10-708: Probabilistic Graphical Models 10-708, Spring 2018

14 : Approximate Inference Monte Carlo Methods

Lecturer: Kayhan Batmanghelich *Scribes: Biswajit Paria, Prerna Chikersal*

1 Introduction

We have already covered exact inference techniques like the elimination, message passing, and junction tree algorithms. This lecture covers approximate inference algorithms.

2 Monte Carlo Sampling

Expected probability of x w.r.t. P is computed as follows:

$$\int f(x)p(x)dx \approx \hat{f} \equiv \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), x^{(s)} \sim P(x) \tag{1}$$

$$\implies E_{P(x^{(s)})}[\hat{f}] = \frac{1}{S}\sum_{s=1}^{S} E_{P(x)}[f(x)] = E_{P(x)}[f(x)] \tag{2}$$

The above estimator you get is unbiased. This is due to the law of large numbers which says that if you compute the empirical average, it will go towards the actual mean.

For the estimator, you just need to sample x from $P(x)$ and substitute in f. However, this estimator also has a variance. Variance of empirical $\hat{f}$ is as follows:

$$var_{P(x^{(s)})}[\hat{f}] = \frac{1}{S^2}\sum_{s=1}^{S} var_{P(x)}[f(x)] = var_{P(x)}[f(x)]/\mathbf{S} \tag{3}$$

From the above we can see that variance of the estimator decays with$\sqrt{S}$. So, if you want an estimator that doesn't vary too much, you need to draw many samples and the behavior of that would be $\frac{1}{\sqrt{S}}$.

π can be approximated using the Monte Carlo method. Here's a video showing how: `https://www.youtube.com/watch?v=VJTFfIqO4TU`.

Monte Carlo method (at least the vanilla version) usually has a high variance. So, it is a very bad method and you should only use it if there's no alternative.

3 Sampling from a given distribution

3.1 Rejection Sampling

How to convert samples from a Uniform[0,1] generator? That is, say you are given a 1D pdf $P(y)$, how do you draw samples from that? If you are given normalized pdf $P(y)$ of a function, you can compute cdf $h(y)$ which a uniform non-decreasing function. If you can inverse $h(y)$ i.e. get $h^{-1}(y)$ then you can get your sample y, as $y = h^{-1}(y)$. But this won't work if $P(y)$ is not normalized. So what else can we do? We can use a method called rejection sampling.

1. Come up with a probability distribution $Q(x)$ that we can easily draw samples from.
2. Find a constant k such that $\frac{\tilde{P}(x)}{kQ(x)} < 1$.
3. Draw a sample from Q such that $P(y = 1|x) = \frac{\tilde{P}(x)}{kQ(x)} < 1$.
4. Accept a sample with $P(y = 1|x)$.

To prove that the accepted samples are coming from normalized $\tilde{P}(x)$, lets calculate the distribution of sample x.

$$\begin{aligned} P(x|y=1) &= \frac{P(y=1|x)Q(x)}{P(y=1)} \\ &= \frac{\frac{\tilde{P}(x)}{k}}{\int P(y=1|x)Q(x)dx} \\ &= \frac{\frac{\tilde{P}(x)}{k}}{\int \frac{\tilde{P}(x)}{k}dx} \\ &= \frac{\frac{\tilde{P}(x)}{k}}{\frac{1}{k}\int \tilde{P}(x)dx} = \frac{\tilde{P}(x)}{Z} \end{aligned}$$

where $Z = \int \tilde{P}(x)dx$.

Pitfalls of Rejection Sampling

Above we saw that rejection sampling works nicely for 1 dimensional probability distributions. But how does it scale for multidimensional probability distributions?

Lets consider an example. D is the number of dimensions.

$$P(x) = \mathcal{N}(0, I), Q(x) = \mathcal{N}(0, \sigma^2 I) \tag{4}$$

where I = identity, and the densities are fully factorisable such that:

$$p(x) = \prod_{i=1}^{D} p(x_i) \tag{5}$$

$$q(x) = \prod_{i=1}^{D} q(x_i) \tag{6}$$

The acceptance rate is:

$$q(y=1|x) = \prod_{i=1}^{D} \frac{p^*(x_i}{k_i q(x_i)} = \prod_{i=1}^{D} q(y=1|x_i) = O(\gamma^D) \tag{7}$$

So, as dimensions D increase, acceptance rate exponentially decreases.

