The Pattern Maximum Likelihood Estimation Problem

Shashank Vatedka

Institute of Network Coding
The Chinese University of Hong Kong
Joint work with Pascal Vontobel

19th April 2017
Overview

- Estimating properties of Markov chains and memoryless sources
- Symmetric properties and performance of plug-in estimators
- Pattern maximum likelihood (PML) estimate
- Approximating the PML estimate using a variational approach
Estimating the transition matrix of a DTMC
An estimation problem

Suppose we have \( X_1, X_2, X_3, \ldots, X_n \) from an irreducible time-homogeneous Markov chain over \( S = \{1, 2, \ldots, k\} \) with transition kernel

\[
p_{x,y} = \Pr[X_{t+1} = y | X_t = x]
\]

and uniform initial distribution.

We know \( k \), but we do not know \( p \).
An estimation problem

Suppose we have $X_1, X_2, X_3, \ldots, X_n$ from an irreducible time-homogeneous Markov chain over $S = \{1, 2, \ldots, k\}$ with transition kernel

$$p_{x, y} = \Pr[X_{t+1} = y | X_t = x]$$

and uniform initial distribution.

We know $k$, but we do not know $p$.

We observe sample path $x_1, x_2, \ldots, x_n$ of length $n$. 
An estimation problem

Suppose we have $X_1, X_2, X_3, \ldots, X_n$ from an irreducible time-homogeneous Markov chain over $S = \{1, 2, \ldots, k\}$ with transition kernel

$$p_{x,y} = \Pr[X_{t+1} = y|X_t = x]$$

and uniform initial distribution.

We know $k$, but we do not know $p$.

We observe sample path $x_1, x_2, \ldots, x_n$ of length $n$.

What can we infer about $p$?

Regime of interest: $n \leq k^2$. 
Let $G$ be an undirected graph.
Let $G$ be an undirected graph.
Let $G$ be an undirected graph.
Let $X_1, X_2, \ldots, X_n$ be a random walk starting from a random initial vertex.
Estimating graphs from random walks

Let $G$ be an undirected graph.
Let $X_1, X_2, \ldots, X_n$ be a random walk starting from a random initial vertex.

**Q:** What can we infer about $G$ from $X_1, X_2, \ldots, X_n$?
This is important in the regime where $n$ is less than $k^2$.

Many parameters such as the degree distribution, eigenvalues of the adjacency matrix, etc., are of interest.
We have a pmf $p$ over $S = \{1, 2, \ldots, k\}$

We observe $X_1, \ldots, X_n$, i.i.d. with each $X_i \sim p$.

What can we infer about $p$?
A simpler problem: the i.i.d. case

We have a pmf $p$ over $S = \{1, 2, \ldots, k\}$

We observe $X_1, \ldots, X_n$, i.i.d. with each $X_i \sim p$.

What can we infer about $p$?

$$\Pr[X^n = x^n] = \prod_{i=1}^{n} p_{x_i} = \prod_{a \in S} p_{a}^{\mu_a}$$

where $\mu_a$ is the number of times $a$ appears in $x^n$. 

If \( n \) is large enough, can find the **empirical estimate** of \( p \) (SML estimate):

For \( a \in S \), let \( \mu_a \) denote the number of times the symbol \( a \) occurs in \( x_1, x_2, \ldots, x_n \).

\[
(p_{\text{SML}})_a = \frac{\mu_a}{\sum_{b \in S} \mu_b} = \frac{\mu_a}{n}
\]
Sequence maximum likelihood estimation

If $n$ is large enough, can find the empirical estimate of $p$ (SML estimate):

For $a \in S$, let $\mu_a$ denote the number of times the symbol $a$ occurs in $x_1, x_2, \ldots, x_n$.

$$ (p_{SML})_a = \frac{\mu_a}{\sum_{b \in S} \mu_b} = \frac{\mu_a}{n} $$

**Problem:** If $n \lesssim k$, we do not get a good estimate.

- If $n < k$, some symbols will never be observed.
- The SML estimate assigns zero probability to such symbols.
Symmetric properties of distributions

- $f(p)$ is symmetric if it is invariant to a relabeling of the alphabet.
- For every $\sigma \in S_k$, $f(p_{\sigma(\cdot)}) = f(p)$.
- Examples: Support size, entropy (Shannon, Renyi), etc.

$$H(p) = -\sum_a p_a \log_2 p_a$$
Symmetric properties of distributions

- $f(p)$ is symmetric if it is invariant to a relabeling of the alphabet.

