Probabilistic Graphical Models

David Sontag

New York University

Lecture 8, March 28, 2012

From last lecture: Variational methods

- Suppose that we have an arbitrary graphical model:

$$p(\mathbf{x};\theta) = \frac{1}{Z(\theta)} \prod_{\mathbf{c}\in C} \phi_c(\mathbf{x_c}) = \exp\Big(\sum_{\mathbf{c}\in C} \theta_c(\mathbf{x_c}) - \ln Z(\theta)\Big)$$

- Finding the *approximating distribution* $q(\mathbf{x}) \in Q$ that minimizes the I-projection to $p(\mathbf{x})$, i.e. $D(q\|p) = \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})}$, is equivalent to

$$\max_{q\in Q} \sum_{\mathbf{c}\in C} E_q[\theta_c(\mathbf{x_c})] + H(q(\mathbf{x}))$$

  where $E_q[\theta_c(\mathbf{x_c})] = \sum_{\mathbf{x_c}} q(\mathbf{x_c})\theta_c(\mathbf{x_c})$ and $H(q(\mathbf{x}))$ is the *entropy* of $q(\mathbf{x})$

- If $p \in Q$, the value of the objective at optimality is **equal to** $\ln Z(\theta)$
- How should we approximate this? We need a compact way of representing $q(\mathbf{x})$ and finding the maxima

From last lecture: Relaxation approaches

We showed two approximation methods, both making use of the *local consistency constraints* M_L on the marginal polytope:

1. Bethe-free energy approximation (for pairwise MRFs):

$$\max_{\mu \in M_L} \quad \sum_{ij \in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j) + \sum_{i \in V} H(\mu_i) - \sum_{ij \in E} I(\mu_{ij})$$

   - Not concave. Can use concave-convex procedure to find local optima
   - Loopy BP, if it converges, finds a saddle point (often a local maxima)

2. Tree re-weighted approximation (for pairwise MRFs):

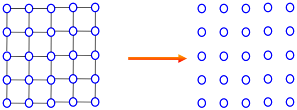
$$(*) \max_{\mu \in M_L} \quad \sum_{ij \in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j) + \sum_{i \in V} H(\mu_i) - \sum_{ij \in E} \rho_{ij} I(\mu_{ij})$$

   - $\{\rho_{ij}\}$ are edge appearance probabilities (must be consistent with some set of spanning trees)
   - This is concave! Find global maximiza using projected gradient ascent
   - Provides an upper bound on log-partition function, i.e. $\ln Z(\theta) \leq (*)$

Two types of variational algorithms: Mean-field and relaxation

$$\max_{q \in Q} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} q(\mathbf{x_c}) \theta_c(\mathbf{x_c}) + H(q(\mathbf{x})).$$

- Although this function is concave and thus in theory should be easy to optimize, we need some compact way of representing $q(\mathbf{x})$
- *Relaxation* algorithms work directly with *pseudomarginals* which may not be consistent with any joint distribution
- *Mean-field* algorithms assume a factored representation of the joint distribution, e.g.



$$q(\mathbf{x}) = \prod_{i \in V} q_i(x_i) \qquad \text{(called } \textit{naive} \text{ mean field)}$$

Naive mean-field

- Suppose that Q consists of all fully factored distributions, of the form $q(\mathbf{x}) = \prod_{i \in V} q_i(x_i)$
- We can use this to simplify

$$\max_{q \in Q} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} q(\mathbf{x_c}) \theta_c(\mathbf{x_c}) + H(q)$$

- First, note that $q(\mathbf{x}_c) = \prod_{i \in c} q_i(x_i)$
- Next, notice that the joint entropy decomposes as a sum of local entropies:

$$\begin{aligned}
H(q) &= -\sum_{\mathbf{x}} q(\mathbf{x}) \ln q(\mathbf{x}) \\
&= -\sum_{\mathbf{x}} q(\mathbf{x}) \ln \prod_{i \in V} q_i(x_i) = -\sum_{\mathbf{x}} q(\mathbf{x}) \sum_{i \in V} \ln q_i(x_i) \\
&= -\sum_{i \in V} \sum_{\mathbf{x}} q(\mathbf{x}) \ln q_i(x_i) \\
&= -\sum_{i \in V} \sum_{x_i} q_i(x_i) \ln q_i(x_i) \sum_{\mathbf{x}_{V \setminus i}} q(\mathbf{x}_{V \setminus i} \mid x_i) = \sum_{i \in V} H(q_i).
\end{aligned}$$

