Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2020

15 Approximate Inference: Variational Methods

In this installment of the notes, we introduce the variational methodology for inference, from which we will see that loopy BP can be viewed as one example. More importantly, we will see the broader possibilities of this methodology for performing approximate inference. Our development continues to focus on undirected graphical models $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

Variational inference focuses on the partition function of a distribution. As we showed in lecture, the problem of computing marginals is equivalent to computing partition functions; i.e., marginals can be obtained via partition functions. To see this directly, suppose we are given a model

$$p_{x_\mathcal{V}}(x_\mathcal{V}) = \frac{\tilde{p}(x_\mathcal{V})}{Z}$$

where $\tilde{p}(\cdot)$ represents the unnormalized distribution, e.g., a factorization over potentials, such as

$$\tilde{p}(x_\mathcal{V}) = \prod_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} \psi_\mathcal{C}(x_\mathcal{C}),$$

where $\mathrm{cl}(\mathcal{G})$ denotes the cliques in $\mathcal{G}$.

Then if we are interested in a marginal $p_{x_\mathcal{S}}(x_\mathcal{S})$ for some $\mathcal{S} \subset \mathcal{V}$, note that

$$p_{x_{\mathcal{V}\setminus\mathcal{S}}|x_\mathcal{S}}(x_{\mathcal{V}\setminus\mathcal{S}}|x_\mathcal{S}) = \frac{p_{x_\mathcal{V}}(x_\mathcal{V})}{p_{x_\mathcal{S}}(x_\mathcal{S})} = \frac{\tilde{p}(x_\mathcal{V})}{Z p_{x_\mathcal{S}}(x_\mathcal{S})} = \frac{\tilde{p}(x_\mathcal{V})}{Z_{\setminus\mathcal{S}}(x_\mathcal{S})}, \tag{1}$$

where we are using

$$Z_{\setminus\mathcal{S}}(x_\mathcal{S}) \triangleq Z p_{x_\mathcal{S}}(x_\mathcal{S})$$

to denote the partition function for the conditional distribution $p_{x_{\mathcal{V}\setminus\mathcal{S}}|x_\mathcal{S}}(x_{\mathcal{V}\setminus\mathcal{S}}|x_\mathcal{S})$. Note from (1) that $Z_{\setminus\mathcal{S}}(x_\mathcal{S})$ can be equivalently expressed in the form

$$Z_{\setminus\mathcal{S}}(x_\mathcal{S}) = \sum_{x_{\mathcal{V}\setminus\mathcal{S}}} \tilde{p}(x_\mathcal{V}),$$

which corresponds to summing out all the variables in the unnormalized distribution $\tilde{p}(x_\mathcal{V})$ except $x_\mathcal{S}$.

Hence, the desired marginal is the ratio of two different partition functions, viz.,

$$p_{x_\mathcal{S}}(x_\mathcal{S}) = \frac{Z_{\setminus\mathcal{S}}(x_\mathcal{S})}{Z}.$$

Evidently, by approximating or bounding partition functions, we can approximate or bound marginals.

15.1 Partition Function Computation as Optimization

In its direct form

$$Z = \sum_{x_\mathcal{V}} \tilde{p}(x_\mathcal{V}) \tag{2}$$

the partition function is difficult to approximate or bound in useful ways. The key to variational methods is rewriting (2) as an *optimization* instead of a *summation*. We introduce the notation

$$E(x_\mathcal{V}) \triangleq -\ln \tilde{p}(x_\mathcal{V}), \tag{3}$$

where we can take $E(x_\mathcal{V}) = \infty$ if $\tilde{p}(x_\mathcal{V}) = 0$. If $\tilde{p}_{\mathsf{x}_\mathcal{V}} = \prod_{i\in\mathcal{V}} \phi_i(x_i) \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j)$ is from a pairwise graphical model, then

$$-E(x_\mathcal{V}) = \sum_{i\in\mathcal{V}} \tilde{\phi}_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \tilde{\psi}_{ij}(x_i, x_j) \tag{4}$$

where $\tilde{\phi}_i(x_i) = \log \phi_i(x_i)$ and $\tilde{\psi}_{ij}(x_i, x_j) = \log \psi_{ij}(x_i, x_j)$.