- For every $\sigma \in S_k$, $f(p_{\sigma(\cdot)}) = f(p)$.

- Examples: Support size, entropy (Shannon, Renyi), etc.

$$H(p) = -\sum_a p_a \log_2 p_a$$

- Want to estimate $f(p)$ from $X_1, \ldots, X_n$.

- Specifically, for $\epsilon, \delta > 0$, want an estimator $\hat{f} : S^n \to \mathbb{R}$ such that

$$\Pr[|f(p) - \hat{f}(X^n)| > \epsilon] < \delta$$
Symmetric properties of distributions

- $f(p)$ is symmetric if it is **invariant to a relabeling** of the alphabet.

- For every $\sigma \in S_k$, $f(p_{\sigma(\cdot)}) = f(p)$.

- Examples: Support size, entropy (Shannon, Renyi), etc.

\[
H(p) = -\sum_a p_a \log_2 p_a
\]

- Want to estimate $f(p)$ from $X_1, \ldots, X_n$.

- Specifically, for $\epsilon, \delta > 0$, want an estimator $\hat{f} : S^n \to \mathbb{R}$ such that

\[
\Pr[|f(p) - \hat{f}(X^n)| > \epsilon] < \delta
\]

- **Sample complexity**: smallest $N$ such that the above holds for all $n \geq N$. Typically take $\delta = 1/3$. 
Estimating symmetric properties: A plug-in approach?

- **Estimating** $f(p)$: Use **ML/favourite estimator**. Different estimator for each $f$. Complexity??
Estimating symmetric properties: A plug-in approach?

- **Estimating** $f(p)$: Use ML/favourite estimator. Different estimator for each $f$. Complexity??

- **Idea**: Find an approximation of $p$, i.e., $\hat{p}$. Compute $f(\hat{p})$ — plug-in estimator.
Estimating symmetric properties: A plug-in approach?

- **Estimating** $f(p)$: Use ML/favourite estimator. Different estimator for each $f$. Complexity??

- **Idea:** Find an approximation of $p$, i.e., $\hat{p}$. Compute $f(\hat{p})$ — plug-in estimator.

- **SML plug-in estimator:** Choose $\hat{p} = p_{SML}$.

- **Problem:** If $n$ is small compared to $k$, then $p_{SML}$ is bad.

- **Q:** Can we do better than the SML estimate?
The Pattern Maximum Likelihood Estimate
An alternative to $\rho_{\text{SML}}$

**Pattern:**

- Given $x = x_1, x_2, \ldots, x_n$, the **index** of symbol $a$ in $x$ is $1$ plus the number of distinct symbols occurring before the first occurrence of $a$ in $x$.

- The **pattern** of $x$ is the string obtained by replacing $x_i$ by the index of $x_i$. 

Example:
Consider $x = \text{abracadabra}$.

The pattern of $x$, $\psi(x) = 12314151231$.

Symbol | Index
-------|-----
a | 1
b | 2
r | 3
c | 4
d | 5
An alternative to $\rho_{SML}$

Pattern:

- Given $x = x_1, x_2, \ldots, x_n$, the index of symbol $a$ in $x$ is 1 plus the number of distinct symbols occurring before the first occurrence of $a$ in $x$.
- The pattern of $x$ is the string obtained by replacing $x_i$ by the index of $x_i$.

Example: Consider $x = abracadabra$.

The pattern of $x$, $\psi(x) = 12314151231$.

Profile:

multiset of number of occurrences of different symbols

$\{\mu_1, \mu_2, \ldots, \mu_n\}$

Profile of abracadabra: $\{5, 2, 2, 1, 1\}$
An alternative to $\rho_{SML}$

Pattern:

- Given $x = x_1, x_2, \ldots, x_n$, the index of symbol $a$ in $x$ is 1 plus the number of distinct symbols occurring before the first occurrence of $a$ in $x$.
- The pattern of $x$ is the string obtained by replacing $x_i$ by the index of $x_i$.

