6.962 Week 6 Tutorial: Approximate Inference Techniques for Graphs with Cycles

Erik Sudderth

March 15, 2001

Outline

1. Introduction to Graphical Models
2. Trees and Belief Propagation
3. Mean Field Theory and Variational Methods
4. Understanding Loopy Belief Propagation
5. Generalized Belief Propagation

Bayesian Estimation

- GIVEN: Noisy observations $\mathcal{Y} = \{y_1, y_2, \ldots\}$ of some "hidden" random variables $\mathcal{X} = \{x_1, x_2, \ldots\}$.

$$\begin{aligned} p(\mathcal{X}) \quad &\Rightarrow \quad \text{prior model} \\ p(\mathcal{Y} \mid \mathcal{X}) \quad &\Rightarrow \quad \text{measurement model} \end{aligned}$$

- Standard estimation problems:

$$\begin{aligned} \widehat{\mathcal{X}}_{\mathrm{MAP}} \quad &= \quad \arg\max_{\mathcal{X}} p(\mathcal{X} \mid \mathcal{Y}) = \arg\max_{\mathcal{X}} p(\mathcal{X})\, p(\mathcal{Y} \mid \mathcal{X}) \\ p(x_i \mid \mathcal{Y}) \quad &= \quad \sum_{\mathcal{X} \setminus x_i} p(\mathcal{X} \mid \mathcal{Y}) = \frac{1}{p(\mathcal{Y})} \sum_{\mathcal{X} \setminus x_i} p(\mathcal{X})\, p(\mathcal{Y} \mid \mathcal{X}) \end{aligned}$$

- Graphical models are a tool for controlling complexity in cases where direct computational costs are prohibitively high:

  Discrete N variables drawn from a finite alphabet with M symbols require $\mathcal{O}(M^N)$ computations.

  Gaussian N vector Gaussian variables of dimension d require $\mathcal{O}((Nd)^3)$ computations.

Graphical Models

- A graph $\mathcal{G}$ is a collection of nodes $\mathcal{S}$ and edges $\mathcal{E}$.
  - Each node $s_i \in \mathcal{S}$ is associated with a random variable $x_i \in \mathcal{X}$.
  - Each edge $(i, j) \in \mathcal{S}$ connects two nodes s_i and s_j.
- Edges are associated with conditional independencies. There are a variety of formalisms for doing this:

  Undirected Graphs (image processing, statistical physics)

  Directed Graphs (artificial intelligence, systems & control)

  Factor Graphs (error correcting codes)
- Many of the results we will discuss have roots in the statistical physics literature, so we will focus on undirected graphs.
- Equivalent ideas can be developed for directed/factor graphs.

Undirected Graphs and Conditional Independence

- For undirected graphs, we define the *neighborhood* of x_i to be

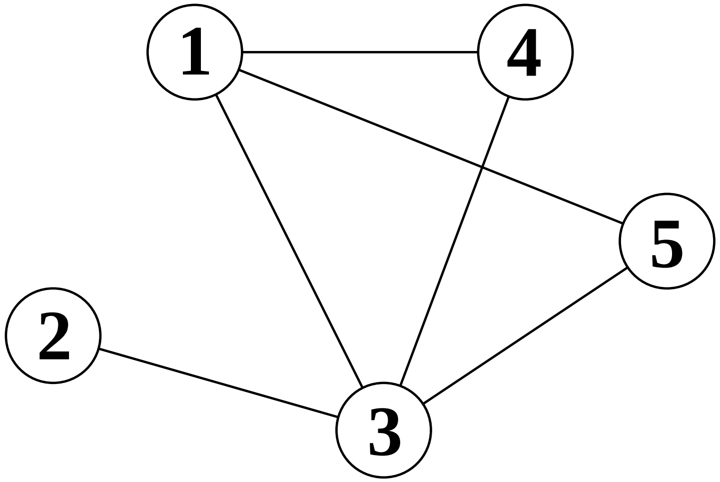
$$\mathcal{N}(x_i) \triangleq \{x_j \in \mathcal{X} \mid (i,j) \in \mathcal{E}\}$$

