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

24 Relationships Among Inference Tasks

In this lecture we describe several probabilistic inference tasks and prove that they are essentially equivalent. This is useful because each inference task leads to different algorithmic frameworks. Understanding the relationships between different inference tasks also gives us a deeper understanding of both the computational tasks themselves and of the probabilistic graphical models on which the tasks are defined.

We will also see that all of these inference tasks are in general computationally hard in the worst case, assuming P $\neq$ NP, the famous complexity theory conjecture (that is widely believed to be true).

24.1 Probabilistic Inference Tasks

In this lecture we will restrict attention to pairwise graphical models. This means that we have an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and each $x \in \mathcal{X}^{\mathcal{V}}$ is assigned probability

$$p(x) = \frac{1}{Z} \prod_{ij \in \mathcal{E}} \psi_{ij}(x_i, x_j)\,.$$

Recall that $\psi_{ij} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ are potential functions and Z is the partition function whose purpose it is to normalize the distribution. It is sometimes convenient to explicitly write additional node-wise potentials $\psi_i(x_i)$ for each $i \in \mathcal{V}$ (and the model is still a "pairwise" GM). We will assume that $\mathcal{X}$ is finite (and small).

The probabilistic inference tasks under consideration in this lecture are, given $(G, \{\psi_{ij}\})$, to:

1. Compute the partition function Z
2. Produce independent samples from p
3. Compute expectations $\mathbb{E}_p[f(X)]$ of bounded functions $f : \mathcal{X}^{\mathcal{V}} \to [0, 1]$
4. Compute (nodewise) marginals $p_i(x_i)$

We will actually be concerned with approximate versions of these tasks. Namely, for any given $\epsilon > 0$: 1) Compute a multiplicative approximation to Z, i.e., return a number $\widehat{Z}$ such that $Z(1+\epsilon)^{-1} \leq \widehat{Z} \leq Z(1+\epsilon)$; 2) Produce samples that are within ϵ total variation distance from the true distribution p; 3) Determine $\mathbb{E}_p[f(X)]$ to within additive error ϵ; 4) Compute marginals to within additive error ϵ.

24.2 Some remarks about conditioning and marginals on multiple variables

24.2.1 Conditioning

In real-world inference tasks one often wants to compute $p_{\mathsf{x}_\mathcal{A}|\mathsf{x}_\mathcal{B}}(x_\mathcal{A}|z_\mathcal{B})$, i.e., the posterior on variables $\mathsf{x}_\mathcal{A}$ given observations $\mathsf{x}_\mathcal{B} = z_\mathcal{B}$; what relationship does this have to the above tasks? Let $\mathcal{B}^c = \mathcal{V} \setminus \mathcal{B}$. Luckily, as we now show, it turns out that the conditional distribution of $\mathsf{x}_{\mathcal{B}^c}$ given fixed $\mathsf{x}_\mathcal{B} = z_\mathcal{B}$ is described by a *new* MRF, and then one can obtain the distribution of $\mathsf{x}_\mathcal{A}$ by marginalizing out variables in $\mathcal{B}^c \setminus \mathcal{A}$.

To start with, thinking of $z_\mathcal{B}$ as fixed (and $x_{\mathcal{B}^c}$ as the variables) means we can drop a factor $p(z_\mathcal{B})$ and write

$$p(x_{\mathcal{B}^c}|z_\mathcal{B}) = \frac{p(x_{\mathcal{B}^c}, z_\mathcal{B})}{p(z_\mathcal{B})} \propto p(x_{\mathcal{B}^c}, z_\mathcal{B})\,.$$

Now, the factorized form[1] for p yields

$$\begin{aligned} p(x_{\mathcal{B}^c}, z_\mathcal{B}) &\propto \prod_{\substack{\mathcal{C}\in\mathrm{cl}(\mathcal{G})\\ \mathcal{C}\cap\mathcal{B}=\varnothing}} \psi_\mathcal{C}(x_\mathcal{C}) \prod_{\substack{\mathcal{C}\in\mathrm{cl}(\mathcal{G})\\ \mathcal{C}\cap\mathcal{B}\neq\varnothing}} \psi_\mathcal{C}(x_{\mathcal{C}\setminus\mathcal{B}}, z_{\mathcal{C}\cap\mathcal{B}}) \\ &= \prod_{\substack{\mathcal{C}\in\mathrm{cl}(\mathcal{G})\\ \mathcal{C}\cap\mathcal{B}=\varnothing}} \psi_\mathcal{C}(x_\mathcal{C}) \prod_{\substack{\mathcal{C}\in\mathrm{cl}(\mathcal{G})\\ \mathcal{C}\cap\mathcal{B}\neq\varnothing}} \widetilde{\psi}_\mathcal{C}(x_{\mathcal{C}\setminus\mathcal{B}})\,, \end{aligned}$$