Example: Consider $x = abracadabra$.

The pattern of $x$,

$$\psi(x) = 12314151231.$$ 

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td>r</td>
<td>3</td>
</tr>
<tr>
<td>c</td>
<td>4</td>
</tr>
<tr>
<td>d</td>
<td>5</td>
</tr>
</tbody>
</table>

Pattern probability:

$$\mathbb{P}(\psi|p) \triangleq \sum_{\sigma} \prod_{i=1}^{k} p_{\sigma(i)}^{\mu_i}$$
An alternative to $\rho_{SML}$

**Pattern:**

- Given $x = x_1, x_2, \ldots, x_n$, the *index* of symbol $a$ in $x$ is 1 plus the number of distinct symbols occurring before the first occurrence of $a$ in $x$.
- The pattern of $x$ is the string obtained by replacing $x_i$ by the index of $x_i$.

**Example:** Consider $x = abracadabra$.

The pattern of $x$,

$$\psi(x) = 12314151231.$$  

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>1</td>
</tr>
<tr>
<td>$b$</td>
<td>2</td>
</tr>
<tr>
<td>$r$</td>
<td>3</td>
</tr>
<tr>
<td>$c$</td>
<td>4</td>
</tr>
<tr>
<td>$d$</td>
<td>5</td>
</tr>
</tbody>
</table>

$$\mathbb{P}(12314151231|p) = \Pr[abcadaeabca] + \cdots + \Pr[abracadabra] + \cdots$$
An alternative to $\rho_{SML}$: The PML estimate

**SML and PML estimates**

- $\rho_{SML}$: is the pmf that maximizes the probability of occurrence of the sequence $x$. 
An alternative to $\rho_{\text{SML}}$: The PML estimate

**SML and PML estimates**

- $\rho_{\text{SML}}$: is the pmf that maximizes the probability of occurrence of the sequence $\mathbf{x}$.

- $\rho_{\text{PML}}$: the Pattern maximum likelihood (PML) estimate is the pmf that maximizes the probability of occurrence of $\psi(\mathbf{x})$. 
An alternative to $\rho_{SML}$: The PML estimate

**SML and PML estimates**

- $\rho_{SML}$: is the pmf that maximizes the probability of occurrence of the sequence $x$.

- $\rho_{PML}$: the Pattern maximum likelihood (PML) estimate is the pmf that maximizes the probability of occurrence of $\psi(x)$.

For convenience, maximize over ordered pmfs, i.e., $p_1 \geq p_2 \geq \ldots \geq p_k$.

$$p^{(\psi)}_{PML} = \arg \max_{p \in \mathcal{P}_k} \mathbb{P}(\psi|p)$$

$$= \arg \max_{p \in \mathcal{P}_k} \sum_{\sigma} \prod_{i=1}^{k} p_{\sigma(i)}^{\mu_i}$$ (1)
Origins of PML: Universal compression of memoryless sources over unknown alphabets by Orlitsky et al.\(^1\)

---

Origins of PML: Universal compression of memoryless sources over unknown alphabets by Orlitsky et al.¹

Universal compression: block redundancy (average number of additional bits required compared to the case when distribution is known)

\[ R(\mathcal{P}) = \inf_q \sup_p \sup_{x \in \mathcal{S}} \log \frac{p(x)}{q(x)} \]

Origins of PML:

- Universal compression of memoryless sources over unknown alphabets by Orlitsky et al.\(^1\)

- Block redundancy (average number of additional bits required compared to the case when distribution is known)

\[
R(\mathcal{P}) = \inf_q \sup_p \sup_{x \in \mathcal{S}} \log \frac{p(x)}{q(x)}
\]

- For sequences, block redundancy

\[
R(I^n_k) = \frac{k - 1}{2} \log \frac{n}{2\pi} + \log \left( \frac{\Gamma(1/2)^k}{\Gamma(k/2)} \right) + o_k(1)
\]

Origins of PML

- Origins of PML: Universal compression of memoryless sources over unknown alphabets by Orlitsky et al.¹

- Universal compression: block redundancy (average number of additional bits required compared to the case when distribution is known)

\[
R(\mathcal{P}) = \inf_{q} \sup_{p} \sup_{x \in S} \log \frac{p(x)}{q(x)}
\]