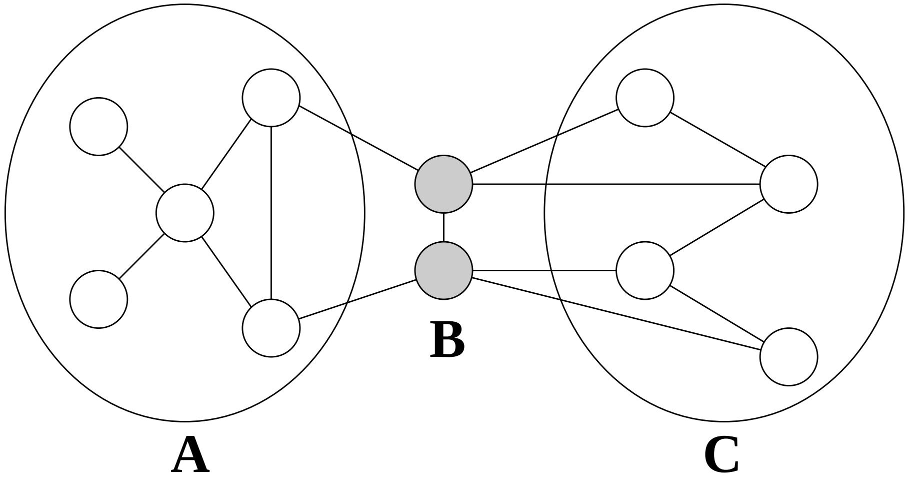
- Conditioned on its immediate neighbors, the probability distribution of a given node is independent of the rest of the graph:

$$p\left(x_i \mid \mathcal{X} \setminus x_i\right) = p\left(x_i \mid \mathcal{N}(x_i)\right)$$

- Alternatively, conditioned on a given set of nodes, the distributions of disjoint subsets of the graph separated by those nodes are independent.

Undirected Graphs

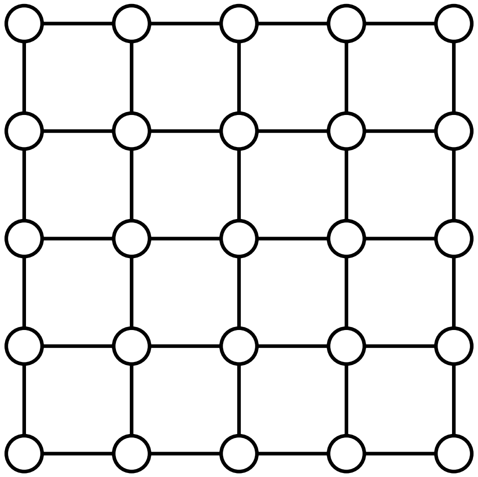




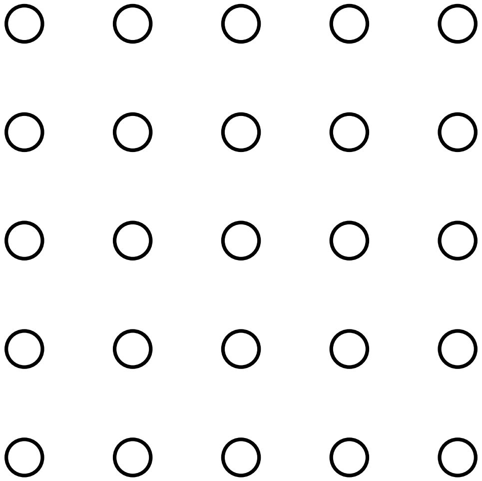
$$
\begin{aligned}
p\,(x_A \mid x_B, x_C\,) &= p\,(x_A \mid x_B\,) \\
p\,(x_A, x_C \mid x_B\,) &= p\,(x_A \mid x_B\,)\,p\,(x_C \mid x_B\,)
\end{aligned}
$$

$$
\begin{aligned}
p\,(x_5 \mid x_1, x_2, x_3, x_4\,) &= p\,(x_5 \mid x_1, x_3\,) \\
p\,(x_2, x_4, x_5 \mid x_1, x_3\,) &= p\,(x_2 \mid x_1, x_3\,)\,p\,(x_4 \mid x_1, x_3\,)\,p\,(x_5 \mid x_1, x_3\,)
\end{aligned}
$$

