Inference and Representation

David Sontag

New York University

Lecture 9, Nov. 11, 2014

Variational methods

- Suppose that we have an arbitrary graphical model:

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

- All of the approaches begin as follows:

$$\begin{aligned}
D(q\|p) &= \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \\
&= -\sum_{\mathbf{x}} q(\mathbf{x}) \ln p(\mathbf{x}) - \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{1}{q(\mathbf{x})} \\
&= -\sum_{\mathbf{x}} q(\mathbf{x}) \Big(\sum_{\mathbf{c}\in C} \theta_c(\mathbf{x_c}) - \ln Z(\theta)\Big) - H(q(\mathbf{x})) \\
&= -\sum_{\mathbf{c}\in C} \sum_{\mathbf{x}} q(\mathbf{x})\theta_c(\mathbf{x_c}) + \sum_{\mathbf{x}} q(\mathbf{x}) \ln Z(\theta) - H(q(\mathbf{x})) \\
&= -\sum_{\mathbf{c}\in C} E_q[\theta_c(\mathbf{x_c})] + \ln Z(\theta) - H(q(\mathbf{x})).
\end{aligned}$$

The log-partition function

- Since $D(q\|p) \geq 0$, we have

$$-\sum_{\mathbf{c}\in C} E_q[\theta_c(\mathbf{x_c})] + \ln Z(\theta) - H(q(\mathbf{x})) \geq 0,$$

  which implies that

$$\ln Z(\theta) \geq \sum_{\mathbf{c}\in C} E_q[\theta_c(\mathbf{x_c})] + H(q(\mathbf{x})).$$

- Thus, *any* approximating distribution $q(\mathbf{x})$ gives a lower bound on the log-partition function (for a BN, this is the log probability of the observed variables)
- Recall that $D(q\|p) = 0$ if and only if $p = q$.Thus, if we allow ourselves to optimize over *all* distributions, we have:

$$\ln Z(\theta) = \max_q \sum_{\mathbf{c}\in C} E_q[\theta_c(\mathbf{x_c})] + H(q(\mathbf{x})).$$

Re-writing objective in terms of moments

$$
\begin{aligned}
\ln Z(\theta) &= \max_q \sum_{\mathbf{c}\in C} E_q[\theta_c(\mathbf{x_c})] + H(q(\mathbf{x})) \\
&= \max_q \sum_{\mathbf{c}\in C} \sum_{\mathbf{x}} q(\mathbf{x})\theta_c(\mathbf{x_c}) + H(q(\mathbf{x})) \\
&= \max_q \sum_{\mathbf{c}\in C} \sum_{\mathbf{x}_c} q(\mathbf{x_c})\theta_c(\mathbf{x_c}) + H(q(\mathbf{x})).
\end{aligned}
$$

- Assume that $p(\mathbf{x})$ is in the exponential family, and let $\mathbf{f}(\mathbf{x})$ be its sufficient statistic vector
- Define $\mu_q = E_q[\mathbf{f}(\mathbf{x})]$ to be the *marginals* of $q(\mathbf{x})$
- We can re-write the objective as

$$
\ln Z(\theta) = \max_{\mu\in M} \max_{q:E_q[\mathbf{f}(\mathbf{x})]=\mu} \sum_{\mathbf{c}\in C} \sum_{\mathbf{x_c}} \theta_c(\mathbf{x_c})\mu_c(\mathbf{x_c}) + H(q(\mathbf{x})),
$$

where M, the **marginal polytope**, consists of all valid marginal vectors

Re-writing objective in terms of moments

- Next, push the max over q instead to obtain:

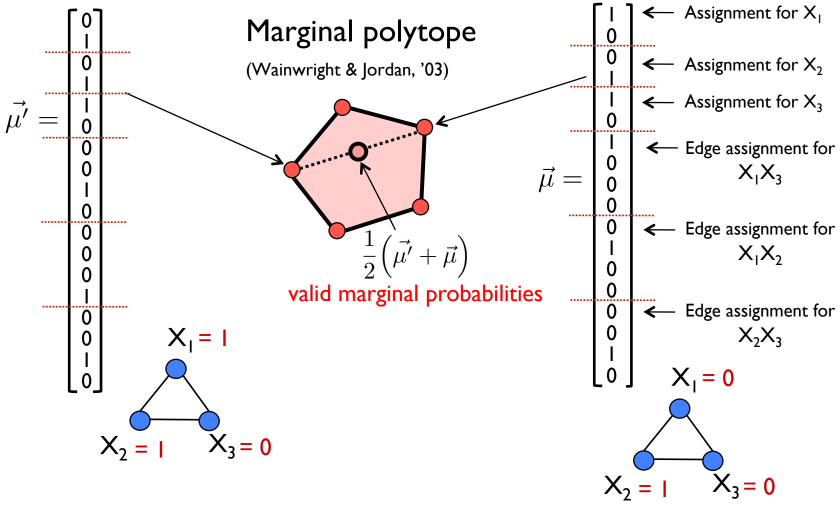
$$\begin{aligned}\ln Z(\theta) &= \max_{\mu} \sum_{\mathbf{c}\in C} \sum_{\mathbf{x_c}} \theta_c(\mathbf{x_c})\mu_c(\mathbf{x_c}) + H(\mu), \text{ where} \\ H(\mu) &= \max_{q:E_q[\mathbf{f}(\mathbf{x})]=\mu} H(q).\end{aligned}$$

- For discrete random variables, the **marginal polytope** M is given by

$$\begin{aligned} M &= \left\{\mu \in \mathbb{R}^d \mid \mu = \sum_{\mathbf{x}\in\mathcal{X}^m} p(\mathbf{x})\mathbf{f}(\mathbf{x}) \text{ for some } p(\mathbf{x}) \geq 0, \sum_{\mathbf{x}\in\mathcal{X}^m} p(\mathbf{x}) = 1\right\} \\ &= \texttt{conv}\left\{\mathbf{f}(\mathbf{x}), \mathbf{x} \in \mathcal{X}^m\right\} \quad (\texttt{conv} \text{ denotes the convex hull operation})\end{aligned}$$

- For a discrete-variable MRF, the sufficient statistic vector $\mathbf{f}(\mathbf{x})$ is simply the concatenation of indicator functions for each clique of variables that appear together in a potential function
- For example, if we have a pairwise MRF on binary variables with $m = |V|$ variables and $|E|$ edges, $d = 2m + 4|E|$

Marginal polytope for discrete MRFs



Relaxation

$$\ln Z(\theta) = \max_{\mu \in M} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} \theta_c(\mathbf{x_c}) \mu_c(\mathbf{x_c}) + H(\mu)$$

- We still haven't achieved anything, because:
  1. The marginal polytope M is complex to describe (in general, exponentially many vertices and facets)
  2. $H(\mu)$ is very difficult to compute or optimize over
- We now make two approximations:
  1. We replace M with a *relaxation* of the marginal polytope, e.g. the local consistency constraints M_L
  2. We replace $H(\mu)$ with a function $\tilde{H}(\mu)$ which approximates $H(\mu)$

Local consistency constraints

- Force every "cluster" of variables to choose a local assignment:

$$\begin{aligned}
\mu_i(x_i) &\geq 0 \quad \forall i \in V, x_i \\
\sum_{x_i} \mu_i(x_i) &= 1 \quad \forall i \in V \\
\mu_{ij}(x_i, x_j) &\geq 0 \quad \forall ij \in E, x_i, x_j \\
\sum_{x_i, x_j} \mu_{ij}(x_i, x_j) &= 1 \quad \forall ij \in E
\end{aligned}$$

- Enforce that these local assignments are globally consistent:

$$\begin{aligned}
\mu_i(x_i) &= \sum_{x_j} \mu_{ij}(x_i, x_j) \quad \forall ij \in E, x_i \\
\mu_j(x_j) &= \sum_{x_i} \mu_{ij}(x_i, x_j) \quad \forall ij \in E, x_j
\end{aligned}$$

- The *local consistency polytope*, M_L is defined by these constraints
- Look familiar? Same local consistency constraints as used in Lecture 6 for the linear programming relaxation of MAP inference!