3.2 Importance Sampling

Rejection sampling is wasteful in the sense that that a lot of samples can be rejected in the case of a bad proposal distribution. A different idea is to retain all the samples and *re-weight* them while computing their mean. The idea can be summarized using the following equations.

$$\int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx \tag{8}$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{p(x^{(s)})}{q(x^{(s)})} \tag{9}$$

where $q(x)$ is a proposal distribution that is positive whenever $p(x)$ is positive, and $\{x^{(s)}\}_{s=1}^{S} \sim q(x)$. We choose a distribution $q(x)$ that is easy to sample from and weight the sth sample by $\frac{p(x^{(s)})}{q(x^{(s)})}$ while taking the mean.

This formulation is also useful when we can compute the probability $P(x)$ but don't have the means to sample from it. However we must be able to sample from our proposal distribution $q(x)$ and also compute the weights $\frac{p(x^{(s)})}{q(x^{(s)})}$.

We now consider a more general scenario where we can compute $p(x)$ and $q(x)$ only upto normalization. We first compute the normalization factor of $p(x)$ as

$$\mathcal{Z}_p = \int \widetilde{p}(x)dx = \int \frac{\widetilde{p}(x)}{q(x)}q(x)dx \tag{10}$$

Then, the expectation can be written as

$$\begin{aligned}
\int f(x)p(x)dx &= \int f(x)\frac{p(x)}{q(x)}q(x)dx && (11)\\
&= \frac{1}{\mathcal{Z}_p}\int f(x)\frac{\widetilde{p}(x)}{q(x)}q(x)dx && (12)\\
&= \frac{\int f(x)\frac{\widetilde{p}(x)}{q(x)}q(x)dx}{\int \frac{\widetilde{p}(x)}{q(x)}q(x)dx} && (13)\\
&= \frac{\int f(x)\frac{\widetilde{p}(x)}{\widetilde{q}(x)}q(x)dx}{\int \frac{\widetilde{p}(x)}{\widetilde{q}(x)}q(x)dx} && (14)\\
&\text{(Replacing } q(x) \text{ by } \widetilde{q}(x)/\mathcal{Z}_q\text{)} && (15)
\end{aligned}$$

It can be approximated as

$$\int f(x)p(x)dx \approx \frac{\sum_l f(x^l)\frac{\widetilde{p}(x^l)}{\widetilde{q}(x^l)}}{\sum_l \frac{\widetilde{p}(x^l)}{\widetilde{q}(x^l)}} = \sum_l f(x_l)w_l \qquad (16)$$

where $x^l \sim q(x)$. While this estimator is consistent i.e. converges to the true expectation in the limit, it is a biased estimator.

Pitfalls of Importance Sampling

The current form of importance sampling doesn't scale well with the number of dimensions. Importance sampling works best when the proposal distribution is $q = p$, or when the weights are uniformly distributed with $w_l = 1/l$. However if we have full access to the original distribution, we wouldn't do importance sampling in the first place.

Unless p is close q, there will be a small number of dominant weights leading to a high variance. This is particularly evident in high-dimensions. To see this, consider a toy example of unnormalized weights $u_i = p(x^i)/q(x^i)$. The variability of two components of u can be computed as

$$\left\langle (u_i - u_j)^2 \right\rangle = \left\langle u_i^2 \right\rangle + \left\langle u_j^2 \right\rangle - 2\left\langle u_i \right\rangle \left\langle u_j \right\rangle \qquad (17)$$

where the expectation is wrt q. The mean of the individual weights $\langle u_i \rangle = \langle u_j \rangle = 1$, and $\left\langle u_i^2 \right\rangle = \left\langle u_j^2 \right\rangle = \left\langle \frac{p(x)}{q(x)} \right\rangle_p$. If $p(x)$ and $q(x)$ both factorize as $p(x) = \prod_{d=1}^D p(x_d)$, $q(x) = \prod_{d=1}^D q(x_d)$, we get $\langle u_i \rangle = \left\langle \frac{p(x_d)}{q(x_d)} \right\rangle_p^D$ yielding,

$$\left\langle (u_i - u_j)^2 \right\rangle = 2\left(\left\langle \frac{p(x_d)}{q(x_d)} \right\rangle_p^D - 1 \right) \qquad (18)$$

which is exponential in D. It can be shown that $\left\langle \frac{p(x_d)}{q(x_d)} \right\rangle_p > 1$, showing that a few components of u dominate in high dimensions.

