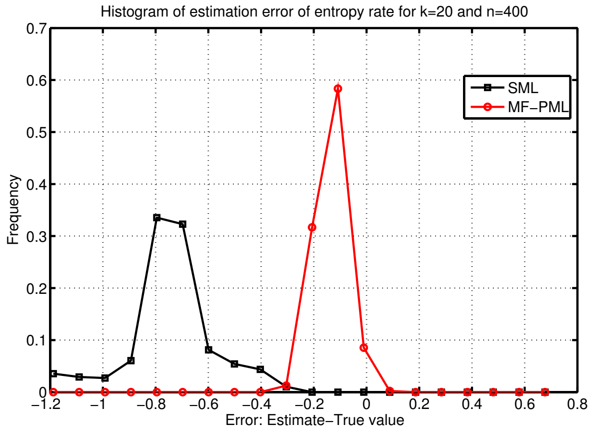

Fig. 8. Histogram of estimation error of entropy rate for $k = 20$ and $n = 400$.

Since $\nu(\pi)$ is uniform, by symmetry arguments, the integral $\int_\pi \pi_i d\nu(\pi)$ is a constant independent of i. Hence,

$$\begin{aligned} L(P|\boldsymbol{\psi}) &= \int_\pi \pi_{\sigma(1)} d\nu(\pi) \times \sum_{\sigma\in\mathcal{K}} \frac{1}{k!} \prod_{(i,j)\in[k]\times[k]} p_{\sigma(i)\sigma(j)}^{\mu_{ij}} \\ &= \alpha \sum_{\sigma\in\mathcal{K}} \prod_{(i,j)\in[k]\times[k]} p_{\sigma(i)\sigma(j)}^{\mu_{ij}}, \qquad (19) \end{aligned}$$

where α is a constant that depends only on k.

Therefore, maximizing (19) with respect to P is the same as maximizing

$$L'(P|\boldsymbol{\psi}) \triangleq \sum_{\sigma\in\mathcal{K}} \prod_{(i,j)\in[k]\times[k]} p_{\sigma(i)\sigma(j)}^{\mu_{ij}}.$$

Hence, the PML distribution gives the ML estimate of the canonical TPM.

Appendix B
Proof of Proposition 1

The proof follows the same ideas as [6, Theorem 2]. From the definition of the PML distribution, for every $\boldsymbol{\psi} \in \mathcal{Z}^{(n)}$, we have

$$\mathbb{P}\big(\boldsymbol{\psi}|p_{\text{PML}}^{(\boldsymbol{\psi})}\big) \geq \mathbb{P}(\boldsymbol{\psi}|p). \tag{20}$$

Now choose any pattern $\boldsymbol{\zeta}$ in $\mathcal{Z}^{(n)}$ that satisfies $\mathbb{P}(\boldsymbol{\zeta}|p) > \delta$. Suppose that $\mathbf{X}^{(n)}$ is a random sequence with a uniform initial distribution and having transition kernel $p_{\text{PML}}^{(\boldsymbol{\zeta})}$. Using (8), we have

$$\begin{aligned} \delta &> \Pr\Big[\big|f\big(p_{\text{PML}}^{(\boldsymbol{\zeta})}\big) - \hat{f}\big(\boldsymbol{\psi}(\mathbf{X}^{(n)})\big)\big| \geq \epsilon\Big] \\ &= \sum_{\boldsymbol{\psi}\in\mathcal{Z}^{(n)}} \mathbb{P}\big(\boldsymbol{\psi}; p_{\text{PML}}^{(\boldsymbol{\zeta})}\big) \cdot 1_{\left\{\left|f\left(p_{\text{PML}}^{(\boldsymbol{\zeta})}\right)-\hat{f}(\boldsymbol{\psi})\right|\geq\epsilon\right\}} \\ &\geq \mathbb{P}\big(\boldsymbol{\zeta}; p_{\text{PML}}^{(\boldsymbol{\zeta})}\big) \cdot 1_{\left\{\left|f\left(p_{\text{PML}}^{(\boldsymbol{\zeta})}\right)-\hat{f}(\boldsymbol{\zeta})\right|\geq\epsilon\right\}} \\ &\geq \mathbb{P}(\boldsymbol{\zeta}|p) \cdot 1_{\left\{\left|f\left(p_{\text{PML}}^{(\boldsymbol{\zeta})}\right)-\hat{f}(\boldsymbol{\zeta})\right|\geq\epsilon\right\}} \qquad (21) \\ &\geq \delta \cdot 1_{\left\{\left|f\left(p_{\text{PML}}^{(\boldsymbol{\zeta})}\right)-\hat{f}(\boldsymbol{\zeta})\right|\geq\epsilon\right\}}, \qquad (22) \end{aligned}$$