We claim that the partition function can be written as

$$\ln Z = \sup_{q\in\mathcal{P}} \varphi(q), \quad \text{where} \quad \varphi(q) = H(q) - \mathbb{E}_q\left[E(x_\mathcal{V})\right], \tag{5}$$

and

$$H(q) = -\sum_{x_\mathcal{V}} q(x_\mathcal{V}) \ln q(x_\mathcal{V}) = -\mathbb{E}_q\left[\ln q(\mathsf{x}_\mathcal{V})\right] \tag{6}$$

denotes the entropy of the distribution q, and where $\mathcal{P}$ denotes the set of all possible distributions over variables $\mathsf{x}_\mathcal{V}$.[1]

To verify (5), we first write

$$\begin{aligned}
\ln Z &= -\ln p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}) - E(x_\mathcal{V}) \\
&= \mathbb{E}_q\left[-\ln p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}) - E(x_\mathcal{V})\right], \quad \text{any } q \in \mathcal{P} && (7) \\
&= \mathbb{E}_q\left[-\ln p_{\mathsf{x}_\mathcal{V}}(\mathsf{x}_\mathcal{V})\right] - \mathbb{E}_q\left[E(\mathsf{x}_\mathcal{V})\right] \\
&= \underbrace{\mathbb{E}_q\left[-\ln p_{\mathsf{x}_\mathcal{V}}(\mathsf{x}_\mathcal{V}) + \ln q(\mathsf{x}_\mathcal{V})\right]}_{=D(q\|p_{\mathsf{x}_\mathcal{V}})} \underbrace{- \mathbb{E}_q\left[E(\mathsf{x}_\mathcal{V})\right] - \mathbb{E}_q\left[\ln q(\mathsf{x}_\mathcal{V})\right]}_{=\varphi(q)}, && (8)
\end{aligned}$$

where to obtain (7) we have used that $\ln Z$ is a constant, so averaging with respect to a distribution has no effect, and in (8) we have used (5), (6) and the definition

$$D(q\|p_{\mathsf{x}_\mathcal{V}}) \triangleq \sum_{x_\mathcal{V}} q(x_\mathcal{V}) \ln \frac{q(x_\mathcal{V})}{p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V})}, \tag{9}$$

[1] This set is referred to as the *probability simplex* over the corresponding alphabet. When necessary for clarity we can use $\mathcal{P}^\mathcal{X}$ to denote the simplex corresponding to variables defined over an alphabet $\mathcal{X}$.

which is the Kullback-Leibler divergence (or information divergence) of $p_{x_\mathcal{V}}$ from q.

In turn, it follows from (8) that

$$\sup_{q\in\mathcal{P}} \varphi(q) = \sup_{q\in\mathcal{P}} \big[\ln Z - D(q\|p_{x_\mathcal{V}})\big] = \ln Z - \inf_{q\in\mathcal{P}} D(q\|p_{x_\mathcal{V}}) = \ln Z,$$

where we have used *Gibb's inequality*[2]

$$D(q\|p_{x_\mathcal{V}}) \geq 0, \quad \text{with equality if and only if } q = p_{x_\mathcal{V}}\,. \tag{10}$$

Note, too, that this further implies that

$$\arg\max_{q\in\mathcal{P}} \varphi(q) = p_{x_\mathcal{V}}\,. \tag{11}$$

Expressing the partition function as an optimization of the form (5) motivates a variety of useful approximation algorithms and bounds, as we illustrate in the remainder of these notes.

15.2 Lower Bounds

The optimization form of the (log) partition function (5) is convenient for obtaining useful lower bounds. In particular, for any subset $\mathcal{S} \subset \mathcal{P}$ it follows immediately that

$$\log Z = \sup_{q\in\mathcal{P}} \varphi(q) \geq \sup_{q\in\mathcal{S}\subset\mathcal{P}} \varphi(q). \tag{12}$$

One of the simplest and most widely used such lower bounds arises by choosing as our subset those distributions corresponding to graphs with no edges, i.e., distributions corresponding to independent random variables, which take the form

$$q(x_\mathcal{V}) = \prod_{i\in\mathcal{V}} q_i(x_i), \tag{13}$$

where the q_i have an interpretation as "approximate" node marginals of the target distribution $p_{x_\mathcal{V}}$.

Eq. (13) is referred to as the *mean-field factorization*, and we denote the subset in this case using the notation $\mathcal{P}_{\mathrm{MF}}$. Despite the simple structure in q, the resulting bound is surprisingly good in many applications, and thus a particularly popular variational approximation in practice.

[2]Eq. (10) is obtained via Jensen's inequality: for a strictly concave function f and a random variable u we have $\mathbb{E}[f(u)] \leq f(\mathbb{E}[u])$ with equality if and only if u is a constant. In particular with $f(\cdot) = \ln(\cdot)$ and $u = p_{x_\mathcal{V}}(x_\mathcal{V})/q(x_\mathcal{V})$ we obtain

$$-D(q\|p_{x_\mathcal{V}}) = \mathbb{E}_q\left[\ln \frac{p_{x_\mathcal{V}}(x_\mathcal{V})}{q(x_\mathcal{V})}\right] \leq \ln \mathbb{E}_q\left[\frac{p_{x_\mathcal{V}}(x_\mathcal{V})}{q(x_\mathcal{V})}\right] = \ln \sum_{x_\mathcal{V}} p(x_\mathcal{V}) = 0.$$

