Lecture 6 Variational Inference

Xuanxi Zhang

1 Inference in a Graphical Model

Suppose we have a graphical model with variables $x_1, \ldots, x_d$. The joint distribution is given by:

$$p(x_1, \ldots, x_d) = \frac{1}{Z} \prod_{C \in G} \psi_C(x_C)$$

where C represents the cliques of graph G.

The main task is to compute the marginals:

$$p(x_S) = \int p(x_1, \ldots, x_d) \prod_{j \notin S} dx_j$$

where S is a (small) subset of the variables $x_1, \ldots, x_d$. In general, such marginalization is intractable as soon as $|S| \ll N$ (i.e., #P-complete problem).
Thus, we need to exploit the structure in G.

2 Belief Propagation Algorithm

Factor Graph Representation A factor graph is a bipartite graph where we draw an edge $x_i - C_j$ if variable x_i appears in factor C_j.

2.1 Message Passing in Factor Graphs

The Belief Propagation (BP) algorithm will pass messages between the variables and factors in the graph iteratively.

Undirected Bipartite Factor Graph

We consider an *undirected bipartite factor graph*:

Introducing Directed Structure

We first introduce a **directed** structure in this graph. The directed messages are represented as:

$$\nu_{i \to j}, \quad \hat{\nu}_{j \to i}$$

where these are "messages" sent locally in the graph.

Message Passing in Iterations

At each iteration t, for each edge (i, j), we define messages as probability distributions over X_i:

$$\nu_{i \to j}^{(t)}, \quad \hat{\nu}_{j \to i}^{(t)} \in \mathcal{P}(X_i)$$

where $\nu_{i \to j}(x_i) \geq 0$ and

$$\int \nu_{i \to j}(x_i) dx_i = 1.$$

Convergence of Messages

Under some conditions, it turns out that these messages converge to a **fixed point** as $t \to \infty$.

Message Marginals and Update Rules

Marginal Interpretation

$$\nu_{i\to j}^{(\infty)} : \text{marginal of } X_i \text{ in a modified graphical model, where factor } j \text{ is missing.}$$

$$\hat{\nu}_{j\to i}^{(\infty)} : \text{marginal of } X_i \text{ in a modified graphical model, where variable } i \text{ only has factor } j.$$

Update Rules for Message Passing

Update rules consist of **local message-passing**:
Neighborhood notation:

$$N(i) \setminus j, \quad N(j) \setminus i$$

Variable to Factor Message Update

$$\nu_{i\to j}^{(t+1)}(x_i) \propto \prod_{j'\in N(i)\setminus j} \hat{\nu}_{j'\to i}^{(t)}(x_i)$$

Factor to Variable Message Update

$$\hat{\nu}_{j\to i}^{(t+1)}(x_i) \propto \int \psi_j(x_{C_j}) \left(\prod_{i'\in N(j)\setminus i} \nu_{i'\to j}^{(t)}(x_{i'}) \right) dx_{i'}$$

Computing Marginals by Message Aggregation

We compute the desired marginals by aggregating messages $\hat{\nu}_{j\to i}$ for all $j \in N(i)$:

$$\nu_i^{(t)}(x_i) \propto \prod_{j\in N(i)} \hat{\nu}_{j\to i}^{(t)}(x_i)$$

Correctness of the Algorithm

Question: When is this algorithm correct?
Intuition: When the original graph G can be *separated* along node i, using the modified graphs $\tilde{G}_{ij}$, for $j \in N(i)$, then we can **"divide and conquer"**.
Key Property: If the graphs $\tilde{G}_{ij_1}$ and $\tilde{G}_{ij_2}$ only share variable x_i, then the marginal $p(x_i)$ satisfies certain properties.

Marginal Computation

$$p(x_i) = \int p(x_1, \ldots, x_d) \prod_{i'\neq i} dx_{i'} = \frac{1}{Z} \int \prod_j \psi_j(x_{C_j}) \prod_{i'\neq i} dx_{i'}$$

$$= \frac{1}{Z} \int \left[\prod_{j\in \tilde{G}_{ij_1}} \psi_j(x_{C_j}) \right] \left[\prod_{j\in \tilde{G}_{ij_2}} \psi_j(x_{C_j}) \right] \psi_{j_1}(x_i, x_{N(j_1)\setminus i}) \psi_{j_2}(x_i, x_{N(j_2)\setminus i})$$

$$= \frac{1}{Z} \left[\int \prod_{j\in \tilde{G}_{ij_1}} \psi_j(x_{C_j}) \cdot \psi_{j_1}(x_i, x_{N(j_1)\setminus i}) \prod_{i'\neq i} dx_{i'} \right] \left[\int \tilde{G}_{ij_2} \right]$$

$$\propto \hat{\nu}_{j_1\to i}(x_i) \cdot \hat{\nu}_{j_2\to i}(x_i)$$

Tree Structure and Exactness of Belief Propagation

- Graphs G that satisfy this separability condition for every node cannot have any cycle. - Therefore, G is a **tree**.

Theorem: (BP is exact on trees) Consider a tree graphical model, with diameter t^* (the maximum distance between any pair of nodes). Then:

Convergence and Exactness of Belief Propagation

(i) Convergence of BP Updates

Belief Propagation (BP) updates converge after at most t^* iterations, for any initial condition. That is, for any edge (i, j) and any $t > t^*$, we have:

$$\nu_{i\to j}^{(t)} = \nu_{i\to j}^*, \quad \hat{\nu}_{j\to i}^{(t)} = \hat{\nu}_{j\to i}^*$$

(ii) Fixed Point Messages Provide Exact Marginals

The fixed point messages provide exact marginals:

$$\nu_i^*(x_i) = p(x_i) \quad \forall i.$$

Proof (Sketch)

Main idea: Induction over the depth of the tree.
For a given node i, let $j' \in N(i) \setminus j$ and $i' \in N(j') \setminus i$.
Induction Step: Assume the result is true for trees of depth $< t^*$.