where the potentials $\widetilde{\psi}$ are obtained from the corresponding original ones by fixing the coordinates in $\mathcal{C} \cap \mathcal{B}$ as specified in $z_\mathcal{B}$. One now makes the observation that the first product contains cliques entirely in $\mathcal{G} \setminus \mathcal{B}$, and indeed so too each potential in the second product depends only on subsets $x_\mathcal{D}$ for $\mathcal{D} \in \mathrm{cl}(\mathcal{G} \setminus \mathcal{B})$. To see this, check that each clique $\mathcal{C} \in \mathrm{cl}(\mathcal{G})$ with $\mathcal{C} \cap \mathcal{B} \neq \varnothing$ corresponds precisely to the clique $\mathcal{C} \setminus \mathcal{B}$ which is entirely in $\mathcal{G} \setminus \mathcal{B}$. This all implies that we can write

$$p(x_{\mathcal{B}^c}, z_\mathcal{B}) \propto \prod_{\mathcal{C}\in\mathrm{cl}(\mathcal{G}\setminus\mathcal{B})} \phi_\mathcal{C}(x_\mathcal{C})$$

for some potentials ϕ (which really are just combinations of cliques from before), which means that $p(x_{\mathcal{B}^c}|z_\mathcal{B})$ is described by an MRF on $x_{\mathcal{B}^c}$.

24.2.2 Marginals on more than one variable

Suppose that you wish to compute the marginal on the first two variables—can this be related to computing a marginal on a single variable? How much additional computation is needed?

[1] We will always assume our MRF distributions factorize over cliques, which (as shown by Hammersley-Clifford) is true for positive distributions. Models satisfying the Global Markov Property that do not factorize are somewhat anomalous—even the independent set model, which is not strictly positive, factorizes into a product of edge-potentials.

We describe how to compute $p(x_1, x_2)$ for a fixed pair $(x_1, x_2) \in \mathcal{X}^2$, using the identity $p(x_1, x_2) = p(x_1)p(x_2|x_1)$. One first computes $p(x_1)$; then, $p(x_2|x_1)$ is obtained as the marginal $\widetilde{p}(x_2)$ in a new model $\widetilde{p}$ obtained as described above.

The procedure can be repeated iteratively to yield any size marginal, and the number of individual marginalization steps is equal to the size of the marginal. Of course, this yields only the probability of a particular configuration, and a marginal of size m requires the procedure to be repeated $O(|\mathcal{X}|^m)$ times in order to get probabilities of all configurations. In certain cases there are slightly more efficient procedures than what is described here.

24.3 The Independent Set Model

For simplicity we will focus on a particular pairwise graphical model known as the Independent Set Model (in statistical physics the model is known as the hard core lattice gas model). Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, an independent set is a subset S of the nodes, so that no edge has both of its endpoints in S. Instead of dealing with a subset it is notationally easier to work with the corresponding indicator vector, so the variables in the model take binary values $x_i \in \mathcal{X} = \{0, 1\}$. Thus, each binary vector $x \in \{0, 1\}^N$ corresponds to the subset of the nodes with $x_i = 1$. The graphical model associated to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is described by the probability distribution

$$p(x) = \frac{1}{Z} \prod_{ij \in \mathcal{E}} \psi(x_i, x_j), \qquad \text{where} \quad \psi(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j = 1 \\ 1 & \text{otherwise.} \end{cases}$$

The potential functions enforce that no two endpoints of an edge can be in the set, i.e. any x not corresponding to an independent set has zero probability. Since each independent set has the same weight, the graphical model describes the uniform distribution over independent sets. The partition function is equal to the number of independent sets, i.e.,

$$Z = |\{\text{independent sets in } \mathcal{G}\}| \,.$$

In the statistical physics setting there is typically an additional node potential λ^{x_i} for each $i \in \mathcal{V}$. Denoting the set of independent sets of $\mathcal{G}$ by $\mathcal{I}$, the distribution of interest is

$$p(x) = \frac{1}{Z} \lambda^{\sum_i x_i} \,, \quad \text{if } x \text{ in } \mathcal{I} \,.$$

The parameter λ is known as the *activity* or *fugacity*. Higher values of λ result in more probability mass being placed on large independent sets.