- For sequences, block redundancy

\[
R(I_k^n) = \frac{k - 1}{2} \log \frac{n}{2\pi} + \log \left( \frac{\Gamma(1/2)^k}{\Gamma(k/2)} \right) + o_k(1)
\]

- (Orlitsky et al.) For compressing patterns, block redundancy

\[
(1.5 \log e) n^{1/3} (1 + o(1)) \leq R(I_{\psi}^n) \leq \pi \sqrt{2/3(\log e)} \sqrt{n}
\]

---

PML plug-in estimator: Compute $p_{\text{PML}}$, and find $f(p_{\text{PML}})$.

---

Estimating symmetric properties using $p_{\text{PML}}$

**PML plug-in estimator:** Compute $p_{\text{PML}}$, and find $f(p_{\text{PML}})$.

Let $\mathcal{Z}^{(n)}$ denote the set of all length-$n$ patterns.

---

**Proposition (Acharya et al.\(^1\))**

Consider any estimator $\hat{f}$ for $f$ that takes as input\(^2\) $\psi(X^{(n)})$. Suppose that for every $\epsilon > 0$, $\delta > 0$ and transition probability distribution $p$, there exists $N$ such that

$$\Pr\left[ |f(p) - \hat{f}(\psi(X^{(n)}))| \geq \epsilon \right] < \delta$$

for all $n \geq N$. Then,

$$\Pr\left[ |f(p_{\text{PML}}) - f(p)| \geq 2\epsilon \right] < \delta \cdot |\mathcal{Z}^{(n)}|$$

for all $n \geq N$.

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

---

The previous result is not as bad as it sounds!

- (Acharya et al.) For symmetric properties such as entropy, support size, distance from uniform distribution, sample complexity is order optimal!
The previous result is not as bad as it sounds!

- (Acharya et al.) For symmetric properties such as entropy, support size, distance from uniform distribution, sample complexity is order optimal!

<table>
<thead>
<tr>
<th>Property</th>
<th>SML</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>$O(k/\epsilon)$</td>
<td>$O\left(\frac{k}{\epsilon \log k}\right)$</td>
</tr>
<tr>
<td>Support size*</td>
<td>$O(k \log(1/\epsilon))$</td>
<td>$O\left(\frac{k}{\log k \log^2(1/\epsilon)}\right)$</td>
</tr>
</tbody>
</table>
The previous result is not as bad as it sounds!

- (Acharya et al.) For symmetric properties such as entropy, support size, distance from uniform distribution, sample complexity is order optimal!

<table>
<thead>
<tr>
<th>Property</th>
<th>SML</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>$O(k/\epsilon)$</td>
<td>$O\left(\frac{k}{\epsilon \log k}\right)$</td>
</tr>
<tr>
<td>Support size*</td>
<td>$O(k \log(1/\epsilon))$</td>
<td>$O\left(\frac{k}{\log k \log^2(1/\epsilon)}\right)$</td>
</tr>
</tbody>
</table>

- **Basic idea:** There exist optimal estimators that give bias $\epsilon$, error probability $1/3$, and satisfy a “bounded difference property”

- Have sample complexity $N_s$

- Use **McDiarmid’s inequality** to show that probability of error $e^{-\Omega(\sqrt{n})}$ can be achieved using $O(N_s)$ samples

- Then, use previous result
Efficiently approximating $p_{\text{PML}}$
Computing $p_{PML}$

$$p_{PML}^{(\psi)} = \arg \max_{p \in \mathcal{P}} \sum_{\sigma} \prod_{a=1}^{k} p_{\sigma(a)}^{\mu_a}.$$
Computing $p_{PML}$

$$p_{PML}^{(\psi)} = \arg \max_{p \in \mathcal{P}} \sum_{\sigma} \prod_{a=1}^{k} p_{\sigma(a)}^{\mu_a}.$$

**Permanent:** Given $k \times k$ matrix $M = (m_{i,j})$,

$$\text{perm}(M) = \sum_{\sigma \in S_k} \prod_{i=1}^{k} a_{i,\sigma(i)}$$
Computing $p_{\text{PML}}$

$$p^{(\psi)}_{\text{PML}} = \arg\max_{p \in \mathcal{P}} \sum_{\sigma} \prod_{a=1}^{k} p^{\mu_a}_{\sigma(a)}.$$ 

