CS 6347

Lecture 9

Variational Methods
(Mean Field)

I-Projections

- Consider a general MRF

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C)$$

- Consider approximate distributions $q \in Q$, and minimize the KL divergence, or equivalently

$$\max_{q \in Q} \left[H(q) + \sum_C \sum_{x_C} q_C(x_C) \log \psi_c(x_c) \right]$$

- This gives a lower bound on $\log Z$, with equality if $p \in Q$

Bethe Free Energy

$$\max_{\tau \in \mathbf{T}} H_B(\tau) + \sum_C \sum_{x_C} \tau_C(x_C) \log \psi_c(x_c)$$

where T is the local marginal polytope, and H_B is an approximate entropy function

$$H_B(\tau) = -\sum_{i \in \mathrm{V}} \sum_{x_i} \tau_i(x_i) \log \tau_i(x_i) - \sum_C \sum_{x_C} \tau_C(x_C) \log \frac{\tau_C(x_C)}{\prod_{i \in C} \tau_i(x_i)}$$

Bethe Free Energy

- Local optima of the Bethe free energy are in correspondence with fixed points of loopy belief propagation
  - At convergence, we can extract approximate marginals from loopy BP and plug them into the Bethe free energy to approximate Z
  - How do we know that these approximate marginals are in the local marginal polytope?

Bipartite Matchings

- Given a bipartite graph, a matching is a subset of the edges, $M \subseteq E$, such that no vertex is incident to more than one edge in M
  - A matching is perfect if every vertex is incident to exactly one edge in M
- Let's suppose we are given a positive vector of edge weights w and define the weight of a matching to be the product of the edge weights that it contains
- As an example of a general graphical model, let's construct a probability distribution over a given graph G such that probability of a matching is proportional to its weight

(done one the board)

Naive Mean Field

$$\max_{q \in Q} \left[H(q) + \sum_{C} \sum_{x_C} q_C(x_C) \log \psi_c(x_c) \right]$$

- Instead of using the Bethe approximation, we could also restrict the set Q to be sufficiently simple
  - Need it to have a compact representation and a tractable entropy function
- One simple, yet popular, choice is the so-called naive mean field approximation

$$q(x) = \prod_{i \in V} q_i(x_i)$$

Naive Mean Field

$$q(x) = \prod_{i \in V} q_i(x_i)$$

- Assumes that the joint distribution factorizes over a completely disconnected graph
- This distribution is compactly represented: we only need $|V|(size\ of\ the\ state\ space\ - 1)$ numbers
- The entropy of this distribution is easy to compute

$$H(q) = -\sum_{i \in V} \sum_{x_i} q_i(x_i) \log q_i(x_i)$$

Naive Mean Field

$$q(x) = \prod_{i \in V} q_i(x_i)$$

- Marginal distributions are also easy to compute

$$q_C(x_C) = \prod_{i \in C} q_i(x_i)$$

- Plugging into the lower bound

$$\log Z \geq -\sum_{i \in V} \sum_{x_i} q_i(x_i) \log q_i(x_i) + \sum_{C} \sum_{x_C} \left[\prod_{i \in C} q_i(x_i) \right] \log \psi_c(x_c)$$

Naive Mean Field

$$\max_q -\sum_{i\in V}\sum_{x_i} q_i(x_i)\log q_i(x_i) + \sum_C \sum_{x_C}\left[\prod_{i\in C} q_i(x_i)\right]\log\psi_c(x_c)$$

such that

$$\sum_{x_i} q_i(x_i) = 1, \qquad for\ all\ i \in V$$

$$q_i(x_i) \geq 0, \qquad for\ all\ i \in V, x_i$$

This is NOT a concave optimization problem!

Naive Mean Field

$$\max_q -\sum_{i\in V}\sum_{x_i} q_i(x_i)\log q_i(x_i) + \sum_C \sum_{x_C}\left[\prod_{i\in C} q_i(x_i)\right]\log \psi_c(x_c)$$

such that

$$\sum_{x_i} q_i(x_i) = 1, \qquad for\ all\ i \in V$$

$$q_i(x_i) \geq 0, \qquad for\ all\ i \in V, x_i$$

Can construct a Lagrangian and apply coordinate ascent

(worked out on the board)

UTD

Naive Mean Field

$$\max_q -\sum_{i\in V}\sum_{x_i} q_i(x_i)\log q_i(x_i) + \sum_C \sum_{x_C}\left[\prod_{i\in C} q_i(x_i)\right]\log \psi_c(x_c)$$

such that

$$\sum_{x_i} q_i(x_i) = 1, \qquad for\ all\ i \in V$$

$$q_i(x_i) \geq 0, \qquad for\ all\ i \in V, x_i$$

Equivalent to the Bethe free energy with additional constraint that $\tau_C(x_C) = \prod_{i\in C} \tau_i(x_i)$