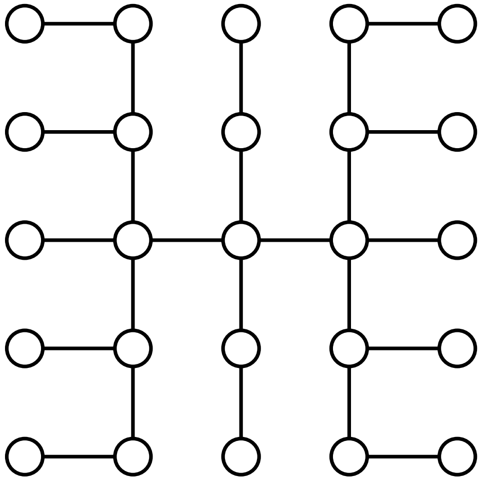
Markov Random Fields



Nearest–Neighbor Grid



Mean Field Approximation



Structured Mean Field

Cliques

- *QUESTION:* How do we determine if a distribution $p(\mathcal{X})$ satisfies the conditional independencies implied by a given graph $\mathcal{G}$?
- To provide an answer, the following definitions will be useful:

Clique A set of nodes in which every node is *directly* connected to every other node in the clique

Maximal Clique A clique which is not a proper subset of any other clique

Hammersley-Clifford Theorem (1971)

$$\begin{aligned} \mathcal{G} \quad &\Rightarrow \quad \text{Undirected graph defined on a set of random variables } \mathcal{X} \\ \mathcal{C} \quad &\Rightarrow \quad \text{Set of all maximal cliques of } \mathcal{G} \\ \psi_C(x_C) \quad &\Rightarrow \quad \text{Arbitrary positive ``clique potential'' function} \end{aligned}$$

- A positive distribution $p(\mathcal{X})$ satisfies the conditional independencies implied by $\mathcal{G}$ if and only if it can be written in the factorized form

$$\begin{aligned} p(\mathcal{X}) &= \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \\ Z &= \sum_{\mathcal{X}} \prod_{C \in \mathcal{C}} \psi_C(x_C) \end{aligned}$$

- We may equivalently define $\phi_C(x_C) \triangleq -\log \psi_C(x_C)$ and write

$$p(\mathcal{X}) = \frac{1}{Z} \exp\left\{ - \sum_{C \in \mathcal{C}} \phi_C(x_C) \right\}$$

Inference on Undirected Graphs

- For convenience, we will assume that all cliques involve at most two nodes, allowing the graph–structured prior $p(\mathcal{X})$ to be written as

$$p(\mathcal{X}) = \frac{1}{Z} \prod_{(i,j)\in\mathcal{E}} \psi_{i,j}(x_i, x_j) \prod_{s_i\in\mathcal{S}} \psi_i(x_i)$$

- If we associate a single measurement with each hidden node, $p(\mathcal{X} \mid \mathcal{Y})$ has the same graphical structure as the prior $p(\mathcal{X})$

$$p(\mathcal{X} \mid \mathcal{Y}) = \frac{1}{Z(\mathcal{Y})} \prod_{(i,j)\in\mathcal{E}} \psi_{i,j}(x_i, x_j) \prod_{s_i\in\mathcal{S}} \psi_i(x_i)\, p(y_i \mid x_i)$$

- Computing marginals $p(x_i)$ and conditional marginals $p(x_i \mid \mathcal{Y})$ are therefore equivalent problems.

$$p(x_i \mid \mathcal{Y}) = \frac{1}{Z(\mathcal{Y})} \sum_{\mathcal{X}\setminus x_i} \prod_{(i,j)\in\mathcal{E}} \psi_{i,j}(x_i, x_j) \prod_{s_i\in\mathcal{S}} \psi_i(x_i)\, p(y_i \mid x_i)$$

Tree–Structured Graphs

- Any tree–structured prior distribution may be factorized as

$$p(\mathcal{X}) = \prod_{(i,j)\in\mathcal{E}} \frac{p(x_i, x_j)}{p(x_i)\, p(x_j)} \prod_{s_i\in\mathcal{S}} p(x_i)$$