**Determinant:** Given $k \times k$ matrix $M = (m_{i,j})$,

$$\det(M) = \sum_{\sigma \in S_k} (-1)^{\text{sgn}(\sigma)} \prod_{i=1}^{k} a_{i,\sigma(i)}.$$
Computing $p_{\text{PML}}$

$$p_{\text{PML}}^{(\psi)} = \arg \max_{p \in \mathcal{P}} \sum_{\sigma} \prod_{a=1}^{k} p_{\sigma(a)}.$$

**Permanent:** Given $k \times k$ matrix $M = (m_{i,j})$,

$$\text{perm}(M) = \sum_{\sigma \in S_k} \prod_{i=1}^{k} a_{i,\sigma(i)}.$$

Pattern probability = $\text{perm}((p_{ij}^{\mu}))$

Computing permanent is hard!
For 0 – 1 matrix, best known Ryser’s algorithm requires $O(k2^k)$ operations.
We use a variational approach as done by Vontobel\(^a\).

\(^a\)P. O. Vontobel, “The Bethe approximation of the pattern maximum likelihood distribution,”
ISIT, Boston, MA, 2012

P. O. Vontobel, “The Bethe and Sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the Valiant-Valiant estimate,” ITA, San Diego, CA, 2014
A variational approach: Reformulating $\rho_{\text{PML}}$

$$Z \triangleq \mathbb{P}(\psi | p) = \sum_{\sigma \in \mathcal{K}} \prod_{i,j} p_{ij}^{\mu_j \sigma_{ij}}$$

where $\sigma_{ij}$ is the $(i,j)$th entry of the permutation matrix $\sigma$

**Objective:** Express this as the minimum of a certain **free energy** function.
A variational approach: Reformulating $\rho_{\text{PML}}$

\[
Z \triangleq \mathbb{P}(\psi|\rho) = \sum_{\sigma \in \mathcal{K}} \prod_{i,j} p_{i}^{\mu_{j}\sigma_{ij}}
\]

where $\sigma_{ij}$ is the $(i, j)$th entry of the permutation matrix $\sigma$. Introduce a “trial” distribution $\beta$ on all permutations on $\{1, 2, \ldots, k\}$.

Define the Gibbs average energy function

\[
U_{G}(\beta; \rho, \psi) \triangleq -\sum_{\sigma \in \mathcal{K}} \beta(\sigma) \log \left( \prod_{i,j} p_{i}^{\mu_{j}\sigma_{ij}} \right)
\]

\[
= -\sum_{\sigma \in \mathcal{K}} \sum_{i,j} \beta(\sigma)_{ij} \log \left( p_{i}^{\mu_{j}} \right),
\]

and the Gibbs entropy function

\[
H_{G}(\beta) \triangleq -\sum_{\sigma \in \mathcal{K}} \beta(\sigma) \log \beta(\sigma).
\]
A variational approach: Reformulating $\rho_{PML}$

\[
U_G(\beta; p, \psi) = \log k - \sum_{\sigma \in \mathcal{K}} \sum_{i,j,l,m} \beta(\sigma) \log \left( p_{l,m}^{\mu ij \sigma_{ij} \sigma_{jm}} \right),
\]

\[
H_G(\beta) = - \sum_{\sigma \in \mathcal{K}} \beta(\sigma) \log \beta(\sigma).
\]

We define the Gibbs free energy function

\[
F_G(\beta; p, \psi) \triangleq U_G(\beta; p, \psi) - H_G(\beta),
\]

It is a fact that\(^2\)

\[
\min_{\beta} F_G(\beta; p, \psi) = - \log Z = - \log \mathbb{P}(\psi|p)
\]

Therefore,

\[
\rho_{PML}^{(\psi)} = \arg \min_{p \in \mathcal{C}} \min_{\beta \in \mathcal{P}} F_G(\beta; p, \psi).
\]

