Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 25:
Reweighted Max-Product & LP Relaxations,
Survey of Advanced Topics
Max Marginals

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

- A **max-marginal** gives the probability of the most likely state in which some variables are constrained to take specified values:

  \[ \nu_s(x_s) = \max_{\{x'|x'_s = x_s\}} p(x'_1, x'_2, \ldots, x'_N) \]

  \[ \nu_{st}(x_s, x_t) = \max_{\{x'|x'_s = x_s, x'_t = x_t\}} p(x'_1, x'_2, \ldots, x'_N) \]

- For a pairwise MRF, a solution \( \hat{x} \) is guaranteed to be one (of possibly many) global MAP estimates if and only if:

  \[ \hat{x}_s \in \arg \max_{x_s} \nu_s(x_s) \quad s \in \mathcal{V} \]

  \[ (\hat{x}_s, \hat{x}_t) \in \arg \max_{x_s, x_t} \nu_{st}(x_s, x_t) \quad (s, t) \in \mathcal{E} \]
Belief Propagation (Max-Product)

Max-Marginals:

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

Messages:

\[ m_{ts}(x_s) \propto \max_{x_t} \psi_{st}(x_s, x_t) \psi_t(x_t) \prod_{u \in \Gamma(t) \setminus s} m_{ut}(x_t) \]
Belief Propagation (Min-Sum)

**Negative Log-Max-Marginals:**

\[ \bar{\nu}_t(x_t) = \phi_t(x_t) + \sum_{u \in \Gamma(t)} \bar{m}_{ut}(x_t) \]

\[ \phi_t(x_t) = -\log \psi_t(x_t) \]

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

\[ \bar{m}_{ts}(x_s) = \min_{x_t} \phi_{st}(x_s, x_t) + \phi_t(x_t) + \sum_{u \in \Gamma(t) \setminus s} \bar{m}_{ut}(x_t) \]
The Generalized Distributive Law

• A commutative semiring is a pair of generalized “multiplication” and “addition” operations which satisfy:
  - Commutative: \( a + b = b + a \) \( a \cdot b = b \cdot a \)
  - Associative: \( a + (b + c) = (a + b) + c \) \( a \cdot (b \cdot c) = (a \cdot b) \cdot c \)
  - Distributive: \( a \cdot (b + c) = a \cdot b + a \cdot c \)
(Why not a ring? May be no additive/multiplicative inverses.)

• Examples:

<table>
<thead>
<tr>
<th>Addition</th>
<th>Multiplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>sum</td>
<td>product</td>
</tr>
<tr>
<td>max</td>
<td>product</td>
</tr>
<tr>
<td>max</td>
<td>sum</td>
</tr>
<tr>
<td>min</td>
<td>sum</td>
</tr>
</tbody>
</table>

• For each of these cases, our factorization-based dynamic programming derivation of belief propagation is still valid
• Leads to *max-product and min-sum belief propagation* algorithms for exact MAP estimation in trees
Max-Product to MAP Estimates

**Global Directed Factorization:**
- Choose some node as the root of the tree, order by depth
- Define directed factorization from root to leaves:

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

\[
p(x) = p(x_{Root}) \prod_s p(x_s \mid x_{Pa(s)})
\]

**Bottom-Up Message Passing:**
- Pass max-product messages recursively from leaves to root
- Find max-marginal of root node:

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

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

**Top-Down Recursive Selection:**
- Take maximizing root, then maximize by depth given parent:

\[
\nu_s(x_s \mid X_t = \hat{x}_t, t = Pa(s)) \propto \psi_{ts}(\hat{x}_t, x_s) \psi_s(x_s) \prod_{u \in \Gamma(s) \setminus t} m_{us}(x_s)
\]
Discriminative Graphical Models

A CRF is trained to match marginals:

\[ p(y \mid x, \theta) = \exp \left\{ \sum_{f \in F} \theta_f^T \phi_f(y_f, x) - A(\theta, x) \right\} \]

- A max-margin Markov network or structural SVM adapts hinge loss, and is trained via MAP estimation
Approximate MAP Estimation

- Greedy coordinate ascent: *Iterative Conditional Modes (ICM)*

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

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

\[ \hat{x}_i = \arg \max_{x_i} q_i(x_i) \]

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

- Limit of both Gibbs sampling and mean field in limit \( \beta \to \infty \)
- Physical interpretation: Temperature \( \beta^{-1} \to 0 \)
- The *simulated annealing* method applies Gibbs sampling as temperature is (very, very slowly) decreased
Marginalization as Convex Optimization

\[ p(x \mid \theta) = \exp\{\theta^T \phi(x) - A(\theta)\} \]
\[ A(\theta) = \log \sum_{x \in \mathcal{X}} \exp\{\theta^T \phi(x)\} \]
\[ \mathcal{M} = \text{conv}\{\phi(x) \mid x \in \mathcal{X}\} \]

