Distributed Parameter Estimation in Probabilistic Graphical Models

Yariv D. Mizrahi[1] Misha Denil[2] Nando de Freitas[2,3,4]
[1]University of British Columbia [2]University of Oxford
[3]Canadian Institute for Advanced Research [4]Google DeepMind

Motivation

We present foundational theoretical results on distributed parameter estimation for undirected probabilistic graphical models. We introduce a general condition on composite likelihood decompositions of these models which guarantees the global consistency of distributed estimators, provided the local estimators are consistent.

Background

- Maximum Likelihood: Intractable in general graphs.

$$\mathcal{L}^{ML}(\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n \,|\, \boldsymbol{\theta})$$

- Composite Likelihood: Approximate ML with tractable factorisation. Terms are coupled through overlap in $\boldsymbol{\theta}^i$s.

$$\mathcal{L}^{CL}(\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{i=1}^{I} f^i(\mathbf{x}_n, \boldsymbol{\theta}^i)$$

- Consensus/Distributed estimation: Estimate coupled terms independently and combine (Liu and Ihler, 2012) or not (Mizrahi et al. 2014) repeated estimates. Asymptotic results for consensus estimation require

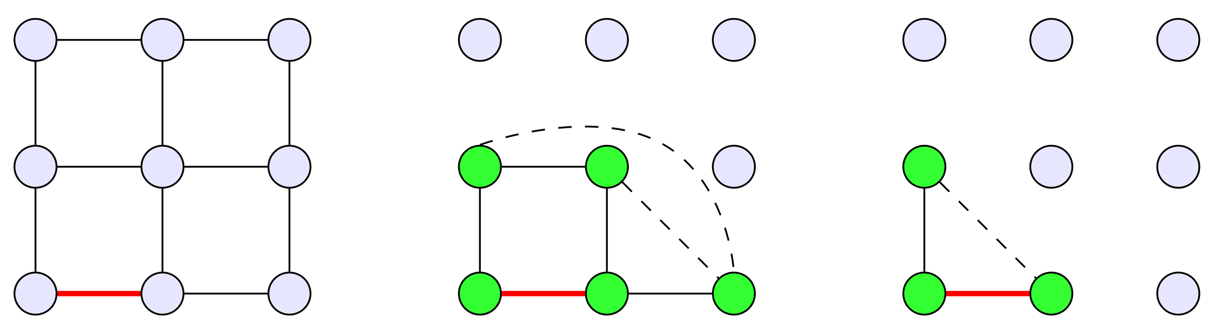
$$(\hat{\boldsymbol{\theta}}^i)_c \xrightarrow{p} \boldsymbol{\theta}_c$$

Strong LAP Argument

Strong LAP Argument Let q be a clique in $\mathcal{G}$ and let $q \subseteq \mathcal{A}_q \subseteq \mathcal{V}$. Suppose $p(\mathbf{x}_\mathcal{V} \,|\, \boldsymbol{\theta})$ and $p(\mathbf{x}_{\mathcal{A}_q} \,|\, \boldsymbol{\alpha})$ are parametrised so that their potentials are normalised with respect to zero and the parameters are identifiable with respect to the potentials. If $\mathcal{A}_q$ satisfies the Strong LAP Condition for q then $\boldsymbol{\theta}_q = \boldsymbol{\alpha}_q$.

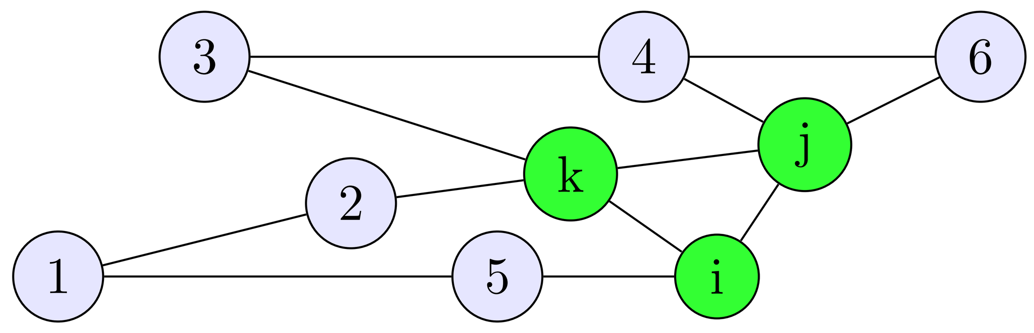
$$\begin{aligned} p(\mathbf{x}_{\mathcal{A}_q} \,|\, \boldsymbol{\theta}) &= \sum_{\mathbf{x}_{\mathcal{V}\setminus\mathcal{A}_q}} p(\mathbf{x}_\mathcal{V} \,|\, \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \sum_{\mathbf{x}_{\mathcal{V}\setminus\mathcal{A}_q}} \exp(-\sum_{c\in\mathcal{C}} E(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c)) \\ &= \frac{1}{Z(\boldsymbol{\theta})} \exp(-E(\mathbf{x}_q \,|\, \boldsymbol{\theta}_q) - \sum_{c\in\mathcal{C}_q\setminus\{q\}} E(\mathbf{x}_c \,|\, \boldsymbol{\theta}_{\mathcal{V}\setminus q})) \end{aligned}$$