Thermodynamic system with state space $\mathcal{K}$. Probability that the system is in state $\sigma$:

$$\gamma(\sigma) = \frac{e^{-E(\sigma)/T}}{Z}$$

where

- $E : \mathcal{K} \rightarrow \mathbb{R}$ is the energy function (Hamiltonian)
- $T$ is the temperature, $\kappa$ is Boltzmann’s constant ($1.38 \times 10^{-23} \text{ JK}^{-1}$)
- $Z = \sum_\sigma e^{-E(\sigma)/(\kappa T)}$ is the Helmholtz free energy
Interpretation

Thermodynamic system with state space $\mathcal{K}$. Probability that the system is in state $\sigma$:

$$
\gamma(\sigma) = \frac{e^{-E(\sigma)/T}}{Z}
$$

where

- $E : \mathcal{K} \to \mathbb{R}$ is the energy function (Hamiltonian)
- $T$ is the temperature, $\kappa$ is Boltzmann’s constant ($1.38 \times 10^{-23} \text{ JK}^{-1}$)
- $Z = \sum_{\sigma} e^{-E(\sigma)/(\kappa T)}$ is the Helmholtz free energy

In our case,

- $E(\sigma) = \sum_{i,j} \sigma_{i,j} \log p_{ij}^{\mu j}$
- $\kappa T = 1$
- $Z = \mathbb{P}(\psi|\rho)$
Interpretation

The Helmholtz average energy function

\[
U_H(\gamma; E) \triangleq \sum_{\sigma} \gamma(\sigma)E(\sigma)
\]

and the Helmholtz entropy function

\[
H_H(\gamma) \triangleq - \sum_{\sigma \in K} \gamma(\sigma) \log \gamma(\sigma).
\]

Then, \( F_H = -\kappa T \log Z = U_H - TH_H \)
The Helmholtz average energy function

\[ U_H(\gamma; E) \triangleq \sum_{\sigma} \gamma(\sigma) E(\sigma) \]

and the Helmholtz entropy function

\[ H_H(\gamma) \triangleq - \sum_{\sigma \in \mathcal{K}} \gamma(\sigma) \log \gamma(\sigma). \]

Then, \[ F_H = -\kappa T \log Z = U_H - TH_H \]

The Gibbs average energy function

\[ U_G(\beta; E) \triangleq - \sum_{\sigma \in \mathcal{K}} \beta(\sigma) E(\sigma), \]

the Gibbs entropy function

\[ H_G(\beta) \triangleq - \sum_{\sigma \in \mathcal{K}} \beta(\sigma) \log \beta(\sigma), \]

and \[ F_G = U_G - TH_G \]
A variational approach to approximating $\rho_{PML}$

\[
\rho_{PML}^{(\psi)} = \arg \min_{p} \min_{\beta} F_G(\beta; p, \psi).
\]

But how do we compute this?

---

A variational approach to approximating $\rho_{\text{PML}}$

\[ \rho_{\text{PML}}^{(\psi)} = \arg \min_p \min_\beta F_G(\beta; p, \psi). \]

But how do we compute this?

**Idea:** Use approximations that are easy to compute\(^3\).

Specifically, perform minimization w.r.t. $\beta$ over an easier set.

A variational approach to approximating $p_{\text{PML}}$

\[ p_{\text{PML}}^{(\psi)} = \arg \min_p \min_\beta F_G(\beta; p, \psi). \]

But how do we compute this?

**Idea:** Use approximations that are easy to compute\(^3\). Specifically, perform minimization w.r.t. $\beta$ over an easier set.

**Mean field approximation:** Choose $\beta$ to be a product distribution. Easy to compute.

**Bethe approximation:** Typically use low-complexity belief propagation algorithms.

---

Generalization to DTMCs
An alternative to $\rho_{\text{SML}}$: The PML estimate

SML and PML estimates

- $\rho_{\text{SML}}$: is the transition kernel that maximizes the probability of occurrence of the sequence $x$. 
SML and PML estimates

- $p_{SML}$: is the transition kernel that maximizes the probability of occurrence of the sequence $x$.

- $p_{PML}$: the Pattern maximum likelihood (PML) estimate is the transition kernel that maximizes the probability of occurrence of $\psi(x)$.
SML and PML estimates

- $p_{SML}$: is the transition kernel that maximizes the probability of occurrence of the sequence $x$.
- $p_{PML}$: the Pattern maximum likelihood (PML) estimate is the transition kernel that maximizes the probability of occurrence of $\psi(x)$.