where (21) follows from (20). Clearly the above can hold only if $1_{\{|f(p_{\text{PML}}^{(\boldsymbol{\zeta})})-\hat{f}(\boldsymbol{\zeta})|\geq\epsilon\}} = 0$, or equivalently, $|f(p_{\text{PML}}^{(\boldsymbol{\zeta})}) - \hat{f}(\boldsymbol{\zeta})| < \epsilon$. Similarly, if $\mathbf{X}^{(n)}$ is distributed according to the transition kernel p, then,

$$\begin{aligned} \delta &> \Pr\Big[\big|f(p) - \hat{f}\big(\boldsymbol{\psi}(\mathbf{X}^{(n)})\big)\big| \geq \epsilon\Big] \\ &= \sum_{\boldsymbol{\psi}\in\mathcal{Z}^{(n)}} \mathbb{P}(\boldsymbol{\psi}|p) \cdot 1_{\{|f(p)-\hat{f}(\boldsymbol{\psi})|\geq\epsilon\}} \\ &\geq \mathbb{P}(\boldsymbol{\zeta}|p) \cdot 1_{\{|f(p)-\hat{f}(\boldsymbol{\zeta})|\geq\epsilon\}} \\ &\geq \delta \cdot 1_{\{|f(p)-\hat{f}(\boldsymbol{\zeta})|\geq\epsilon\}}. \qquad (23) \end{aligned}$$

Again, the above can hold only if $1_{\{|f(p)-\hat{f}(\boldsymbol{\zeta})|\geq\epsilon\}} = 0$. We can therefore conclude that for all $\boldsymbol{\zeta}$ satisfying $\mathbb{P}(\boldsymbol{\zeta}|p) > \delta$, we have $|f(p_{\text{PML}}^{(\boldsymbol{\zeta})}) - \hat{f}(\boldsymbol{\zeta})| < \epsilon$ and $|f(p) - \hat{f}(\boldsymbol{\zeta})| < \epsilon$. Using the triangle inequality, we have

$$|f(p_{\text{PML}}^{\boldsymbol{\zeta}}) - f(p)| < 2\epsilon \tag{24}$$

for all $\boldsymbol{\zeta}$ satisfying $\mathbb{P}(\boldsymbol{\zeta}|p) > \delta$.

We now have

$$
\begin{aligned}
&\Pr\Big[\big|f(p)-f\big(p_{\mathrm{PML}}^{(\boldsymbol{\psi})}\big)\big|\geq 2\epsilon\Big]\\
&=\sum_{\boldsymbol{\psi}\in\mathcal{Z}^{(n)}}\mathbb{P}(\boldsymbol{\psi}|p)\cdot 1_{\left\{\left|f(p)-f\left(p_{\mathrm{PML}}^{(\boldsymbol{\psi})}\right)\right|\geq 2\epsilon\right\}}\\
&=\sum_{\boldsymbol{\psi}:\mathbb{P}(\boldsymbol{\psi}|p)>\delta}\mathbb{P}(\boldsymbol{\psi}|p)\cdot 1_{\left\{\left|f(p)-f\left(p_{\mathrm{PML}}^{(\boldsymbol{\psi})}\right)\right|\geq 2\epsilon\right\}}\\
&\qquad+\sum_{\boldsymbol{\psi}:\mathbb{P}(\boldsymbol{\psi}|p)\leq\delta}\mathbb{P}(\boldsymbol{\psi}|p)\cdot 1_{\left\{\left|f(p)-f\left(p_{\mathrm{PML}}^{(\boldsymbol{\psi})}\right)\right|\geq 2\epsilon\right\}}\\
&=\sum_{\boldsymbol{\psi}:\mathbb{P}(\boldsymbol{\psi}|p)\leq\delta}\mathbb{P}(\boldsymbol{\psi}|p)\cdot 1_{\left\{\left|f(p)-f\left(p_{\mathrm{PML}}^{(\boldsymbol{\psi})}\right)\right|\geq 2\epsilon\right\}} \qquad (25)\\
&\leq\delta\cdot|\mathcal{Z}^{(n)}|, \qquad (26)
\end{aligned}
$$