$$p(\mathbf{x}_{\mathcal{A}_q} \,|\, \boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \exp(-E(\mathbf{x}_q \,|\, \boldsymbol{\alpha}_q) - \sum_{c\in\mathcal{C}_q\setminus\{q\}} E(\mathbf{x}_c \,|\, \boldsymbol{\alpha}_c))$$



Relative Path Connectivity

Nodes $i, j \in \mathcal{A}$ are path connected with respect to $\mathcal{V} \setminus \mathcal{A}$ if there is a path from i to j through $\mathcal{V} \setminus \mathcal{A}$. Here $\mathcal{A} = \{i, j, k\}$; (k, j) are path connected via $\{3, 4\}$ and (k, i) are path connected via $\{2, 1, 5\}$, but the pair (i, j) are path disconnected with respect to $\mathcal{V} \setminus \mathcal{A}$.



Strong LAP Condition Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph and let $q \in \mathcal{C}$ be a clique of interest. We say that a set $\mathcal{A}$ such that $q \subseteq \mathcal{A} \subseteq \mathcal{V}$ satisfies the strong LAP condition for q if there exist $i, j \in q$ such that i and j are path-disconnected with respect to $\mathcal{V} \setminus \mathcal{A}$.

Normalized Potentials

A Gibbs potential $E(\mathbf{x}_c|\boldsymbol{\theta}_c)$ is said to be normalised with respect to zero if $E(\mathbf{x}_c|\boldsymbol{\theta}_c) = 0$ whenever there exists $t \in c$ such that $\mathbf{x}_t = 0$.

Existence and Uniqueness of Normalized Potentials Griffeath (1976).

References

- Y. Mizrahi, M. Denil, and N. de Freitas. Linear and parallel learning of Markov random fields. In *International Conference on Machine Learning*, 2014.
- Q. Liu and A. Ihler. Distributed parameter estimation via pseudo-likelihood. In *International Conference on Machine Learning*, 2012.
- D. Griffeath. Introduction to random fields. In *Denumerable Markov Chains*, volume 40 of *Graduate Texts in Mathematics*, pages 425–458. Springer, 1976.

Contributions

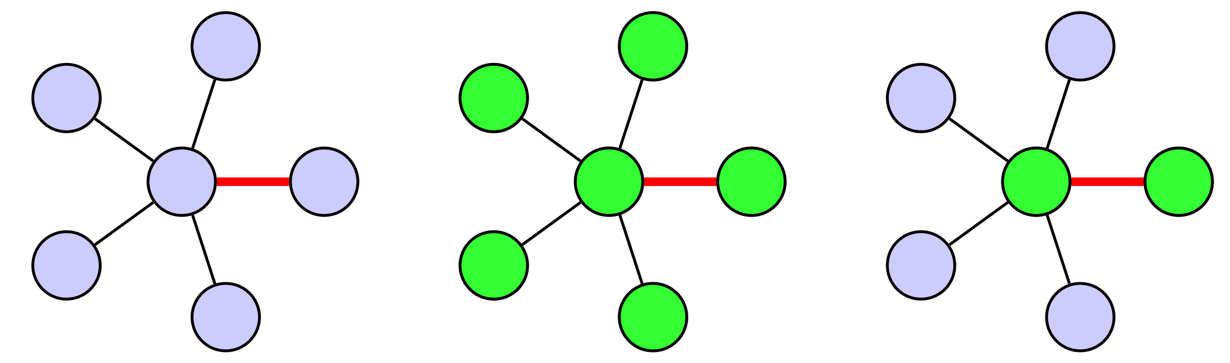
Generalize the LAP algorithm of Mizrahi et al. (2013) to general composite likelihood factorisations.

Provide a strong condition which guarantees a key assumption used in the asymptotic theory of consensus estimators.

Unify the work on LAP and consensus estimation under a common framework.

Provide a new tool for deriving consistent distributed parameter estimation algorithms for Markov random fields.

Auxiliary MRF Structures



- **Left:** Graph with clique of interest highlighted.
- **Centre:** Structure of marginal distribution over smallest neighbourhood satisfying LAP.
- **Right:** Structure of marginal distribution over smallest neighbourhood satisfying Strong LAP.

