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 $\mathcal{S} = \{1, 2, \ldots, k\}$ 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 $\mathcal{S} = \{1, 2, \ldots, k\}$ 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 $\mathcal{S} = \{1, 2, \ldots, k\}$ with transition kernel

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

and uniform initial distribution.

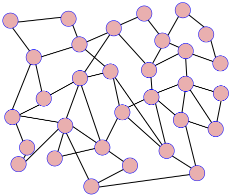
We know k, but we do not know p.

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

What can we infer about p?
Regime of interest: $n \leq k^2$.

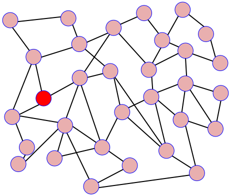
Estimating graphs from random walks

Let $\mathcal{G}$ be an undirected graph.

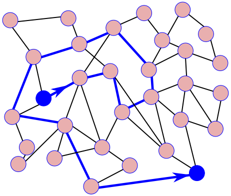


Estimating graphs from random walks

Let $\mathcal{G}$ be an undirected graph.

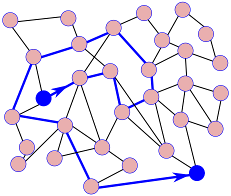


Let $\mathcal{G}$ be an undirected graph.
Let $X_1, X_2, \ldots, X_n$ be a random walk starting from a random initial vertex.



Estimating graphs from random walks

Let $\mathcal{G}$ 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 $\mathcal{G}$ 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.

A simpler problem: the i.i.d. case

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

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 $\mathcal{S} = \{1, 2, \ldots, k\}$

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

What can we infer about p?

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

where μ_a is the number of times a appears in x^n.

Sequence maximum likelihood estimation

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

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

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

Sequence maximum likelihood estimation

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

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

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

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 $\hat{f} : \mathcal{S}^n \to \mathbb{R}$ 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 $\hat{f} : \mathcal{S}^n \to \mathbb{R}$ 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 $\hat{p} = p_{\text{SML}}$.
- **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 p_{SML}

Pattern:

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

An alternative to p_{SML}

Pattern:

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

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

The pattern of $\mathbf{x}$,

$$\psi(\mathbf{x}) = 12314151231.$$

Symbol	Index
a	1
b	2
r	3
c	4
d	5

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 p_{SML}

Pattern:

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

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

The pattern of $\mathbf{x}$,

$$\psi(\mathbf{x}) = 12314151231.$$

Symbol	Index
a	1
b	2
r	3
c	4
d	5

Pattern probability:

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

An alternative to p_{SML}

Pattern:

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

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

The pattern of $\mathbf{x}$,

$$\psi(\mathbf{x}) = 12314151231.$$

Symbol	Index
a	1
b	2
r	3
c	4
d	5

$$\mathbb{P}(12314151231|p) = \Pr[abcadaeabca] + \cdots + \Pr[abracadabra] + \cdots$$

An alternative to p_{SML}: The PML estimate

SML and PML estimates

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

An alternative to p_{SML}: The PML estimate

SML and PML estimates

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

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

An alternative to p_{SML}: The PML estimate

SML and PML estimates

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

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

$$\begin{aligned} p_{\mathrm{PML}}^{(\psi)} &= \arg\max_{p \in \mathcal{P}_k} \mathbb{P}(\psi | p) \\ &= \arg\max_{p \in \mathcal{P}_k} \sum_{\sigma} \prod_{i=1}^{k} p_{\sigma(i)}^{\mu_i} \end{aligned} \tag{1}$$

Origins of PML

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

[1] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory,* Jul. 2004

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)

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

[1] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory,* Jul. 2004

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)

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

- For sequences, block redundancy

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

[1] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, Jul. 2004

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)

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

- 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

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

[1] A. Orlitsky, N. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, Jul. 2004

Estimating symmetric properties using p_{PML}

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

[1] J. Acharya, H. Das, A. Orlitsky, and A.T. Suresh, “A Unified Maximum Likelihood Approach for Optimal Distribution Property Estimation,” arXiv, Dec 2016

Estimating symmetric properties using p_{PML}

PML plug-in estimator: Compute p_{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] $\psi(\mathbf{X}^{(n)})$. *Suppose that for every* $\epsilon > 0$, $\delta > 0$ *and transition probability distribution* p, *there exists* N *such that*

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

for all $n \geq N$. *Then,*

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

for all $n \geq N$.

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

[1] J. Acharya, H. Das, A. Orlitsky, and A.T. Suresh, "A Unified Maximum Likelihood Approach for Optimal Distribution Property Estimation," arXiv, Dec 2016

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!

Property	SML	Optimal
Entropy	$O(k/\epsilon)$	$O\left(\frac{k}{\epsilon \log k}\right)$
Support size*	$O(k \log(1/\epsilon))$	$O\left(\frac{k}{\log k} \log^2(1/\epsilon)\right)$

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!