where (25) follows from (24). This completes the proof of the proposition. □

Appendix C
Proof of Lemma 1

Let us define $b_{il}\triangleq\beta_{il}(1)$. With some manipulations, we can write (15) as

$$
\begin{aligned}
&F_{\mathrm{TMF}}(\mathbf{b};p,\boldsymbol{\psi})\\
&=-\sum_{\boldsymbol{\sigma}\in\{0,1\}^{k\times k}}\sum_{i,j,l,m}\Big(\prod_{(x,y)\in[k]\times[k]}b_{xy}^{\sigma_{xy}}(1-b_{xy})^{1-\sigma_{xy}}\\
&\qquad\qquad\times\log\big(p_{lm}^{\mu_{ij}\sigma_{il}\sigma_{jm}}\big)\Big)\\
&\quad-\sum_{\boldsymbol{\sigma}\in\{0,1\}^{k}}\big(\prod_{i,l}b_{il}^{\sigma_{il}}(1-b_{il})^{1-\sigma_{il}}\log 1_{\mathcal{K}}(\boldsymbol{\sigma})\big)\\
&\quad+\sum_{i,l}\big(b_{il}\log b_{il}+(1-b_{il})\log(1-b_{il})\big)+\log k\\
&=-\sum_{i,j,l,m}b_{il}b_{jm}\log p_{lm}^{\mu_{ij}}\\
&\quad-\sum_{\boldsymbol{\sigma}\in\{0,1\}^{k\times k}}\prod_{i,l}b_{il}^{\sigma_{il}}(1-b_{il})^{1-\sigma_{il}}\log 1_{\mathcal{K}}(\boldsymbol{\sigma})\\
&\quad+\sum_{i,l}\big(b_{il}\log b_{il}+(1-b_{il})\log(1-b_{il})\big)+\log k.
\end{aligned}
\tag{27}
$$

We can interpret b_{il} as the probability that the permutation $\boldsymbol{\sigma}$ maps i to l. Observe that $F_{\mathrm{TMF}}(\mathbf{b};p,\boldsymbol{\psi})$ is finite only if the probability $\prod_{i,l}b_{il}^{\sigma_{il}}(1-b_{il})^{1-\sigma_{il}}$ is zero for every $\boldsymbol{\sigma}\notin\mathcal{K}$.

For a binary-valued matrix $\boldsymbol{\sigma}$ and $1\leq j,m\leq k$, let us define another matrix $\boldsymbol{\sigma}^{(jm)}$ as follows

$$
(\boldsymbol{\sigma}^{(jm)})_{il}=\begin{cases}\sigma_{il} & \text{if } (i,l)\neq(j,m)\\ 1-\sigma_{il} & \text{if } (i,l)=(j,m).\end{cases}
$$

In other words, $\boldsymbol{\sigma}^{(jm)}$ is obtained from $\boldsymbol{\sigma}$ by flipping the (j,m)th entry. Since $\mathcal{K}$ denotes the set of all $k\times k$ permutation matrices, the following statement is obvious:

Fact 1. *Let $\boldsymbol{\sigma}\in\mathcal{K}$. Then, for every $1\leq j,m\leq k$, we have $\boldsymbol{\sigma}^{(jm)}\notin\mathcal{K}$.*

We also have the following statement

Fact 2. *$F_{\mathrm{TMF}}(\mathbf{b};p,\boldsymbol{\psi})$ is finite only if $b_{il}\in\{0,1\}$ for every i,l.*

Proof. As remarked earlier, $F_{\mathrm{TMF}}(\mathbf{b};p,\boldsymbol{\psi})$ is finite only if the product distribution $\{(1-b_{il},b_{il})\}$ assigns zero probability for all $\boldsymbol{\sigma}\notin\mathcal{K}$. If $\mathbf{b}$ assigns a positive probability to some $\boldsymbol{\sigma}\in\mathcal{K}$ and $0<b_{il}<1$ for some (i,l), then clearly $\mathbf{b}$ also assigns a nonzero probability to $\boldsymbol{\sigma}^{(il)}$. However, $\boldsymbol{\sigma}^{(il)}\notin\mathcal{K}$ by Fact 1 and hence $F_{\mathrm{TMF}}(\mathbf{b};p,\psi)=-\infty$. □

From facts 1 and 2, the following statement is obvious

Corollary 1. *$F_{\mathrm{TMF}}(\mathbf{b};p,\boldsymbol{\psi})$ is finite if and only if $(b_{il})_{i,l}\in\mathcal{K}$.*