To solve[3]

$$\ln Z_{\mathrm{MF}} \triangleq \sup_{q \in \mathcal{P}_{\mathrm{MF}}} \varphi(q) \tag{14}$$

$$= \sup_{q_{\mathcal{V}}} \left\{ \sum_{i \in \mathcal{V}} \mathbb{E}_{q_i} \left[\tilde{\phi}_i(x_i) \right] + \sum_{(i,j) \in \mathcal{E}} \mathbb{E}_{q_i q_j} \left[\tilde{\psi}_{ij}(x_i, x_j) \right] + \sum_{i \in \mathcal{V}} H(q_i) \right\} \tag{15}$$

subject to

$$q_i(x_i) \geq 0 \qquad \text{for all } i \in \mathcal{V} \text{ and } x_i \in \mathcal{X} \tag{16}$$

$$\sum_{x_i \in \mathcal{X}} q_i(x_i) = 1 \qquad \text{for all } i \in \mathcal{V}, \tag{17}$$

we can use Lagrange multipliers.

We let λ_i be the Lagrange multiplier associated with (17). We omit introducing Lagrange multipliers for (16), since these constraints will be satisfied automatically by our solution.

In turn, the associated Lagrangian is

$$\begin{aligned}
\tilde{\varphi}(q, \lambda_{\mathcal{V}}) &= \varphi(q) + \sum_{i \in \mathcal{V}} \lambda_i \left[\sum_{x_i \in \mathcal{X}} q_i(x_i) - 1 \right] \\
&= \sum_{i \in \mathcal{V}} \mathbb{E}_{q_i} \left[\tilde{\phi}_i(x_i) \right] + \sum_{(i,j) \in \mathcal{E}} \mathbb{E}_{q_i q_j} \left[\tilde{\psi}_{ij}(x_i, x_j) \right] + \sum_{i \in \mathcal{V}} H(q_i) + \sum_{i \in \mathcal{V}} \lambda_i \left[\sum_{x_i \in \mathcal{X}} q_i(x_i) - 1 \right] \\
&= \sum_{i \in \mathcal{V}} \sum_{x_i \in \mathcal{X}} q_i(x_i)\, \tilde{\phi}_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \sum_{x_i, x_j \in \mathcal{X}} q_i(x_i)\, q_j(x_j)\, \tilde{\psi}_{ij}(x_i, x_j) \\
&\qquad - \sum_{i \in \mathcal{V}} \sum_{x_i \in \mathcal{X}} q_i(x_i) \ln q_i(x_i) + \sum_{i \in \mathcal{V}} \lambda_i \left[\sum_{x_i \in \mathcal{X}} q_i(x_i) - 1 \right],
\end{aligned}$$

whence

$$\frac{\partial \tilde{\varphi}(q, \lambda_{\mathcal{V}})}{\partial q_k(x_k)} = \tilde{\phi}_k(x_k) + \left[\sum_{j \in \mathcal{N}(k)} \sum_{x_j \in \mathcal{X}} q_j(x_j)\, \tilde{\psi}_{kj}(x_k, x_j) \right] - \left[\ln q_k(x_k) + 1 \right] + \lambda_k. \tag{18}$$

Setting (18) to zero and replacing the index k with i for convenience, we find that the optimizing q satisfies

$$q_i(x_i) \propto \exp \left\{ \tilde{\phi}_i(x_i) + \sum_{j \in \mathcal{N}(i)} \sum_{x_j \in \mathcal{X}} q_j(x_j)\, \tilde{\psi}(x_i, x_j) \right\}. \tag{19}$$

[3]Note that $H(q) = \sum_{i \in \mathcal{V}} H(q_i)$ due to the marginal independence structure in q.

Since (19) is an implicit equation for q, we typically resort to the following iterative procedure for solving it: for $t = 0, 1, \ldots,$

$$q_i^{t+1}(x_i) \quad \longleftarrow \quad \frac{\exp\left\{\tilde{\phi}_i(x_i) + \sum_{j \in \mathcal{N}(i)} \sum_{x_j \in \mathcal{X}} q_j^t(x_j)\, \tilde{\psi}(x_i, x_j)\right\}}{\sum_{x_i' \in \mathcal{X}} \exp\left\{\tilde{\phi}_i(x_i') + \sum_{j \in \mathcal{N}(i)} \sum_{x_j \in \mathcal{X}} q_j^t(x_j)\, \tilde{\psi}(x_i', x_j)\right\}} \tag{20}$$

starting from some initial guess $q_{\mathcal{V}}^0$. When this converges to the optimizing q^*, our lower bound is $\ln Z_{\mathrm{MF}} = \varphi(q^*)$. In general, analyzing convergence of the iteration (20) is involved, like for analysis of loopy BP. One can typically only expect to arrive at a local optimum (unless the objective function in (15) is convex, convergence to q^* is guaranteed). Even if one does converge to the global optimum q^*, the quality of the approximation of the marginals is unclear in most scenarios.

