CS242: Probabilistic Graphical Models
Lecture 20: Variational Methods and Naïve Mean Field

Instructor: Erik Sudderth
Brown University Computer Science
November 11, 2014
Simplifying Assumptions

\[ p(x) = \frac{1}{Z} \prod_{(s,t) \in E} \psi_{st}(x_s, x_t) \prod_{s \in V} \psi_s(x_s) \quad x_s \in \{1, \ldots, K_s\} \]

- Inference problems with *discrete hidden variables*
- All dependencies are at most *pairwise*, represented by node potentials (vectors) and edge potentials (matrices)
- Support for *arbitrary graph structures*

**How Important are These Assumptions?**

- All algorithms are straightforward to generalize to higher-order factors, but become harder to derive, explain, and implement
- All algorithms can be generalized to continuous Gaussian variables
- For continuous non-Gaussian variables, some algorithms have no known generalizations, others are more complex (examples will follow)
Reminder: Information Theory

- The **entropy** is a natural measure of the inherent uncertainty (difficulty of compression) of some random variable:

\[ H(p_x) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \geq 0 \quad \text{discrete entropy (concave)} \]

- The **relative entropy** or **Kullback-Leibler (KL) divergence** is a non-negative, but asymmetric, “distance” between a given pair of probability distributions:

\[ D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad D(p \parallel q) \geq 0 \]

The KL divergence equals zero if and only if \( p(x) = q(x) \) almost everywhere.

- The **mutual information** measures dependence between a pair of random variables:

\[ I(p_{xy}) = D(p_{xy} \parallel p_x p_y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \geq 0 \]

Zero if and only if variables are independent.
Variational Approximate Inference

\[ p(x) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s) \]

- Choose a family of approximating distributions that is tractable. The simplest example:
  \[ q(x) = \prod_{s \in \mathcal{V}} q_s(x_s) \quad \text{approximate as independent, form of marginals unconstrained} \]

- Define a distance to measure the quality of different approximations. Two possibilities:
  \[ D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad D(q \parallel p) = \sum_x q(x) \log \frac{q(x)}{p(x)} \]

- Find the approximation minimizing the chosen distance:
  \[ q^* = \arg \min_q D(p \parallel q) \quad q^* = \arg \min_q D(q \parallel p) \]
Fully Factorized Variational Approximations

\[ p(x) = \text{any distribution on } N \text{ variables} \]

\[ q(x) = \prod_{s \in \mathcal{V}} q_s(x_s) \]

\[ \text{True Node Marginals:} \]

\[ p_s(x_s) = \sum_{x_{\mathcal{V}\setminus s}} p(x_s, x_{\mathcal{V}\setminus s}) \]

\[ x_{\mathcal{V}\setminus s} = \{x_t \mid t \in \mathcal{V}, t \neq s\} \]

- In information theory, if the true distribution of a source is \( p(x) \) and we model it via \( q(x) \), the compression cost (measured in extraneous bits/nats) equals

\[ D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log p(x) - p(x) \log q(x) \]

\[ = \sum_{s \in \mathcal{V}} H(p_s) - H(p) + \sum_{s \in \mathcal{V}} D(p_s \parallel q_s) \]

Marginal KLD:

\[ q_s^*(x_s) = p_s(x_s) \]

True marginals are best approx, but how do we compute them?
Derivation: Fully Factorized Optima

\[ D(p \parallel q) = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) \]

\[ = -H(p) - \sum_x p(x) \log \prod_s q_s(x_s) \]

\[ = -H(p) - \sum_x \sum_s p(x) \log q_s(x_s) \]

\[ = -H(p) + \sum_s \sum_{x_s} \left[ p_s(x_s) \log q_s(x_s) + p_s(x_s) \log p(x_s) - p_s(x_s) \log q(x_s) \right] \]

\[ = \sum_s H(p_s) - H(p) + \sum_s \sum_{x_s} p_s(x_s) \log \frac{p_s(x_s)}{q_s(x_s)} \]

\[ = D(p_s \parallel q_s) \]
Reminder: Jensen’s Inequality

\[ f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \]

Valid for both discrete (expectations are sums) and continuous (expectations are integrals) random variables, for any convex function \( f \).

\[ \ln(\mathbb{E}[X]) \geq \mathbb{E}[\ln(X)] \]

The logarithm is concave.
Mean Field Variational Bounds

- Consider posterior distribution of hidden variables $x$ given observations $y$:

$$D(q(x) \| p(x \mid y)) = \sum_x q(x) \log \frac{q(x)}{p(x \mid y)}$$

- Motivate this distance via \textit{marginal likelihood} of observed data:

$$\log p(y) = \log \sum_x p(x, y) = \log \sum_x q(x) \frac{p(x, y)}{q(x)} \geq \sum_x q(x) \log \frac{p(x, y)}{q(x)}$$