Pattern probability:

$$P(\psi|p) \triangleq \frac{1}{k} \sum_{\sigma} \prod_{i=1}^{k} \prod_{j=1}^{k} p_{\sigma(i),\sigma(j)}^{\mu_{ij}}$$
An alternative to $\rho_{\text{SML}}$: The PML estimate

**SML and PML estimates**
- $\rho_{\text{SML}}$: is the transition kernel that maximizes the probability of occurrence of the sequence $x$.
- $\rho_{\text{PML}}$: the Pattern maximum likelihood (PML) estimate is the transition kernel that maximizes the probability of occurrence of $\psi(x)$.

**Pattern probability:**

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

**PML estimate:**

$$p^{(\psi)}_{\text{PML}} = \arg \max_p P(\psi|p).$$
The traditional mean-field approximation

\[ p_{\text{PML}}^{(\psi)} = \arg \min_{p \in \mathcal{P}} \min_{\beta \in \mathcal{P}'} F_G(\beta; p, \psi). \]

Choose \( \beta \) to be a product distribution on \( k \times k \) binary matrices, i.e.,

\[ \beta(\sigma) = \prod_{i,l} \beta_{il}(\sigma_{il}). \]

\[ F_{\text{TMF}}(\beta; p, \psi) = -\sum_{\sigma \in \{0,1\}^{k \times k}} \left( \left( \prod_{i,l} \beta_{il}(\sigma_{il}) \right) \log \left( 1_K(\sigma) \prod_{i,j,l,m} p_{l,m}^{i,j} \sigma_{il} \sigma_{jm} \right) \right) \]

\[ + \sum_{i,l} \sum_{\sigma_{il}=0} 1 \beta_{il}(\sigma_{il}) \log \beta_{il}(\sigma_{il}) + \log k. \]

The traditional mean-field PML estimate is

\[ p_{\text{TFMFPML}}^{(\psi)} = \arg \min_{p \in \mathcal{C}} \min_{\beta} F_{\text{TMF}}(\beta; p, \psi). \]

However, we show that this actually reduces to the SML estimate.
A modified mean field estimate

Inspired by mean field approach used by Chertkov and Yedidia\textsuperscript{4} for approximating permanent of a nonnegative matrix.

In the MF approximation, impose constraint that $\sum_l \beta_{il}(1) = \sum_i \beta_{il}(1) = 1$. Define $b_{il} \triangleq \beta_{il}(1)$.

$$F_{MF}(\cdot; p, \psi) : \mathcal{D} \rightarrow \mathbb{R}$$

$$F_{MF}(b; p, \psi) = - \sum_{i,j,l,m} b_{il} b_{jm} \log p_{lm}^{\mu_{ij}} - \sum_{i,l} b_{il} \log p_{ll}^{\mu_{ii}}$$

$$+ \sum_{i,l} (b_{il} \log b_{il} + (1 - b_{il}) \log(1 - b_{il})) + \log k. \quad (2)$$

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

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

Empirical results
Empirical results

We have a low-complexity algorithm to compute MFPML estimate.

**Figure:** Histogram of estimation error of absolute second largest eigenvalue of transition matrix for $k = 20$ and $n = 400$. 

- **SML**
- **MF−PML**
Empirical results

We have a low-complexity algorithm to compute MFPML estimate.

Figure: Histogram of estimation error of entropy rate for $k = 20$ and $n = 400$. 
Empirical results

We have a low-complexity algorithm to compute MFPML estimate.

Figure: Histogram of estimation error of absolute second largest eigenvalue of transition matrix for $k = 50$ and $n = 2000$. 
Empirical results

We have a low-complexity algorithm to compute MFPML estimate.

Figure: Histogram of estimation error of entropy rate for $k = 50$ and $n = 2000$. 
Points to ponder on

- Good reasons to study PML estimates for Markov chains.
- Obtaining efficient approximations is hard.
- Bethe approximation: Complexity blows up very quickly.
- Ideally want algorithms to work for large $k$.
- Even the mean field PML estimate becomes difficult to implement for very large $k$. 
Thank you!