- This allows the conditional distribution to be factorized as

$$\begin{aligned} p(\mathcal{X} \mid \mathcal{Y}) &= \frac{1}{p(\mathcal{Y})} \prod_{(i,j)\in\mathcal{E}} \frac{p(x_i, x_j)}{p(x_i)\, p(x_j)} \prod_{s_i\in\mathcal{S}} p(x_i)\, p(y_i \mid x_i) \\ &\triangleq \frac{1}{p(\mathcal{Y})} \prod_{(i,j)\in\mathcal{E}} \psi_{i,j}(x_i, x_j) \prod_{s_i\in\mathcal{S}} \psi_i(x_i) \end{aligned}$$

- Using Bayes' rule and the Markov properties of $\mathcal{G}$, we have

$$\begin{aligned} p(x_i \mid \mathcal{Y}) &= \frac{p(x_i)\, p(\mathcal{Y} \mid x_i)}{p(\mathcal{Y})} = \frac{p(x_i)\, p(y_i \mid x_i)}{p(\mathcal{Y})} \prod_{s_j\in\mathcal{N}(s_i)} p\left(\mathcal{Y}_{j\setminus i} \mid x_i\right) \\ &= \alpha \psi_i(x_i) \prod_{s_j\in\mathcal{N}(s_i)} p\left(\mathcal{Y}_{j\setminus i} \mid x_i\right) \end{aligned}$$

Belief Propagation on Tree–Structured Graphs

- Suppose we associate the conditional likelihoods $p\left(\mathcal{Y}_{j\setminus i} \mid x_i\right)$ with a "message" $m_{j\to i}(x_i)$ that s_j sends to s_i
- Each message $m_{j\to i}(x_i)$ is an M–dimensional vector of real numbers giving the likelihood of each possible value of x_j conditioned on the observations in the subtree rooted at x_i
- Belief Propagation (BP) operates through an iterative "message–passing" procedure:

$$\begin{aligned} p(x_i \mid \mathcal{Y}) &= \alpha \psi_i(x_i) \prod_{s_j \in \mathcal{N}(s_i)} m_{j\to i}(x_i) \\ m_{j\to i}(x_i) &= \sum_{x_j} \psi_{i,j}(x_i, x_j)\, \psi_j(x_j) \prod_{s_k \in \mathcal{N}(s_j)\setminus s_i} m_{k\to j}(x_j) \end{aligned}$$

- On trees, BP converges to the *exact* $p(x_i \mid \mathcal{Y})$ once messages have been allowed to propagate across the entire diameter of the graph
- BIG computational savings $\Rightarrow \mathcal{O}(M^2 N)$ versus $\mathcal{O}(M^N)$ operations

Belief Propagation on Graphs with Cycles

- When $\mathcal{G}$ has cycles, the conditional independencies used to derive the BP algorithm no longer hold

Junction Tree Algorithm Cluster nodes until you have a tree–structured "super–graph" and then run the BP algorithm

- Gives exact answers, but creates intractably large clusters for most interesting architectures

Loopy Belief Propagation Use the BP message passing as an iterative procedure and hope for convergence

- Messages lose their strict probabilistic interpretation, so the standard BP derivation provides no justification for this procedure
- If cycles are long, we expect conditional independencies to "approximately" hold, so loopy BP may give decent approximations
- Excellent empirical performance for *some* problems motivates further investigation

Approximating Distributions

- If $p(\mathcal{X} \mid \mathcal{Y})$ is intractable, we could consider approximating it by a tractable distribution $q(\mathcal{X} \mid \mathcal{Y}, \lambda)$

$$\lambda \Rightarrow \text{parameterizes a class of tractable distributions}$$

- We would like the "best" $q(\mathcal{X} \mid \mathcal{Y}, \lambda)$. One reasonable metric is

$$\begin{aligned} \lambda^* &= \arg\min_{\lambda} D\left(p(\mathcal{X} \mid \mathcal{Y}) \,||\, q(\mathcal{X} \mid \mathcal{Y}, \lambda)\right) \\ &= \arg\min_{\lambda} \sum_{\mathcal{X}} p(\mathcal{X} \mid \mathcal{Y}) \log \frac{p(\mathcal{X} \mid \mathcal{Y})}{q(\mathcal{X} \mid \mathcal{Y}, \lambda)} \end{aligned}$$