- Express log-partition as optimization over all distributions \( Q \)

\[ A(\theta) = \sup_{q \in Q} \left\{ \sum_{x \in \mathcal{X}} \theta^T \phi(x) q(x) - \sum_{x \in \mathcal{X}} q(x) \log q(x) \right\} \]

Jensen’s inequality gives arg max: \( q(x) = p(x \mid \theta) \)

- More compact to optimize over relevant sufficient statistics:

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^T \mu + H(p(x \mid \theta(\mu))) \right\} \]
\[ \mu = \sum_{x \in \mathcal{X}} \phi(x) q(x) = \sum_{x \in \mathcal{X}} \phi(x) p(x \mid \theta(\mu)) \]
MAP Estimation as Convex Optimization

\[ p(x \mid \theta) = \exp \{ \theta^T \phi(x) - A(\theta) \} \]

\[ \max_{x \in \mathcal{X}} \theta^T \phi(x) = \max_{x \in \mathcal{X}} p(x \mid \theta) \]

\[ \max_{x \in \mathcal{X}} \theta^T \phi(x) = \max_{\mu \in \mathcal{M}} \theta^T \mu \]

\[ \mathcal{M} = \text{conv} \{ \phi(x) \mid x \in \mathcal{X} \} \]

- This is a linear program: Maximization of a linear function over a convex polytope, with one vertex for each \( x \in \mathcal{X} \)
- No need to directly consider entropy for MAP estimation
- MAP also arises as limit of standard variational objective:

\[ \max_{x \in \mathcal{X}} \theta^T \phi(x) = \lim_{\beta \to \infty} \frac{A(\beta \theta)}{\beta} \]

\[ A(\beta \theta) = \sup_{\mu \in \mathcal{M}} \left\{ \beta \theta^T \mu + H(p(x \mid \theta(\mu))) \right\} \]

Convexity allows order of limit and optimization to interchange
Tree-Based Outer Approximations

- For some graph $G$, denote true marginal polytope by $\mathbb{M}(G)$
- Associate marginals with nodes and edges, and impose the following *local consistency* constraints $\mathbb{L}(G)$
  \[
  \sum_{x_s} \mu_s(x_s) = 1, \quad s \in \mathcal{V} \quad \mu_s(x_s) \geq 0, \mu_{st}(x_s, x_t) \geq 0
  \]
  \[
  \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s), \quad (s, t) \in \mathcal{E}, x_s \in \mathcal{X}_s
  \]
- For any graph, this is a *convex* outer bound: $\mathbb{M}(G) \subseteq \mathbb{L}(G)$
- For any tree-structured graph $T$, we have $\mathbb{M}(T) = \mathbb{L}(T)$
MAP Linear Programming Relaxations

\[
\max_{x \in \mathcal{X}} \theta^T \phi(x) = \max_{\mu \in \mathcal{M}(G)} \theta^T \mu \leq \max_{\tau \in \mathcal{L}(G)} \theta^T \tau
\]

- Spanning tree polytope has linear number of constraints, so we can solve linear program in polynomial time
- If we find “integral” vertex of original polytope, we have certificate guaranteeing solution of original MAP problem
- Otherwise, “round” solution to find approximate MAP estimate

Possible Efficient Solution: Reweighted Max-Product BP

\[
m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t)^{1/\rho_{st}} \frac{q_t(x_t)}{m_{st}(x_t)} \\
m_{ts}(x_s) \propto \max_{x_t} \psi_{st}(x_s, x_t)^{1/\rho_{st}} \frac{q_t(x_t)}{m_{st}(x_t)}
\]

Edge appearance weights as in reweighted sum-product
Informal summary of results of Wainwright et al., Weiss et al.:

- Zero-temperature limit of “convexified” sum-product algorithms are guaranteed to solve MAP LP relaxation
- Reweighted max-product closely related, but not identical
- Standard max-product only approximates LP relaxation
Current Research: Structure Learning

Unknown Graphs for Known Variables

• Objective: Likelihood with MDL or Bayesian penalty
• Classic approach: Stochastic search in space of graphs
• Modern approach: Convex optimization with sparsity priors, which encourage some parameters to be set to zero

Deep Learning

• Hierarchical models, with observations at finest scale, and many layers of hidden variables
• Classic neural networks: Directed graphical models
• Modern restricted Boltzmann machines: Undirected models
• Challenge: Extraordinarily non-convex, extensive heuristics (partially understood) required to avoid local optima

Bayesian Nonparametrics

• Allow model complexity to grow as observations observed
• “Infinite” models via stochastic process priors on distributions