Self-reducibility The independent set model satisfies an important property known as self-reducibility, which amounts to closure of the set of models when conditioning on some of the variables. In other words, fixing a given variable, say x_1, results in a new graphical model *from the same class*. To see this, observe that if we set $x_1 = 0$, then there are no constraints on any of the other variables and the resulting distribution on $X_{\mathcal{V}\setminus\{1\}}$ is just the same independent set model on the graph obtained by removing node 1. Similarly, if we set $x_1 = 1$, then there is zero probability that any of the neighbors of x_1 are equal to 1, and hence we can remove both node 1 and the neighbors of node 1.

Writing $p_{\mathcal{G}}$ for the independent set model on $\mathcal{G}$ to make the dependence on the graph explicit, what we have argued is that

$$p_{\mathcal{G}}(x_{\mathcal{V}\setminus\{1\}}|x_1 = 0) = p_{\mathcal{G}\setminus\{1\}}(x_{\mathcal{V}\setminus\{1\}}) \tag{1a}$$

and

$$p_{\mathcal{G}}(x_{\mathcal{V}\setminus\{1\}}|x_1 = 1) = p_{\mathcal{G}\setminus(\{1\}\cup\mathcal{N}(1))}(x_{\mathcal{V}\setminus(\{1\}\cup\mathcal{N}(1))}) \cdot \mathbb{1}_{x_{\mathcal{N}(1)}=\mathbf{0}} \,. \tag{1b}$$

It is a very good exercise to carefully justify these equalities.

24.4 Fully Polynomial Approximation Schemes

We need a way to think about algorithmic tractability when computing something approximately, and we first define a class of "efficient" algorithms whose runtime has reasonable dependence on the desired error as well as size of the problem. We give the definitions in terms of the partition function Z in order to be concrete, but these definitions extend to computation of a real-valued function in the natural way.

Fully Polynomial Time Approximate Scheme (fptas). Given a sequence of undirected graphical models $\mathcal{G}_n$ specified by $(\mathcal{V}_n, \mathcal{E}_n, (\psi_C)_{C\in \mathrm{cl}(\mathcal{G}_n)})$ and desired accuracy ϵ, an fptas computes in time bounded by a polynomial in $1/\epsilon$ and n an estimate $\widehat{Z}(\mathcal{G}_n, \epsilon)$ satisfying

$$Z(\mathcal{G}_n)(1-\epsilon) \le \widehat{Z}(\mathcal{G}_n, \epsilon) \le Z(\mathcal{G}_n)(1+\epsilon)\,.$$

One typically takes $\mathcal{V}_n = \{1, \dots, n\}$ so that n measures the "input size" but it can occasionally make sense to measure the input size in another way.

We will also be interested in randomized algorithms, which must be allowed to occasionally produce an answer quite far from the truth.

Fully Polynomial Randomized Approximate Scheme (fpras). Given a sequence of undirected graphical models $\mathcal{G}_n$ specified by $(\mathcal{V}_n, \mathcal{E}_n, (\psi_C)_{C\in \mathrm{cl}(\mathcal{G}_n)})$ and desired accuracy ϵ, an fpras computes an estimate $\widehat{Z}(\mathcal{G}_n, \epsilon)$ in time bounded by a polynomial in $1/\epsilon$ and n such that

$$\mathbb{P}\left(Z(\mathcal{G}_n)(1-\epsilon) \le \widehat{Z}(\mathcal{G}_n, \epsilon) \le Z(\mathcal{G}_n)(1+\epsilon)\right) \ge \frac{3}{4}.$$

24.5 Reductions Between Inference Tasks

The following result uses the notion of self-reducibility, which we will not define in the most general possible way. For our purposes, as discussed above a family of models (e.g. the independent set model) being self-reducible means that fixing the value of a node results in a model *from the same family* on a smaller graph.