\textit{Definition of Marginal} \hspace{1cm} \textit{Multiply by One} \hspace{1cm} \textit{Jensen’s Inequality}

$$\log p(y) \geq -D(q(x) \| p(x \mid y)) + \log p(y)$$

- Minimizing KL $D(q \| p)$ \textit{maximizes a lower bound} on data likelihood
- Optimizing over all possible distributions $q(x)$ recovers true posterior
- Optimizing over restricted family $q(x)$ recovers approximate posterior
Inference and Free Energies

- Boltzmann distribution from statistical mechanics (at unit temperature):
  \[ p(x) = \frac{1}{Z} \exp\{-E(x)\} \]
  relates probabilities \( p(x) \) to energies \( E(x) \)

- For any joint distribution on states, express KL divergence as:
  \[
  D(q || p) = \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x)
  = -H(q) + \sum_x q(x) E(x) + \log Z
  \]
  Variational inference is “energy minimization”
  
  - Negative Entropy
  - Average Energy
  - Normalization (Partition Function)

- Dropping the fixed normalizer \( Z \), KL divergence becomes a free energy
- **GOAL:** Find families of distributions with “tractable” free energies
Average Energies versus Free Energies

\[ p(x) = \frac{1}{Z} \exp\{-E(x)\} \quad x \in \mathcal{X} \]

- Minimizing average energy recovers mode (MAP estimate):
  \[ q^*(x) = \arg \min \sum_x q(x) E(x) = \begin{cases} 1 & \text{if } x = \hat{x} \\ 0 & \text{if } x \neq \hat{x} \end{cases} \]
  \[ \hat{x} = \arg \min_x E(x) = \arg \max_x p(x) \]

- Maximizing entropy (minimizing neg-entropy) recovers uniform distribution:
  \[ q^*(x) = \arg \max_q H(q) = |\mathcal{X}|^{-1} \]

- Minimizing free energy recovers full joint distribution (and marginals):
  \[ q^*(x) = \arg \min_q -H(q) + \sum_x q(x) E(x) = \arg \min_q D(q \mid\mid p) = p(x) \]
Naïve Mean Field Free Energy

- Express pairwise MRF in exponential family (“energy”) form:

\[
p(x) = \frac{1}{Z} \exp \left\{ - \sum_{(s,t) \in \mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s \in \mathcal{V}} \phi_s(x_s) \right\}
\]

\[
\phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t) \\
\phi_s(x_s) = -\log \psi_s(x_s)
\]

- A naïve mean field method approximates distribution as fully factorized:

\[
q(x) = \prod_{s \in \mathcal{V}} q_s(x_s)
\]

\[
q_s(x_s = k) = \mu_{sk} \geq 0, \quad \sum_{k=1}^{K_s} \mu_{sk} = 1.
\]

- For factorized distributions, the entropy has a simple form:

\[
H(q) = \sum_{s \in \mathcal{V}} H_s(q_s) = -\sum_{s \in \mathcal{V}} \sum_{x_s} q_s(x_s) \log q_s(x_s)
\]

- For a pairwise MRF, the average entropy (expected sufficient stats) equals

\[
\sum_{x} q(x) E(x) = \sum_{(s,t) \in \mathcal{E}} \sum_{x_s, x_t} q_s(x_s) q_t(x_t) \phi_{st}(x_s, x_t) + \sum_{s \in \mathcal{V}} \sum_{x_s} q_s(x_s) \phi_s(x_s)
\]
Naïve Mean Field Updates

Free Energy:  \( F(\mu) = -H(\mu) + E(\mu) \) \( \quad q(x) = \prod_{s \in \mathcal{V}} q_s(x_s) \)

Entropy: \( H(\mu) = - \sum_{s \in \mathcal{V}} \sum_k \mu_{sk} \log \mu_{sk} \) \( \quad q_s(x_s = k) = \mu_{sk} \)

Average Energy: \( E(\mu) = \sum_{(s,t) \in \mathcal{E}} \sum_{k,\ell} \mu_{sk} \mu_{tl} \phi_{st}(k, \ell) + \sum_{s \in \mathcal{V}} \sum_k \mu_{sk} \phi_s(k) \)

• Constraints which any feasible solution must satisfy:
  \[ \sum_k \mu_{sk} = 1 \text{ for all nodes } s \in \mathcal{V}. \]

• Lagrangian encoding objective and constraints:
  \[ \mathcal{L}(\mu) = -H(\mu) + E(\mu) + \sum_{s \in \mathcal{V}} \lambda_s \left( 1 - \sum_k \mu_{sk} \right) \]

• Coordinate descent: Optimize one node marginal fixing others, iterate.
Naïve Mean Field Updates