Local consistency constraints are *exact* for trees

- The marginal polytope depends on the specific sufficient statistic vector $\mathbf{f}(\mathbf{x})$
- **Theorem:** The local consistency constraints *exactly* define the marginal polytope for a tree-structured MRF
- **Proof:** Consider any pseudo-marginal vector $\vec{\mu} \in M_L$. We will specify a distribution $p_T(\mathbf{x})$ for which $\mu_i(x_i)$ and $\mu_{ij}(x_i, x_j)$ are the pairwise and singleton marginals of the distribution p_T
- Let X_1 be the root of the tree, and direct edges away from root. Then,

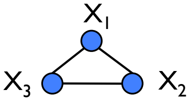
$$p_T(\mathbf{x}) = \mu_1(x_1) \prod_{i \in V \setminus X_1} \frac{\mu_{i,pa(i)}(x_i, x_{pa(i)})}{\mu_{pa(i)}(x_{pa(i)})}.$$

- Because of the local consistency constraints, each term in the product can be interpreted as a conditional probability.

Example for non-tree models

- For non-trees, the local consistency constraints are an *outer bound* on the marginal polytope
- Example of $\vec{\mu} \in M_L \backslash M$ for a MRF on binary variables:

$$\mu_{ij}(x_i, x_j) = \begin{array}{cc l} X_j = 0 & X_j = 1 & \\ \boxed{0} & \boxed{.5} & X_i = 0 \\ \boxed{.5} & \boxed{0} & X_i = 1 \end{array}$$



- To see that this is not in M, note that it violates the following triangle inequality (valid for marginals of MRFs on **binary variables**):

$$\sum_{x_1 \neq x_2} \mu_{1,2}(x_1, x_2) + \sum_{x_2 \neq x_3} \mu_{2,3}(x_2, x_3) + \sum_{x_1 \neq x_3} \mu_{1,3}(x_1, x_3) \leq 2.$$

Maximum entropy (MaxEnt)

- Recall that $H(\mu) = \max_{q:E_q[\mathbf{f}(\mathbf{x})]=\mu} H(q)$ is the entropy of the *maximum entropy distribution* with marginals μ
- This yields the optimization problem:

$$\max_q H(q(\mathbf{x})) = -\sum_{\mathbf{x}} q(\mathbf{x}) \log q(\mathbf{x})$$

$$\text{s.t.} \qquad \sum_{\mathbf{x}} q(\mathbf{x}) f_i(\mathbf{x}) = \alpha_i$$

$$\sum_{\mathbf{x}} q(\mathbf{x}) = 1 \quad \text{(strictly concave w.r.t. } q(\mathbf{x})\text{)}$$

- E.g., when doing inference in a pairwise MRF, the α_i will correspond to $\mu_l(x_l)$ and $\mu_{lk}(x_l, x_k)$ for all $(l,k) \in E, x_l, x_k$

What does the MaxEnt solution look like?

- To solve the MaxEnt problem, we form the Lagrangian:

$$L = -\sum_{\mathbf{x}} q(\mathbf{x}) \log q(\mathbf{x}) - \sum_i \lambda_i \left(\sum_{\mathbf{x}} q(\mathbf{x}) f_i(\mathbf{x}) - \alpha_i \right) - \lambda_{sum} \left(\sum_{\mathbf{x}} q(\mathbf{x}) - 1 \right)$$

- Then, taking the derivative of the Lagrangian,

$$\frac{\partial L}{\partial q(\mathbf{x})} = -1 - \log q(\mathbf{x}) - \sum_i \lambda_i f_i(\mathbf{x}) - \lambda_{sum}$$

- And setting to zero, we obtain:

$$q^*(\mathbf{x}) = \exp\left(-1 - \lambda_{sum} - \sum_i \lambda_i f_i(\mathbf{x}) \right) = e^{-1-\lambda_{sum}} e^{-\sum_i \lambda_i f_i(\mathbf{x})}$$

- From constraint $\sum_{\mathbf{x}} q(\mathbf{x}) = 1$ we obtain $e^{1+\lambda_{sum}} = \sum_{\mathbf{x}} e^{-\sum_i \lambda_i f_i(\mathbf{x})} = Z(\lambda)$
- We conclude that the maximum entropy distribution has the form (substituting $\vec{\theta}$ for $-\vec{\lambda}$)

$$q^*(\mathbf{x}) = \frac{1}{Z(\theta)} \exp(\theta \cdot \mathbf{f}(\mathbf{x}))$$