Final Steps in Proof of BP Exactness on Trees

$$p(x_i) \simeq \psi_j(x_i) \int dx_1 \dots dx_{i-1}\, dx_{i+1} \dots dx_d \prod_j \psi_j(x_{C_j})$$

Using the **induction hypothesis**, this simplifies to:

$$\simeq \psi_j(x_i) \int \prod_{j'\in N(i)\setminus j} \psi_{j'}(x_{C_j}) \prod_{i'\in N(j')\setminus i} p(x_{i'})\, dx_{i'}$$

Applying the **tree structure property**:

$$\simeq \psi_j(x_i) \prod_{j'\in N(i)\setminus j} \int \psi_{j'}(x_{C_j}) \prod_{i'\in N(j')\setminus i} p(x_{i'})\, dx_{i'}$$

which leads to:

$$\simeq \nu_{j\to i}$$

Thus, we conclude:

$$\nu_{i\to j} \simeq \nu_i^*(x_i)$$

(More details in M&M, Theorem 14.1.)

Remarks

- We can use the same algorithm to compute marginals (and hence conditionals) over several variables. - Complexity of the BP algorithm on trees:

Complexity of BP on Trees

- BP is **linear** in the depth of the tree. - BP is **exponential** in the size of the factors, $N(j)$.

Belief Propagation on General Graphs and Free Energy

Natural Question: How about general graphs? Does BP work?

Observation

As an algorithm, nothing prevents us from running the iterative BP on a generic graph, even if it has loops!

What Happens in This Case?

- **Answer (Worst-case):** We can build counter-examples where BP does not converge to the true marginals.
(Pearl '88)
- **Answer (Average-case):** In practice, BP fails "gently"; the answer is nearly correct for graphs that "look like" trees.

Open Questions: - Does BP always stop? - How far is BP from the true marginals?

Bethe Free Entropy and Variational Principle

Let $p(x)$ be a general Gibbs distribution over a factor graph, and let $q(x)$ be another *positive* distribution.

Kullback-Leibler Divergence

Consider the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(q\|p) := \mathbb{E}_q\left[\log \frac{q(x)}{p(x)}\right]$$

(also known as relative entropy).
Fact: $D_{\mathrm{KL}} > 0$ and $D_{\mathrm{KL}} = 0$ if and only if $q = p$.

Jensen's Inequality and KL Divergence

Indeed, we have:

$$D_{\mathrm{KL}}(q\|p) = -\mathbb{E}_q \log \frac{p}{q} \geq -\log \mathbb{E}_q \left(\frac{p}{q}\right) = 0.$$

(by Jensen's inequality).

KL Divergence for a Gibbs Distribution

When $p(x)$ is a Gibbs distribution of the form:

$$p(x) = \frac{1}{Z} e^{-E(x)}$$

the KL divergence becomes:

$$D_{\mathrm{KL}}(q\|p) = \mathbb{E}_q[E(x)] + \mathbb{E}_q[\log q] + \log Z.$$

Recognizing terms:

$$D_{\mathrm{KL}}(q\|p) = U(q) - H(q) + \log Z \geq 0.$$

where $U(q)$ is the expected energy, and $H(q)$ is the entropy.

Variational Principle and Mean-Field Approximation

For a generic q, we have:

$$U(q) - H(q) \geq -\log Z$$

with equality if and only if $q = p$.

Variational Principle

- Consider $q \in \mathcal{F}$, a variational family, and optimize the left-hand side (LHS) over $\mathcal{F}$. - The family $\mathcal{F}$ is chosen such that the optimization is tractable.

Mean-Field Variational Model

Consider first a separable approximation:

$$q(x) = \prod_i q_i(x_i)$$

Free Entropy of q?

Energy Function

Recall that:

$$p(x) = \frac{1}{Z} \prod_j \psi_j(x_{C_j})$$

Thus, the energy function is:

$$E(x) = -\sum_j \log \psi_j(x_{C_j}).$$

Expectation of Energy Under q

Using the fact that q is a product measure, we obtain:

$$U(q) = -\sum_j \int \log \psi_j(x_{C_j}) \prod_{i \in C_j} q_i(x_i) dx_i.$$

Mean-Field Entropy and Approximation

The entropy term in the mean-field approximation is:

$$H(q) = -\sum_i \int q(x_i) \log q(x_i)\, dx_i.$$

Mean-Field Approximation Reformulation

The mean-field approximation minimizes:

$$\min_{q \text{ separable}} [U(q) - H(q)] \quad \Leftrightarrow \quad \min_{q_{\text{sep}}} D_{\text{KL}}(q \| p).$$

Remarks on Mean-Field Approximation

- This mean-field approximation does **not** agree with:

$$q(x) = \prod_i p(x_i)$$

(the product of the marginals of p).
This would be true if instead we considered:

$$\min_{q_{\text{sep}}} D_{\text{KL}}(p \| q).$$

Practical Relevance

- Even though the mean-field approximation (MF) is crude (it assumes *no model of interactions*), it often provides useful information. - Example: In spin glasses, MF helps estimate **average magnetization**.