Tighter Lower Bounds

Tighter lower bounds can be obtained by choosing larger sets $\mathcal{S}$ in (12) than $\mathcal{P}_{\mathrm{MF}}$. One obvious choice corresponds to the set $\mathcal{P}_{\mathcal{T}}$ of distributions over a particular spanning tree $\mathcal{T}$ over the nodes $\mathcal{V}$. Another would be the set $\mathcal{P}_{\mathrm{tree}}$ of distributions over all N^{N-2} possible spanning trees[4] over the $|\mathcal{V}| = N$ nodes. Evidently,

$$\mathcal{P}_{\mathrm{MF}} \subset \mathcal{P}_{\mathcal{T}} \subset \mathcal{P}_{\mathrm{tree}} \subset \mathcal{P}$$

so the associated partition function bounds satisfy

$$\ln Z_{\mathrm{MF}} \leq \ln Z_{\mathcal{T}} \leq \ln Z_{\mathrm{tree}} \leq \ln Z$$

However, in practice, these refined lower bounds (and the many other possibilities) are not widely used.

15.3 Larger Cliques

To this point we have restricted our attention to distributions that factor into at most edgewise potentials, corresponding to (4). However, extending to distributions that factor over larger cliques is conceptually straightforward. Here we consider the general case in which (4) is replaced by

$$-E(x_{\mathcal{V}}) = \sum_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \tilde{\psi}_{\mathcal{C}}(x_{\mathcal{C}}), \tag{21}$$

[4]This result on the number of spanning trees is referred to as Cayley's Theorem.

where $\mathrm{cl}^*(\mathcal{G})$ are the maximal cliques of the graph $\mathcal{G}$.

In this case, our objective function in (5) becomes

$$\varphi(q) = \mathbb{E}_q \left[\sum_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \tilde{\psi}_{\mathcal{C}}(\mathsf{x}_{\mathcal{C}}) \right] + H(q). \tag{22}$$

While deriving upper bounds for this general case requires additional tools, we can obtain lower bounds following essentially the same approach developed earlier, albeit with a little more bookkeeping.

For example, the mean-field approximation is obtained as follows. As before, we introduce Lagrange multiplier λ_i for constraint $\sum_{x_i \in \mathcal{X}} q_i(x_i) = 1$, for $i \in \mathcal{V}$. Then the Lagrangian is

$$\begin{aligned}
\tilde{\varphi}(q, \lambda_{\mathcal{V}}) &= \mathbb{E}_q \left[\sum_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \tilde{\psi}_{\mathcal{C}}(\mathsf{x}_{\mathcal{C}}) \right] + \sum_{i \in \mathcal{V}} H(q_i) + \sum_{i \in \mathcal{V}} \lambda_i \left(\sum_{x_i \in \mathcal{X}} q_i(x_i) - 1 \right) \\
&= \sum_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \mathbb{E}_{q_{\mathcal{C}}} \left[\tilde{\psi}_{\mathcal{C}}(\mathsf{x}_{\mathcal{C}}) \right] + \sum_{i \in \mathcal{V}} H(q_i) + \sum_{i \in \mathcal{V}} \lambda_i \left(\sum_{x_i \in \mathcal{X}} q_i(x_i) - 1 \right),
\end{aligned}$$

where

$$q_{\mathcal{C}}(x_{\mathcal{C}}) = \prod_{i \in \mathcal{C}} q_i(x_i).$$

In turn, we obtain

$$\frac{\partial \tilde{\varphi}(q, \lambda_{\mathcal{V}})}{\partial q_i(x_i)} = \sum_{\substack{\{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G}): \\ i \in \mathcal{C}\}}} \sum_{x_{\mathcal{C} \setminus \{i\}}} \tilde{\psi}_{\mathcal{C}}(x_{\mathcal{C}}) \prod_{\substack{\{j \in \mathcal{C}: \\ j \neq i\}}} q_j(x_j) - \ln q_i(x_i) - 1 + \lambda_i,$$

which when we equate to zero yields the mean field equation

$$q_i(x_i) \propto \exp \left\{ \sum_{\substack{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G}): \\ i \in C}} \sum_{x_{\mathcal{C} \setminus \{i\}}} \tilde{\psi}_{\mathcal{C}}(x_{\mathcal{C}}) \prod_{\substack{\{j \in \mathcal{C}: \\ j \neq i\}}} q_j(x_j) \right\},$$

which can be solved iteratively as in the pairwise case.