Exercise 1. Let's add a nodewise potential (called a nodewise external field) to yield models of the form

$$p(x) = \frac{1}{Z} \exp\Big(\sum_{ij} \theta_{ij} x_i x_j + \sum_{i \in \mathcal{V}} \theta_i x_i \Big) .$$

Show that the marginal on $x_{\mathcal{V}\setminus\{1\}}$ conditioned on $x_1 = +1$ remains within the family described in the last displayed equation. (Hint: simply write down $p(x_{\mathcal{V}\setminus\{i\}} | x_i = +1)$.)

Exercise 2. Consider the family of Ising models without external fields, consisting of distributions of the form

$$p(x) = \frac{1}{Z} \exp\Big(\sum_{ij} \theta_{ij} x_i x_j \Big) ,$$

for graph $\mathcal{G}$ and parameters (θ_{ij}). Show that the marginal on $x_{\mathcal{V}\setminus\{1\}}$ conditioned on $x_1 = +1$ is *not* from the family described in the last displayed equation.

Theorem 1. *If a family of graphical models is self-reducible, then (the approximate versions of) 1) computing the partition function Z, 2)sampling, 3) computing expectations, and 4) nodewise marginals are all computationally equivalent, i.e., an fptas (or fpras) for any one of these can be used as a subroutine a polynomial number of times in order to solve any of the others thus yielding an fptas (or fpras).*

The arguments we give show computational equivalence for exact deterministic versions of these tasks, except for any algorithm using sampling as a subroutine. The arguments can be modified in a straightforward way to show full equivalence of approximate versions of all of the tasks.

We will write, for example, 1 ⇒ 2 to mean that an efficient algorithm for inference task 1 can be used to produce an efficient algorithm for inference task 2. The specifics—exact or approximate, deterministic or randomized—will be clear from the arguments.

Remark 1. It is worth emphasizing that we want to show computational equivalence for inference tasks on a specific family of models. For instance, that the ability to compute marginals in the independent set model allows to compute the partition function of independent set models (with roughly the same computational cost). This is clearly stronger than showing, for instance, that the ability to compute marginals in any Markov random field allows to compute partition functions in independent set models; importantly, this sort of lossy reduction would not enable us to “close the loop” in the reductions. This is where self-reducibility comes into play: we will need to perform computations on models obtained by conditioning on the value of a node, and in general conditioning on the value of a node may result in a graphical model from a different family.

Proof of Theorem 1:

$2 \Rightarrow 3$. This follows from Chebyshev’s inequality, which shows concentration for a sum of i.i.d. random variables. Note that due to randomness in the samples, here we can only guarantee an fpras (randomized approximation algorithm) for task 3. Suppose that we can produce i.i.d. samples $X^{(1)}, \dots, X^{(M)}$, $X^{(m)} \sim p$. Then we estimate $\mathbb{E}_p[f(X)]$ by the empirical average of the samples, $\frac{1}{M}\sum_{m=1}^{M} f(X^{(m)})$. It is left to the reader to solve for the appropriate M in terms of desired accuracy and error probability.

$3 \Rightarrow 4$. By computing the expectation of the specific function $f(x) = \mathbb{1}_{x_i=a}$ for each $a \in \mathcal{X}$, we obtain the marginal $p_i(x_i = a) = \mathbb{E}_p[f(X)]$.

$4 \Rightarrow 2$. We assume that you can compute a marginal in any independent set model in polynomial time, and use this to produce a sample from p in polynomial time. We specialize to the independent set model for ease of exposition.

Begin by sampling $X_1 \sim p_1(\cdot)$, i.e., just sample X_1 according to its marginal. Now,

$$p_{\mathcal{G}}(x_{\mathcal{V}}) = p_1(x_1)p(x_{\mathcal{V}\setminus\{1\}}|x_1)$$

so we will try to sample the remaining nodes conditional on the outcome of X_1. But by (1) the latter factor is just the independent set distribution in a different graphical model (with fewer nodes). If $x_1 = 0$, the new model has node 1 removed, and if $x_1 = 1$ then all neighbors $j \in \mathcal{N}(1)$ have value $x_j = 0$ and are then also removed. So we can now sample a new variable conditional on the outcome of X_1, and by carrying out this procedure iteratively we determine values for all of the variables.

We have now shown that tasks 2, 3, and 4 are computationally equivalent, and it remains to relate partition function computation to these.