Property	SML	Optimal
Entropy	$O(k/\epsilon)$	$O\left(\frac{k}{\epsilon \log k}\right)$
Support size*	$O(k \log(1/\epsilon))$	$O\left(\frac{k}{\log k} \log^2(1/\epsilon)\right)$

- Basic idea: There exist optimal estimators that give bias ϵ, 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_{PML}

Computing p_{PML}

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

Computing p_{PML}

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

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

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

Computing p_{PML}

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

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_{PML}

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

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

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

Pattern probability = $\text{perm}((p_i^{\mu_j}))$

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 p_{PML}

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

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

Objective: Express this as the minimum of a certain free energy function.

A variational approach: Reformulating p_{PML}

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

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

Introduce a "trial" distribution β on all permutations on $\{1, 2, \ldots, k\}$.

Define the Gibbs average energy function

$$\begin{aligned} U_{\mathrm{G}}(\beta; p, \psi) &\triangleq -\sum_{\boldsymbol{\sigma} \in \mathcal{K}} \beta(\boldsymbol{\sigma}) \log \left(\prod_{i,j} p_i^{\mu_j \sigma_{ij}} \right) \\ &= -\sum_{\boldsymbol{\sigma} \in \mathcal{K}} \sum_{i,j} \beta(\boldsymbol{\sigma}) \sigma_{ij} \log \left(p_i^{\mu_j} \right), \end{aligned}$$

and the Gibbs entropy function

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

A variational approach: Reformulating p_{PML}

$$U_{\mathrm{G}}(\beta; p, \psi) = \log k - \sum_{\boldsymbol{\sigma} \in \mathcal{K}} \sum_{i,j,l,m} \beta(\boldsymbol{\sigma}) \log \left(p_{l,m}^{\mu_{ij} \sigma_{il} \sigma_{jm}} \right),$$

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

We define the Gibbs free energy function

$$F_{\mathrm{G}}(\beta; p, \psi) \triangleq U_{\mathrm{G}}(\beta; p, \psi) - H_{\mathrm{G}}(\beta),$$

It is a fact that[2]

$$\min_{\beta} F_{\mathrm{G}}(\beta; p, \psi) = -\log Z = -\log \mathbb{P}(\psi | p)$$

Therefore,

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

[2] J.S. Yedidia, W. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, Jul. 2005

Interpretation

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

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

where

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

Interpretation

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

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

where

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

In our case,

- $E(\sigma) = \sum_{i,j} \sigma_{i,j} \log p_i^{\mu_j}$
- $\kappa T = 1$
- $Z = \mathbb{P}(\psi | p)$

Interpretation

The Helmholtz average energy function

$$U_{\mathrm{H}}(\gamma; E) \triangleq \sum_{\sigma} \gamma(\sigma) E(\sigma)$$

and the Helmholtz entropy function

$$H_{\mathrm{H}}(\gamma) \triangleq -\sum_{\boldsymbol{\sigma} \in \mathcal{K}} \gamma(\boldsymbol{\sigma}) \log \gamma(\boldsymbol{\sigma}).$$

Then, $F_H = -\kappa T \log Z = U_{\mathrm{H}} - T H_{\mathrm{H}}$

Interpretation

The Helmholtz average energy function

$$U_{\mathrm{H}}(\gamma; E) \triangleq \sum_{\sigma} \gamma(\sigma) E(\sigma)$$

and the Helmholtz entropy function

$$H_{\mathrm{H}}(\gamma) \triangleq -\sum_{\boldsymbol{\sigma} \in \mathcal{K}} \gamma(\boldsymbol{\sigma}) \log \gamma(\boldsymbol{\sigma}).$$

Then, $F_H = -\kappa T \log Z = U_{\mathrm{H}} - TH_{\mathrm{H}}$

The Gibbs average energy function

$$U_{\mathrm{G}}(\beta; E) \triangleq -\sum_{\boldsymbol{\sigma} \in \mathcal{K}} \beta(\boldsymbol{\sigma}) E(\sigma),$$

the Gibbs entropy function

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

and $F_G = U_G - TH_G$

A variational approach to approximating p_{PML}

$$p^{(\psi)}_{\text{PML}} = \arg\min_{p} \min_{\beta} F_{\mathrm{G}}(\beta; p, \psi).$$

But how do we compute this?

[3] J.S. Yedidia, W. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, Jul. 2005

A variational approach to approximating p_{PML}

$$p_{\mathrm{PML}}^{(\psi)} = \arg\min_{p} \min_{\beta} F_{\mathrm{G}}(\beta; p, \psi).$$

But how do we compute this?

Idea: Use approximations that are easy to compute[3].
Specifically, perform minimization w.r.t. β over an easier set.