In other words, the minimizing β puts all its mass on some permutation matrix. Since the second and third terms of (27) are then zero, we infer that minimizing F_{TMF} is equivalent to computing

$$
\begin{aligned}
\min_{\mathbf{b}}F_{\mathrm{TMF}}(\mathbf{b};p,\boldsymbol{\psi})&=\min_{\mathbf{b}\in\mathcal{K}\}}\sum_{i,j,l,m}-b_{il}b_{jm}\log p_{lm}^{\mu_{ij}}\\
&=\min_{\boldsymbol{\sigma}}\sum_{i,j}-\log p_{\sigma(i)\sigma(j)}^{\mu_{ij}},
\end{aligned}
\tag{28}
$$

where the above minimization is over all permutation matrices $\boldsymbol{\sigma}$. In other words, the minimizer of the mean-field free energy function corresponds to a perfect matching. For a fixed p, let $\boldsymbol{\sigma}^*$ be the minimizer of F_{TMF}. Furthermore, let $\sigma^{-1}(l)$ denote the unique $i\in[k]$ for which $\sigma_{il}=1$. Using Lagrange multipliers, it can be easily verified that[7]

$$
\begin{aligned}
&(\arg\min_{p}\min_{\mathbf{b}}F_{\mathrm{TMF}}(\mathbf{b};p,\boldsymbol{\psi}))_{lm}=\\
&\begin{cases}\frac{1}{k} & \text{if } \mu_{\sigma^{-1}(l)\sigma^{-1}(j)}=0 \text{ for all } 1\leq j\leq k\\ c_l\mu_{\sigma^{-1}(l)\sigma^{-1}(m)} & \text{otherwise}\end{cases}
\end{aligned}
$$

where the proportionality constants c_l are such that $\sum_m p_{lm}=1$ for all l. Hence, we can conclude that the traditional mean-field approximation is simply the (sequence) maximum-likelihood estimate of the transition matrix. □

References

[1] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," in *Proc. 20th Conf. Uncertainty in Artificial Intelligence*, Banff, Canada, 2004, pp. 426–435.

[2] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1469–1481, Jul. 2004.

[3] A. Orlitsky, Sajama, N. P. Santhanam, K. Viswanathan, and J. Zhang, "Algorithms for modeling distributions over large alphabets," in *Proc. 2004 IEEE Int. Symp. Inf. Theory*, Chicago, IL, 2004, p. 304.

[4] P. O. Vontobel, "The Bethe approximation of the pattern maximum likelihood distribution," in *Proc. 2012 IEEE Int. Symp. Inf. Theory*, Boston, MA, Jul. 2012, pp. 2012–2016.

[5] ——, "The Bethe and Sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the Valiant-Valiant estimate," in *Proc. 2014 Inf. Theory and Applications Workshop*, San Diego, CA, Feb. 2014, pp. 1–10.

[6] J. Acharya, H. Das, A. Orlitsky, and A. T. Suresh, "Estimating symmetric properties of distributions: A maximum likelihood approach," *draft manuscript*, 2015.

[7] Note that the following is true only up to a relabeling of the states since the resulting transition matrix may not be canonical.

[7] A. Dhulipala and A. Orlitsky, “Universal compression of Markov and related sources over arbitrary alphabets,” *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4182–4190, Sep. 2006.

[8] M. Mohri, M. Riley, and A. T. Suresh, “Automata and graph compression,” in *Proc. 2015 IEEE Int. Symp. Inf. Theory*, Hong Kong, Jun. 2015, pp. 2989–2993.

[9] W. Choi and W. Szpankowski, “Compression of graph structures: Fundamental limits, algorithms, and experiments,” *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 620–638, Feb. 2012.

[10] S. Kancharla, J. Neville, T. Courtade, and V. Kostina, “Graph compression: One-year report,” NSF Center for Science of Information, Tech. Rep., 2012. [Online]. Available: https://www.soihub.org/docs/team-graph-walks.pdf

[11] P. O. Vontobel, “The Bethe permanent of a nonnegative matrix,” *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1866–1901, Mar. 2013.

[12] J. S. Yedidia, W. Freeman, and Y. Weiss, “Constructing free-energy approximations and generalized belief propagation algorithms,” *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.

[13] M. Chertkov and A. Yedidia, “Approximating the permanent with fractional belief propagation,” *J. Machine Learning Research*, vol. 14, no. 1, pp. 2029–2066, 2013.