Entropy for tree-structured models

- Suppose that p is a tree-structured distribution, so that we are optimizing only over marginals $\mu_{ij}(x_i, x_j)$ for $ij \in T$
- We conclude from the previous slide that the $\arg\max_{q:E_q[\mathbf{f}(\mathbf{x})]=\mu} H(q)$ is a tree-structured MRF
- The entropy of q as a function of its marginals can be shown to be

$$H(\vec{\mu}) = \sum_{i \in V} H(\mu_i) - \sum_{ij \in T} I(\mu_{ij})$$

where

$$\begin{aligned} H(\mu_i) &= -\sum_{x_i} \mu_i(x_i) \log \mu_i(x_i) \\ I(\mu_{ij}) &= \sum_{x_i, x_j} \mu_{ij}(x_i, x_j) \log \frac{\mu_{ij}(x_i, x_j)}{\mu_i(x_i)\mu_j(x_j)} \end{aligned}$$

- Can we use this for non-tree structured models?

Bethe-free energy approximation

- The Bethe entropy approximation is (for any graph)

$$H_{bethe}(\vec{\mu}) = \sum_{i \in V} H(\mu_i) - \sum_{ij \in E} I(\mu_{ij})$$

- This gives the following variational approximation:

$$\max_{\mu \in M_L} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} \theta_c(\mathbf{x_c}) \mu_c(\mathbf{x_c}) + H_{bethe}(\vec{\mu})$$

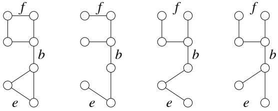
- For non tree-structured models this is not concave, and is hard to maximize
- Loopy belief propagation, if it converges, finds a saddle point!

Concave relaxation

- Let $\tilde{H}(\mu)$ be an *upper bound* on $H(\mu)$, i.e. $H(\mu) \leq \tilde{H}(\mu)$
- As a result, we obtain the following **upper bound** on the log-partition function:

$$\ln Z(\theta) \leq \max_{\mu \in M_L} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} \theta_c(\mathbf{x_c}) \mu_c(\mathbf{x_c}) + \tilde{H}(\mu)$$

- An example of a **concave** entropy upper bound is the **tree-reweighted** approximation (Jaakkola, Wainwright, & Wilsky, '05), given by specifying a distribution over spanning trees of the graph



Letting $\{\rho_{ij}\}$ denote edge appearance probabilities, we have:

$$H_{TRW}(\vec{\mu}) = \sum_{i \in V} H(\mu_i) - \sum_{ij \in E} \rho_{ij} I(\mu_{ij})$$

Comparison of LBP and TRW

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):

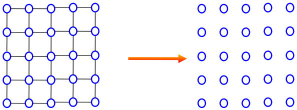
$$(*) \max_{\mu \in M_L} \quad \sum_{ij \in E} \sum_{x_i, x_j} \mu_{ij}(x_i, x_j) \theta_{ij}(x_i, x_j) + \sum_{i \in V} H(\mu_i) - \sum_{ij \in E} \rho_{ij} I(\mu_{ij})$$

   - $\{\rho_{ij}\}$ 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.



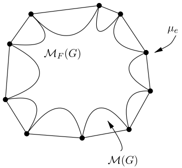
$$q(\mathbf{x}) = \prod_{i \in V} q_i(x_i) \qquad \text{(called } \textit{naive} \text{ mean field)}$$

Naive mean-field

- Using the same notation as in the rest of the lecture, naive mean-field is:

$$
\begin{aligned}
(*) \max_{\mu} \sum_{\mathbf{c} \in C} \sum_{\mathbf{x_c}} \theta_c(\mathbf{x_c}) \mu_c(\mathbf{x}_c) + \sum_{i \in V} H(\mu_i) \quad \text{subject to} \\
\mu_i(x_i) &\geq 0 \quad \forall i \in V, x_i \in \mathrm{Val}(X_i) \\
\sum_{x_i \in \mathrm{Val}(X_i)} \mu_i(x_i) &= 1 \quad \forall i \in V \\
\mu_c(\mathbf{x}_c) &= \prod_{i \in c} \mu_i(x_i)
\end{aligned}
$$

- Corresponds to optimizing over an *inner bound* on the marginal polytope:



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

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