- *GOOD NEWS:* If we choose $q(\mathcal{X} \mid \mathcal{Y}, \lambda) = \prod_i q_i(x_i|\lambda_i)$ to be the class of fully factorized distributions, minimizing $D(p \,||\, q)$ recovers the *exact* marginals $q_i(x_i|\lambda_i) = p(x_i \mid \mathcal{Y})$
- *BAD NEWS:* $D(p \,||\, q)$ involves averages with respect to the intractable distribution $p(\mathcal{X} \mid \mathcal{Y})$, and is as hard to deal with as the original problem

A Tractable Approximation

$$\lambda^* = \arg\min_{\lambda} D\left(q \,||\, p\right) = \arg\min_{\lambda} \sum_{\mathcal{X}} q\left(\mathcal{X} \mid \mathcal{Y}, \lambda\right) \log \frac{q\left(\mathcal{X} \mid \mathcal{Y}, \lambda\right)}{p\left(\mathcal{X} \mid \mathcal{Y}\right)}$$

- *GOOD NEWS:* $D\left(q \,||\, p\right)$ takes expectations with respect to the *tractable* distribution $q\left(\mathcal{X} \mid \mathcal{Y}, \lambda\right)$, so this minimization is possible for certain approximating classes
- *BAD NEWS:* Since it weights distance by the approximating distribution, it is not clear if $D\left(q \,||\, p\right)$ will give good approximations. One justification:

$$\begin{aligned}
\log p\left(\mathcal{Y}\right) &= \log \sum_{\mathcal{X}} p\left(\mathcal{X}, \mathcal{Y}\right) \\
&= \log \sum_{\mathcal{X}} q\left(\mathcal{X} \mid \mathcal{Y}, \lambda\right) \frac{p\left(\mathcal{X}, \mathcal{Y}\right)}{q\left(\mathcal{X} \mid \mathcal{Y}, \lambda\right)} \\
&\geq \sum_{\mathcal{X}} q\left(\mathcal{X} \mid \mathcal{Y}, \lambda\right) \log \left[\frac{p\left(\mathcal{X}, \mathcal{Y}\right)}{q\left(\mathcal{X} \mid \mathcal{Y}, \lambda\right)}\right]
\end{aligned}$$

Classical Mean Field Theory

$$p(\mathcal{X}) = \frac{1}{Z} \exp\left\{ - \sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(x_i, x_j) - \sum_{s_i \in \mathcal{S}} \phi_i(x_i) \right\}$$

- The mean field approximation chooses the simplest possible approximating distribution by removing *all* of the edges

$$\begin{aligned} q(\mathcal{X}) &= \prod_i q_i(x_i) \\ D(q \,||\, p) &= \sum_{(i,j)\in\mathcal{E}} \sum_{x_i, x_j} q_i(x_i)\, q_j(x_j)\, \phi_{i,j}(x_i, x_j) + \sum_{s_i\in\mathcal{S}} \sum_{x_i} q_i(x_i)\, \phi_i(x_i) \\ &\quad + \sum_{s_i\in\mathcal{S}} \sum_{x_i} q_i(x_i) \log q_i(x_i) \end{aligned}$$

- Notice that the $\phi_{i,j}(x_i, x_j)$ terms cause the optimization of the $q_i(x_i)$ distributions at different nodes to become coupled

Minimizing the Mean Field Equations

- The mean field approximation removes intractable dependencies in the original graph by adding a set of extra parameters which must then be optimized
- Although the final model $q(\mathcal{X})$ fully decouples the nodes, the optimization process allows the edges in $p(\mathcal{X})$ to be (approximately) accounted for
- For nonhomogeneous MRFs, we use Lagrange multipliers to enforce the normalization constraint $\sum_{x_i} q_i(x_i) = 1$. Taking derivates gives

$$q_i(x_i) = \alpha \psi_i(x_i) \prod_{s_j \in \mathcal{N}(s_i)} \prod_{x_j} \psi_{i,j}(x_i, x_j)^{q_j(x_j)}$$

- We can attempt to solve these equations by iteratively passing $q_i(x_i)$ terms between nodes (reminiscent of BP messages)
- Unfortunately, there are no guarantees of convergence to a global optimum.