Naive mean-field

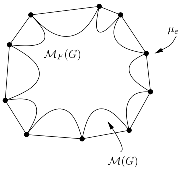
- Putting these together, we obtain the following variational objective:

$$(*) \max_q \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} \theta_c(\mathbf{x_c}) \prod_{i \in c} q_i(x_i) + \sum_{i \in V} H(q_i)$$

  subject to the constraints

$$q_i(x_i) \geq 0 \quad \forall i \in V, x_i \in \mathrm{Val}(X_i)$$
$$\sum_{x_i \in \mathrm{Val}(X_i)} q_i(x_i) = 1 \quad \forall i \in V$$

- Corresponds to optimizing over an *inner bound* on the marginal polytope, given by $\mu_{ij}(x_i, x_j) = \mu_i(x_i)\mu_j(x_j)$ and the above constraints:



- We obtain a *lower bound* on the partition function, i.e. $(*) \leq \ln Z(\theta)$

Naive mean-field for pairwise MRFs

- How do we maximize the variational objective?

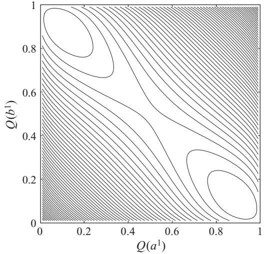
$$(*) \max_q \sum_{ij \in E} \sum_{x_i, x_j} \theta_{ij}(x_i, x_j) q_i(x_i) q_j(x_j) - \sum_{i \in V} \sum_{x_i} q_i(x_i) \ln q_i(x_i)$$

- This is a non-convex optimization problem, with many local maxima!
- Nonetheless, we can greedily maximize it using **block coordinate descent**:
  1. Iterate over each of the variables $i \in V$. For variable i,
  2. Fully maximize (*) with respect to $\{q_i(x_i), \forall x_i \in \mathrm{Val}(X_i)\}$.
  3. Repeat until convergence.
- Constructing the Lagrangian, taking the derivative, setting to zero, and solving yields the update: (*shown on blackboard*)

$$q(x_i) = \frac{1}{Z_i} \exp \left\{ \theta_i(x_i) + \sum_{j \in N(i)} q_j(x_j) \theta_{ij}(x_i, x_j) \right\}$$

How accurate will the approximation be?

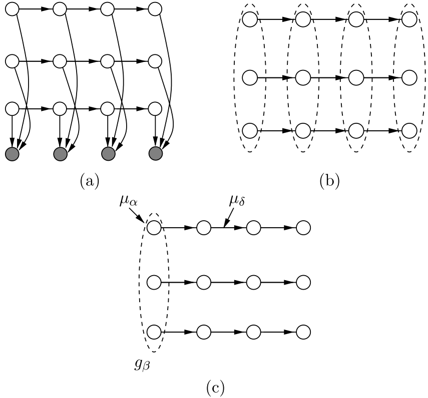
- Consider a distribution which is an XOR of two binary variables A and B: $p(a, b) = 0.5 - \epsilon$ if $a \neq b$ and $p(a, b) = \epsilon$ if $a = b$
- The contour plot of the variational objective is:



- Even for a single edge, mean field can give very wrong answers!
- Interestingly, once $\epsilon > 0.1$, mean field has a single maximum point at the uniform distribution (thus, exact)

Structured mean-field approximations

- Rather than assuming a fully-factored distribution for q, we can use a *structured* approximation, such as a spanning tree
- For example, for a factorial HMM, a good approximation may be a product of chain-structured models:



(a) (b) (c)

Obtaining true bounds on the marginals

- Suppose we can obtain *upper* and *lower* bounds on the partition function
- These can be used to obtain upper and lower bounds on marginals
- Let $Z(\theta_{x_i})$ denote the partition function of the distribution on $\mathbf{X}_{\mathbf{V}\setminus i}$ where $X_i = x_i$
- Suppose that $L_{x_i} \leq Z(\theta_{x_i}) \leq U_{x_i}$
- Then,

$$
\begin{aligned}
p(x_i;\theta) &= \frac{\sum_{\mathbf{x}_{\mathbf{V}\setminus i}} \exp(\theta(\mathbf{x}_{\mathbf{V}\setminus i}, x_i))}{\sum_{\hat{x}_i} \sum_{\mathbf{x}_{\mathbf{V}\setminus i}} \exp(\theta(\mathbf{x}_{\mathbf{V}\setminus i}, \hat{x}_i))} \\
&= \frac{Z(\theta_{x_i})}{\sum_{\hat{x}_i} Z(\theta_{\hat{x}_i})} \\
&\leq \frac{U_{x_i}}{\sum_{\hat{x}_i} L_{\hat{x}_i}}.
\end{aligned}
$$

- Similarly, $p(x_i;\theta) \geq \frac{L_{x_i}}{\sum_{\hat{x}_i} U_{\hat{x}_i}}$.