[3] J.S. Yedidia, W. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, Jul. 2005

A variational approach to approximating p_{PML}

$$p_{\text{PML}}^{(\psi)} = \arg\min_{p} \min_{\beta} F_{\text{G}}(\beta; p, \psi).$$

But how do we compute this?

Idea: Use approximations that are easy to compute[3].
Specifically, perform minimization w.r.t. β over an easier set.

Mean field approximation: Choose β to be a product distribution. Easy to compute.

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

[3] J.S. Yedidia, W. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, Jul. 2005

Generalization to DTMCs

An alternative to p_{SML}: The PML estimate

SML and PML estimates

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

An alternative to p_{SML}: The PML estimate

SML and PML estimates

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

An alternative to p_{SML}: The PML estimate

SML and PML estimates

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

Pattern probability:

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

An alternative to p_{SML}: The PML estimate

SML and PML estimates

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

Pattern probability:

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

PML estimate:

$$p_{\text{PML}}^{(\psi)} = \arg\max_{p} \mathbb{P}(\psi|p).$$

The traditional mean field approximation

$$p^{(\psi)}_{\mathrm{PML}} = \arg\min_{p \in \mathcal{P}} \min_{\beta \in \mathcal{P}'} F_{\mathrm{G}}(\beta; p, \psi).$$

Choose β to be a product distribution on $k \times k$ binary matrices, i.e., $\beta(\sigma) = \prod_{i,l} \beta_{il}(\sigma_{il})$.

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

The traditional mean-field PML estimate is

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

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[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 $b_{il} \triangleq \beta_{il}(1)$.

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

$$F_{\mathrm{MF}}(\mathbf{b}; p, \psi) = -\sum_{\substack{i,j,l,m \\ j\neq i \\ m\neq l}} b_{il} b_{jm} \log p_{lm}^{\mu_{ij}} - \sum_{i,l} b_{il} \log p_{ll}^{\mu_{ii}} + \sum_{i,l} \big(b_{il} \log b_{il} + (1 - b_{il}) \log(1 - b_{il})\big) + \log k. \tag{2}$$

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

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

[4] M. Chertkov and A. Yedidia, "Approximating the permanent with fractional belief propagation," *J. Machine Learning Research*, 2013.

Empirical results

Empirical results

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

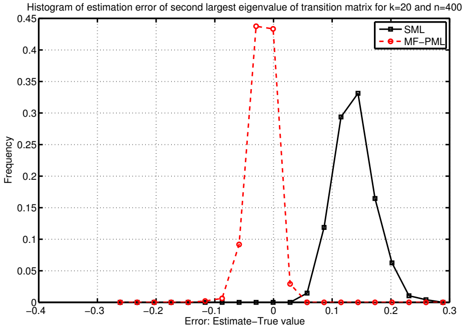


Figure: Histogram of estimation error of absolute second largest eigenvalue of transition matrix for $k = 20$ and $n = 400$.

Empirical results

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

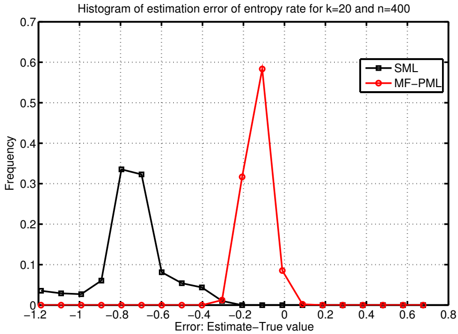


Figure: Histogram of estimation error of entropy rate for $k = 20$ and $n = 400$.

Empirical results

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

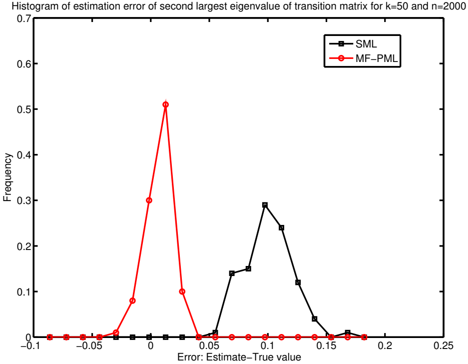


Figure: Histogram of estimation error of absolute second largest eigenvalue of transition matrix for $k = 50$ and $n = 2000$.

Empirical results

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

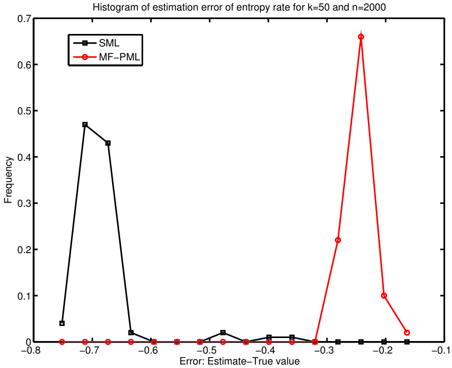


Figure: Histogram of estimation error of entropy rate for $k = 50$ and $n = 2000$.

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!