Variational Methods

- Our justification for the use of $D(q \,||\, p\,)$ in terms of maximizing a lower bound on $p(\mathcal{Y})$ is a simple example of a *variational method.*
- More generally, we choose a family of tractable functions $g(x; \lambda)$ which each approximate $f(x)$, and then attempt to find λ^* such that $g(x; \lambda^*)$ "best" approximates $f(x)$.
- The notion of "best approximation" is often made precise by choosing $g(x; \lambda)$ which bound $f(x)$, and then optimizing that bound.
- Structured Mean Field methods notice that we do not have to remove *all* of a graph's edges to make calculations tractable.
  - Alternate between calculating the variational parameters λ^* of the best subgraph and running a tractable exact algorithm on the resulting subgraph (Markov chain, tree, etc.)
- No general methods for picking tractable variational classes which are also good approximations (often must exploit specific graphical structures, functional forms of clique potentials, etc.)

Gibbs Free Energy

$$G \triangleq \sum_{\mathcal{X}} p(\mathcal{X}) \left[\sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(x_i, x_j) + \sum_{s_i\in\mathcal{S}} \phi_i(x_i) \right] - \left[-\sum_{\mathcal{X}} p(\mathcal{X}) \log p(\mathcal{X}) \right]$$

- Exactly minimizing G with respect to $p(\mathcal{X})$ recovers

$$p(\mathcal{X}) = \frac{1}{Z} \exp\left\{ -\sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(x_i, x_j) - \sum_{s_i\in\mathcal{S}} \phi_i(x_i) \right\}$$

- For complex MRFs, this minimization is intractable. Physicists often minimize an approximate free energy to produce $q(\mathcal{X}) \approx p(\mathcal{X})$
- If we assume $q(\mathcal{X}) = \prod_i q_i(x_i)$, we get the mean field free energy

$$\begin{aligned} G_{\mathrm{MF}} &= \sum_{\mathcal{X}} \prod_i q_i(x_i) \left[\sum_{(j,k)\in\mathcal{E}} \phi_{j,k}(x_j, x_k) + \sum_{s_k\in\mathcal{S}} \phi_k(x_k) \right] \\ &\quad + \sum_{\mathcal{X}} \prod_i q_i(x_i) \log \prod_i q_i(x_i) \end{aligned}$$

Bethe Free Energy

- For tree–structured graphs, the exact free energy is given by

$$
\begin{aligned}
p\left(\mathcal{X}\right) &= \prod_{(i,j)\in\mathcal{E}} \frac{p\left(x_i, x_j\right)}{p\left(x_i\right) p\left(x_j\right)} \prod_{s_i\in\mathcal{S}} p\left(x_i\right) \\
G_{\mathrm{B}} &= \sum_{(i,j)\in\mathcal{E}} \sum_{x_i, x_j} q_{i,j}\left(x_i, x_j\right) \left[\log \frac{q_{i,j}\left(x_i, x_j\right)}{q_i\left(x_i\right) q_j\left(x_j\right)} + \phi_{i,j}\left(x_i, x_j\right)\right] \\
&\quad + \sum_{s_i\in\mathcal{S}} \sum_{x_i} q_i\left(x_i\right) \left[\log q_i\left(x_i\right) + \phi_i\left(x_i\right)\right]
\end{aligned}
$$

- We cannot write the free energy for graphs with cycles solely in terms of $q_i\left(x_i\right)$ and $q_{i,j}\left(x_i, x_j\right)$ because of the partition function Z
- Bethe Approximation $\Rightarrow$ Use the tree–structured free energy G_{B} even though $\mathcal{G}$ is *not* tree–structured.