Conditional LAP

Previous work on LAP (Mizrahi et al. 2013) required joint estimation in marginals over 1-neighbourhoods.

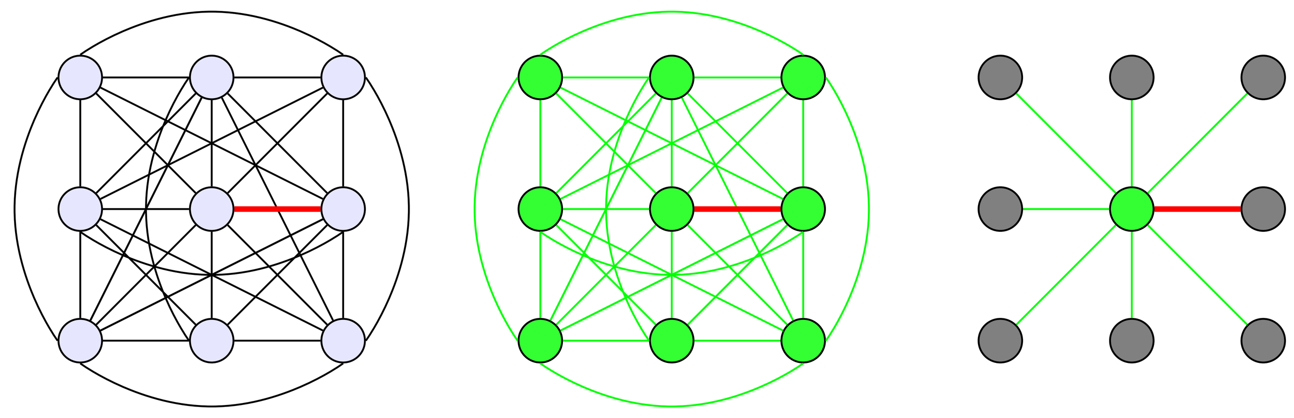
Strong LAP lets us use marginals over smaller domains.

Conditional LAP lets us use conditionals instead of marginals.

Conditional LAP Argument Let q be a clique in $\mathcal{G}$ and let $x_j \in q \subseteq \mathcal{A}_q \subseteq \mathcal{V}$. If $\mathcal{A}_q$ satisfies the Strong LAP Condition for q then $p(\mathbf{x}_\mathcal{V} \,|\, \boldsymbol{\theta})$ and $p(x_j \,|\, \mathbf{x}_{\mathcal{A}_q\setminus\{x_j\}}, \boldsymbol{\alpha})$ share the same normalised potential for q.

$$p(x_j \,|\, \mathbf{x}_{\mathcal{A}_q\setminus\{x_j\}}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_{\mathcal{A}_q} \,|\, \boldsymbol{\theta})}{\sum_{x_j} p(\mathbf{x}_{\mathcal{A}_q} \,|\, \boldsymbol{\theta})}$$

Conditional LAP makes dense graphs tractable, but introduces a different model/data efficiency tradeoff.



- **Left:** Dense graph with clique of interest highlighted.
- **Centre:** Auxiliary MRF required by Strong LAP with joint estimation.
- **Right:** Auxiliary MRF for Conditional LAP (black nodes are conditioned on).

Connection to Distributed Pseudo-Likelihood

Conditional LAP argument says we can use conditional distributions (as in pseudo-likelihood) provided their domains satisfy the Strong LAP Condition. In fact, the pseudo-likelihood domains also satisfy Strong LAP.

Distributed Pseudo-Likelihood Let $q = \{x_1, x_2, .., x_m\}$ be a clique of interest, with 1-neighbourhood $\mathcal{A}_q = q \cup \{\mathcal{N}(x_i)\}_{x_i \in q}$. Then for any $x_j \in q$, the set $q \cup \mathcal{N}(x_j)$ satisfies the Strong LAP Condition for q. Moreover, $q \cup \mathcal{N}(x_j)$ satisfies the Strong LAP Condition for all cliques in the graph that contain x_j.

Statistical Efficiency

Let $\mathcal{A}$ be a set of nodes which satisfies the Strong LAP Condition for q. Let $\hat{\theta}_\mathcal{A}$ be the ML parameter estimate of the marginal over $\mathcal{A}$. If $\mathcal{B}$ is a superset of $\mathcal{A}$, and $\hat{\theta}_\mathcal{B}$ is the ML parameter estimate of the marginal over $\mathcal{B}$. Then (asymptotically):

$$|\theta_q - (\hat{\theta}_\mathcal{B})_q| \leq |\theta_q - (\hat{\theta}_\mathcal{A})_q|.$$

Bottom line: estimating with larger neighbourhoods is never worse, but can be much more expensive.

Asymptotic results of Liu and Ihler (2013) apply.