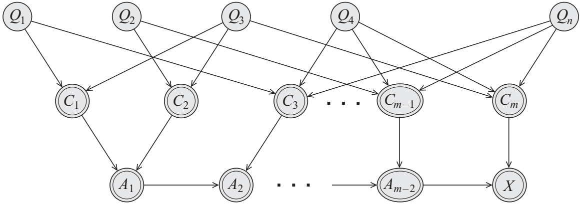
Software packages

1. libDAI
   - `http://www.libdai.org`
   - Mean-field, loopy sum-product BP, tree-reweighted BP, double-loop GBP
2. Infer.NET
   - `http://research.microsoft.com/en-us/um/cambridge/projects/infernet/`
   - Mean-field, loopy sum-product BP
   - Also handles continuous variables

Approximate marginal inference

- Nearly all approximate marginal inference algorithms are either:
  1. Variational algorithms (e.g., mean-field, TRW, loopy BP)
  2. **Monte-carlo methods (e.g., likelihood reweighting, MCMC)**
- **Unconditional sampling:** how can one estimate marginals in a BN if there is no evidence?
  - Topologically sort the variables, forward sample (using topological sort), and compute empirical marginals
  - Since these are indepedent samples, can use a Chernoff bound to quantify accuracy. *Small additive error with just a few samples!*
  - Doesn't contradict hardness results because **unconditional**
- **Conditional sampling:** what about computing $p(X \mid e) = p(X, e)/p(e)$?
  - Could try using forward sampling for both numerator and denominator, but in expectation would need at least $1/p(e)$ samples before $\hat{p}(e) \neq 0$
  - Thus, forward sampling won't work for conditional inference. We need new techniques.

- **Input:** 3-SAT formula with n literals $Q_1, \ldots Q_n$ and m clauses $C_1, \ldots, C_m$



- $p(X = 1) = \sum_{\mathbf{q},\mathbf{c},\mathbf{a}} p(\mathbf{Q} = \mathbf{q}, \mathbf{C} = \mathbf{c}, \mathbf{A} = \mathbf{a}, X = 1)$ is equal to the number of satisfying assignments times $\frac{1}{2^n}$
- Thus, $p(X = 1) > 0$ if and only if the formula has a satisfying assignment
- This shows that *exact marginal inference* is NP-hard

- Might there exist polynomial-time algorithms that can *approximately* answer marginal queries, i.e. for some ϵ, find ρ such that

$$\rho - \epsilon \leq p(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) \leq \rho + \epsilon \quad ?$$

- Suppose such an algorithm exists, for any $\epsilon \in (0, \frac{1}{2})$. Consider the following:
  1. Start with $\mathbf{E} = \{ X = 1 \}$
  2. For $i = 1, \ldots, n$:
  3. Let $q_i = \arg\max_q \; p(Q_i = q \mid \mathbf{E})$
  4. $\mathbf{E} \leftarrow \mathbf{E} \cup (Q_i = q_i)$
- At termination, $\mathbf{E}$ is a satisfying assignment (if one exists). Pf by induction:
  - In iteration i, if $\exists$ satisfying assignment extending $\mathbf{E}$ for **both** $q_i = 0$ *and* $q_i = 1$, then choice in line 3 does not matter
  - Otherwise, suppose $\exists$ satisfying assignment extending $\mathbf{E}$ for $q_i = 1$ but not for $q_i = 0$. Then, $p(Q_i = 1 \mid \mathbf{E}) = 1$ and $p(Q_i = 0 \mid \mathbf{E}) = 0$
  - Even if approximate inference returned $p(Q_i = 1 \mid \mathbf{E}) = 0.501$ and $p(Q_i = 0 \mid \mathbf{E}) = .499$, we would still choose $q_i = 1$
- Thus, it is even NP-hard to *approximately* perform marginal inference!