This problem can be alleviated by resampling the weights (see Barber, Ch 27).

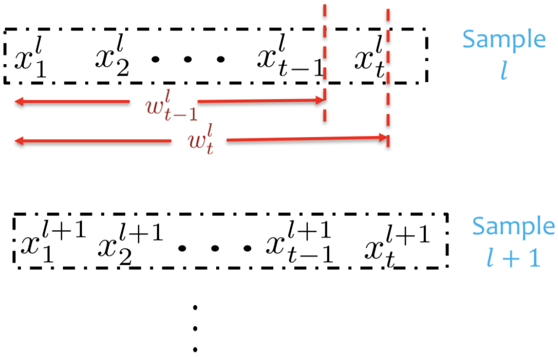

Figure 1: Importance sampling for high dimensional distributions with chain structured graphical models.

4 Using the structure for high-dimensional distributions

Another way of importance sampling is to use the structure of the graphical model. Suppose we want to sample from a high dimensional distribution $p(x_1, \ldots, x_t)$ corresponding to a chain graphical model. We can use the structure of the graphical model and sample each variable x_i in the natural order as depicted in figure 1.

The idea is to use a proposal distribution that is dependent on the variable being sampled: $q(x_t|x_{1:t-1})$. We can write the unnormalized weights in a recursive manner as

$$\widetilde{w}_t^l = \frac{p^*(x_{1:t}^l)}{q(x_{1:t})} = \frac{p^*(x_t^l|x_{1:t-1}^l)}{q(x_t^l|x_{1:t-1}^l)} \frac{p^*(x_{1:t-1}^l)}{q(x_{1:t-1})} \tag{19}$$

$$= \widetilde{w}_{t-1}^l \alpha_t^l \tag{20}$$

where p^* is the unnormalized distribution, and

$$\alpha_t^l = \frac{p^*(x_t^l|x_{1:t-1}^l)}{q(x_t^l|x_{1:t-1}^l)}. \tag{21}$$

Now that the problem is reduced to finding a proposal distribution for an one-dimensional problem, we can use rejection sampling or naive-importance sampling which work fine in low-dimensions.

4.1 Particle Filters

Consider the chain graphical but now with observations as shown in figure 2. Suppose we wish to sample $h_{1:t}|v_{1:t}$. The conditional distribution of $h_{1:t}$ is also known as the *filtered distribution*.

Now we also need to account for the probability of the observations. We modify α_t^l as,

$$\alpha_t^l = \frac{p(v_t|h_t)p(h_t^l|h_{t-1}^l)}{q(h_t^l|h_{1:t-1}^l)} \tag{22}$$

$$= p(v_t|h_t) \quad \text{(Choosing } q \text{ as the same distribution as } p\text{)} \tag{23}$$

resulting in

$$\widetilde{w}_t^l = \widetilde{w}_{t-1}^l p(v_t|h_t). \tag{24}$$

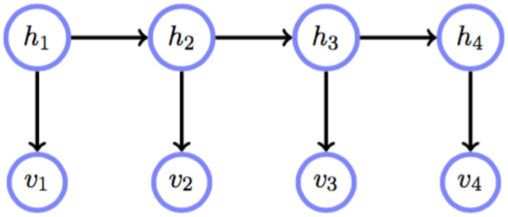

Figure 2: Chain graphical model with observations.

We denote the filtered distribution by ρ. Then

$$\begin{aligned}\rho(h_t) &\propto p(h_t|v_{1:t}) && (25)\\ &\propto p(v_t|h_t)\int_{h_{t-1}} p(h_t|h_{t-1})\rho(h_{t-1}) && (26)\end{aligned}$$

$\rho(h_{t-1})$ can be approximated as

$$\rho(h_{t-1}) \approx \sum_l w_{t-1}^l \delta(h_{t-1}, h'_{t-1}). \tag{27}$$

Substituting 27 in 26, we have

$$\rho(h_t) \approx \frac{1}{Z} p(v_t|h_t) \sum_l p(h_t|h_{t-1}^l) w_{t-1}^l \tag{28}$$

where Z is a normalization factor. Although $\rho(h_{t-1})$ is a spiky approximation $\rho(h_t)$ gets smoothened out due to the emission and transition factors.