Minimization of Bethe Free Energy

- In order to minimize G_{B}, we add Lagrange multipliers to enforce the various marginalization constraints

$$\begin{aligned}
\lambda_{i,j}(x_j) \quad &\Longleftrightarrow \quad \sum_{x_i} q_{i,j}(x_i, x_j) = q_j(x_j) \\
\lambda_{j,i}(x_i) \quad &\Longleftrightarrow \quad \sum_{x_j} q_{i,j}(x_i, x_j) = q_i(x_i) \\
\gamma_i \quad &\Longleftrightarrow \quad \sum_{x_i} q_i(x_i) = 1 \\
\gamma_{ij} \quad &\Longleftrightarrow \quad \sum_{x_i, x_j} q_{i,j}(x_i, x_j) = 1
\end{aligned}$$

Bethe Free Energy & Belief Propagation

- Taking the derivative of the resulting Lagrangian and manipulating

$$
\begin{aligned}
q_i(x_i) &= \alpha \exp\left\{ \phi_i(x_i) + \frac{1}{|\mathcal{N}(s_i)| - 1} \sum_{s_j \in \mathcal{N}(s_i)} \lambda_{j,i}(x_i) \right\} \\
q_{i,j}(x_i, x_j) &= \alpha \exp\left\{ \phi_{i,j}(x_i, x_j) + \phi_i(x_i) + \phi_j(x_j) + \lambda_{i,j}(x_j) + \lambda_{j,i}(x_i) \right\}
\end{aligned}
$$

- Recall that the BP update equations are given by

$$
\begin{aligned}
p(x_i \mid \mathcal{Y}) &= \alpha \psi_i(x_i) \prod_{s_j \in \mathcal{N}(s_i)} m_{j \to i}(x_i) \\
m_{j \to i}(x_i) &= \sum_{x_j} \psi_{i,j}(x_i, x_j)\, \psi_j(x_j) \prod_{s_k \in \mathcal{N}(s_j) \setminus s_i} m_{k \to j}(x_j)
\end{aligned}
$$

- These equations are *exactly* equivalent if we make the following association between Lagrange multipliers and messages:

$$
\lambda_{i,j}(x_j) = \sum_{s_k \in \mathcal{N}(s_j) \setminus s_i} \log m_{k \to j}(x_j)
$$

Duality of KL and Free Energy

- The mean field approximation can be derived by minimizing either an approximate free energy G_{MF} or the KL divergence $D\left(q \,||\, p\right)$ between the fully factorized and true distributions.
- QUESTION: Is there a KL interpretation of the Bethe tree approximation?
- Minimizing G_{B} is equivalent to minimizing an approximate $D\left(q \,||\, p\right)$, where the approximation arises from

$$\log q\left(\mathcal{X}\right) \approx \sum_{(i,j)\in\mathcal{E}} \log \frac{q_{i,j}\left(x_i, x_j\right)}{q_i\left(x_i\right) q_j\left(x_j\right)} + \sum_{s_i\in\mathcal{S}} \log q_i\left(x_i\right)$$

  (derived from the Möbius inversion formula)
- These are a few examples of a much deeper duality between free energy and relative entropy.

Generalized Belief Propagation

- We have interpreted BP messages as exponentiated Lagrange multipliers which enforce a set of local consistency constraints.
- BP only considers pairwise consistency. This ensures that the beliefs at all pairs of nodes will be consistent, but does *not* guarantee that the beliefs at larger clusters of nodes will be consistent.
- Yedidia, Freeman, and Weiss (NIPS 2000) have introduced a new class of generalized belief propagation (GBP) algorithms which enforce the consistency of larger clusters by passing more messages
  - Clustering can produce good estimates for problems where BP gives poor results or fails to converge
  - Changing cluster sizes allows computational complexity and solution accuracy to be balanced
- GBP algorithms correspond to a class of higher–order free energy approximations known as Kikuchi approximation

Double Loop Algorithms

- Yuille has recently introduced a "double–loop" algorithm which is guaranteed to converge to a local minimum of the Bethe free energy.
- Functions by iterating between an "inner" and an "outer" loop:
  - Inner loop determines a set of Lagrange multipliers
  - Outer loop updates beliefs based on Lagrange multipliers
- Derivation depends on decomposition of free energy into a sum of convex and concave parts
- Loopy BP can be viewed as an approximation to double–loop in which the two loops are collapsed together
- GBP–style generalizations of double–loop are also proposed

Summary

- For tree–structured graphs, BP messages have a direct probabilistic interpretation as likelihoods
- For graphs with cycles, BP messages can be interpreted as Lagrange multipliers which attempt to enforce local consistency constraints
- Variational methods, including the mean field approximation, are closely related to BP, and may offer attractive alternatives in certain situations
- KL divergence and Gibbs free energy are connected in a number of very fundamental ways