$1 \Rightarrow 4$. We write $Z(\mathcal{G}, x_i = 1)$ (or $Z(\mathcal{G}, x_i = 0)$) to denote the number of independent sets in $\mathcal{G}$ with $x_i = 1$ (or $x_i = 0$, respectively). Now,

$$p_i(x_i = 1) = \frac{Z(\mathcal{G}, x_i = 1)}{Z} = \frac{Z(\mathcal{G}, x_i = 1)}{Z(\mathcal{G}, x_i = 1) + Z(\mathcal{G}, x_i = 0)}\,.$$

Each of the partition functions in turn simplifies to

$$Z(\mathcal{G}, x_i = 0) = Z(\mathcal{G} \setminus \{i\}) \quad \text{and} \quad Z(\mathcal{G}, x_i = 1) = Z(\mathcal{G} \setminus (\{i\} \cup \mathcal{N}(i)))\,,$$

and crucially these are just partition functions in independent set models on smaller graphs. Thus, the ability to compute partition functions allows to compute marginals.

$4 \Rightarrow 1$. We write the telescoping product

$$\begin{aligned}\frac{1}{Z(\mathcal{G})} &= \frac{Z(\mathcal{G} \setminus \{1\})}{Z(\mathcal{G})} \cdot \frac{Z(\mathcal{G} \setminus \{1,2\})}{Z(\mathcal{G} \setminus \{1\})} \cdots \frac{Z(\varnothing)}{Z(\mathcal{G} \setminus \{1, \dots, N-1\})} \\ &= p(x_1 = 0) p_{\mathcal{G}\setminus\{1\}}(x_2 = 0) \cdots p_{\mathcal{G}\setminus\{1,\dots,N-1\}}(x_N = 0)\,.\end{aligned}$$

Each of these factors is a marginal in an appropriate model, so we have converted the problem of computing the partition function to that of computing marginals. □

24.6 Hardness of Inference

The equivalences described above are useful not only for reasoning about algorithms, but also for reasoning about computational hardness. We will show hardness for sampling, and deduce hardness for all the other tasks.

We consider the Ising model on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with all edge weights equal to $-\beta$. This is known as the *antiferromagnetic* Ising model, and its probability mass function can be written as

$$p(x) \propto \exp\Big(-\beta \sum_{ij\in\mathcal{E}} x_i x_j\Big) \propto \exp\Big(-\beta \cdot |\mathcal{E}| + 2\beta \cdot \mathrm{cut}_{\mathcal{G}}(x)\Big) \propto \exp\Big(2\beta \cdot \mathrm{cut}_{\mathcal{G}}(x)\Big)\,. \quad (2)$$

Here

$$\mathrm{cut}_{\mathcal{G}}(x) = |\{ij \in \mathcal{E} : x_i \neq x_j\}|$$

is the number of edges with different values at the endpoints.

If we take β sufficiently large, then most of the probability mass is on the maximum cut, and finding the (size of the) maximum cut of a graph is known to be NP-hard.

Claim 1. *Consider the antiferromagnetic Ising model* (2)*. Taking $\beta \geq n^2$ yields that for $\mathsf{x} \sim p$,*

$$\mathbb{P}(\mathrm{cut}_{\mathcal{G}}(\mathsf{x}) = \mathrm{maxcut}(\mathcal{G})) \geq 1/2\,.$$

It follows that the ability to produce samples from an antiferromagnetic Ising model allows to compute (with high probability) the maximum cut of a graph.

One can avoid taking β so large by instead constructing a new graph as follows: replicate each node k times (where k is sufficiently large), and each edge $ij \in \mathcal{E}$ now becomes a complete bipartite graph connecting the k nodes arising from i to the k nodes arising from j. This amplifies the mass assigned to configurations corresponding to maximum cuts, while keeping β fixed. It is a useful exercise to do this calculation on one's own.

We conclude that, unfortunately, inference is hard in general. The hardness for antiferromagnetic Ising models is in contrast to the situation for ferromagnetic Ising models: a famous paper of Jerrum and Sinclair gives an fpras for the partition function of such models. Similarly, we know that these inference tasks are tractable in tree models via belief propagation, and in models with bounded treewidth via the junction tree algorithm. The classification of useful model subclasses for which inference is tractable is an ongoing and rich area of research.