**Free Energy:**
\[
F(\mu) = -H(\mu) + E(\mu)
\]
\[
q(x) = \prod_{s \in \mathcal{V}} q_s(x_s)
\]

**Entropy:**
\[
H(\mu) = - \sum_{s \in \mathcal{V}} \sum_{k} \mu_{sk} \log \mu_{sk}
\]
\[
q_s(x_s = k) = \mu_{sk}
\]

**Average Energy:**
\[
E(\mu) = \sum_{(s,t) \in \mathcal{E}} \sum_{k,\ell} \mu_{sk} \mu_{t\ell} \phi_{st}(k, \ell) + \sum_{s \in \mathcal{V}} \sum_{k} \mu_{sk} \phi_s(k)
\]

- Optimal value of distribution at node \(s\), fixing marginals for other nodes:
\[
\log \mu_{sk} + 1 - \lambda_s = -\phi_s(k) - \sum_{t \in \Gamma(s)} \sum_{\ell} \mu_{t\ell} \phi_{st}(k, \ell)
\]
\[
\mu_{sk} \propto \psi_s(k) \prod_{t \in \Gamma(s)} \exp \left\{ - \sum_{\ell} \mu_{t\ell} \phi_{st}(k, \ell) \right\}
\]

- Lagrange multiplier induces normalization to sum to one
Derivation: Naïve Mean Field Updates

\[ L(\mu) = -H(\mu) + E(\mu) + \sum_{\mu_{sk}} \lambda_s (1 - \sum_k \mu_{sk}) \]

\[ \frac{\partial}{\partial \mu_{sk}} [-H(\mu)] = \frac{1}{\mu_{sk}} \log \mu_{sk} = 1 + \log \mu_{sk} \]

\[ \frac{\partial}{\partial \mu_{sk}} E(\mu) = \Phi_s (k) + \sum_{\text{ter}(s)} \sum_l \mu_{te} \Phi_{st}(k, l) \]

\[ \frac{\partial}{\partial \mu_{sk}} E(\mu) = 1 + \log \mu_{sk} + \Phi_s (k) + \sum_{\text{ter}(s)} \sum_l \mu_{te} \Phi_{st}(k, l) - \lambda_s = 0 \]

\[ \log \mu_{sk} + 1 - \lambda_s = -\Phi_s (k) - \sum_{\text{ter}(s)} \sum_l \mu_{te} \Phi_{st}(k, l) \]

\[ e^{(1-\lambda_s) \mu_{sk}} = \prod_{\text{ter}(s)} \exp \left[ -\sum_l \mu_{te} \Phi_{st}(k, l) \right] \]

\[ \mu_{sk} = \Phi_s (k) \prod_{\text{ter}(s)} \exp \left[ -\sum_l \mu_{te} \Phi_{st}(k, l) \right] \]

Choose \( \lambda_s \) to rescale so normalizati satisfied
Mean Field as Message Passing

\[ p(x) = \frac{1}{Z} \prod_{(s,t) \in E} \psi_{st}(x_s, x_t) \prod_s \psi_s(x_s) \]

\[ \phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t) \]

\[ q_i(x_i) \propto \psi_i(x_i) \prod_{j \in \Gamma(i)} m_{ji}(x_i) \]

\[ m_{ji}(x_i) \propto \exp \left\{ -\sum_{x_j} \phi_{ij}(x_i, x_j) q_j(x_j) \right\} \]

- Compared to belief propagation, has identical formula for estimating marginals from messages, but a different message update equation
- If neighboring marginals degenerate to single state, recover Gibbs sampling message
Reminder: Sum-Product Belief Propagation

\[
p(x) = \frac{1}{Z} \prod_{(s, t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s)
\]

\[
q_t(x_t) \propto \psi_t(x_t) \prod_{u \in \Gamma(t)} m_{ut}(x_t)
\]

Belief Propagation (Sum-Product) Messages:

\[
m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \psi_t(x_t) \prod_{u \in \Gamma(t) \setminus s} m_{ut}(x_t)
\]

\[
m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)}
\]

Replaces geometric (log-domain) mean by arithmetic mean, and divides by incoming message to avoid “double-counting” information
**Mean Field versus Belief Propagation**

\[ p(x) = \frac{1}{Z} \prod_{(s,t) \in E} \psi_{st}(x_s, x_t) \prod_{s \in V} \psi_s(x_s) \]

\[ \phi_{st}(x_s, x_t) = -\psi_{st}(x_s, x_t) \]

\[ q_t(x_t) \propto \psi_t(x_t) \prod_{u \in \Gamma(t)} m_{ut}(x_t) \]

**Mean Field:**

\[ m_{ts}(x_s) \propto \exp \left\{ -\sum_{x_t} \phi_{st}(x_s, x_t) q_t(x_t) \right\} \]

\[ m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)} \]

**Belief Propagation:**

- Guaranteed to converge for general graphs, always lower-bounds partition function, but approximate even on trees
- Produces exact marginals for any tree, but for general graphs no guarantees of convergence or accuracy
- **Goal:** Can we justify and generalize loopy BP?