6.867 Machine Learning — Fall 2019
Recitation Note: Sampling Methods

Huitao Shen
huitao@mit.edu

Nov 15, 2019

1 Introduction

Recall that in generative models,

- Model distributions: $p(z)$, $p(x|z)$.
- Target distributions: $p(x)$ or $\partial p(x)/\partial\theta$ for training; $p(z|x)$ for inference. Both distributions are typically intractable.

How to deal with intractable $p(x)$ or $p(z|x)$?

- Approach 1: Variational Inference

  Use $q(z|x;\phi) \approx p(z|x)$ for approximate inference, which can be trained as a byproduct as ELBO maximization, because

  $$\ln p(x) - \underbrace{\mathbb{E}_{z\sim q(z|x)}\left[\ln \frac{p(x,z)}{q(z|x)}\right]}_{\text{ELBO}} = \mathrm{KL}\left(q(z|x)||p(z|x)\right). \tag{1}$$

  We have talked about this approach in last week's recitation. The method is biased because in most scenarios $q(z|x) \neq p(z|x)$.

- Approach 2: Sampling

  The idea of sampling is very straightforward—to approximate the intractable summation (or integration if z is continuous) with finite samples:

  $$p(x) = \sum_z p(x|z)p(z) = \mathbb{E}_{z\sim p(z)}\left[p(x|z)\right] \approx \underbrace{\frac{1}{N}\sum_{i=1}^{N} p(z^{(i)}|x)}_{\hat{p}(x)}, \tag{2}$$

  where $z^{(1)}, \ldots, z^{(N)} \sim p(z)$ independently. $p(z|x)$ can be sampled similarly. The method is unbiased and its correctness is guaranteed by central limit theorem. However, the approximation, i.e. estimator $\hat{p}(x)$, potentially has a large variance when the sample size is small. In the following, we will focus on different sampling techniques that reduce the variance.

Thanks to the strong expressive power of neural networks, the bias in variational inference can be really small, while the variance in sampling methods is still annoyingly big. Therefore, to date the variational inference is more popular than sampling in dealing with intractable distributions. Nevertheless, we should still learn these sampling methods because it is possible one could find an efficient sampling method in the future with the helpful of neural networks (this is actually an active research area). Perhaps more importantly, these algorithms are extremely powerful and useful on their own.

2 Sampling Techniques

2.1 Importance Sampling (Exercise 10, Problem 2)

Instead of sampling $p(z)$, one can sample from a proposal distribution $q(z|x)$:

$$p(x) = \sum_z p(x|z)p(z) = \sum_z \frac{p(z)}{q(z|x)} q(z|x)q(z|x) = \mathbb{E}_{z\sim q(z|x)} \left[\frac{p(z)}{q(z|x)} p(x|z) \right], \tag{3}$$

By construction, $q(z|x) \approx p(z)p(x|z)$ and it is not hard to see the variance of the estimator is zero if $q(z|x) = p(z)p(x|z)$ exactly.

The problem with importance sampling is that if $q(z|x)$ undershoots $p(z)p(x|z)$, the variance will be very large and potentially infinity. When the dimension of z is high, this will almost always happen unless we have an exceptionally good proposal distribution, which is only possible for few well-known target distributions. Therefore, importance sampling is usually very unstable at high dimensions.

Finally, note that importance sampling is not restricted to sampling margin distribution $p(x) = \mathbb{E}_{z\sim p(z)}[p(x|z)]$, but can be used to sample a generic expectation.

2.2 Markov Chain Monte Carlo (MCMC)

A Markov chain is a series of random variables $\{z^{(t)}, t \in \mathbb{N}\}$, such that $p(z^{(t+1)}|z^{(t)}, \dots, z^{(0)}) = p(z^{(t+1)}|z^{(t)})$. In other words, the transition probability from the current state $z^{(t)}$ to the next state $z^{(t+1)}$ only depends on the current state, but not the state history. This Markov property is thus also called "memoryless" property.

The idea of MCMC sampling is to construct a Markov chain, such that the stationary distribution of the Markov chain is the same as the target distribution.

2.2.1 Metropolis-Hastings

In Metropolis-Hastings algorithm, there is a proposal distribution $q(z'|z)$, and an acceptance rate

$$p(z \to z') = \min\left(1, \frac{q(z|z')}{q(z'|z)} \frac{p(z')}{p(z)}\right). \tag{4}$$

Starting from the current state $z^{(t)}$, one first samples a proposed state z' from $q(z'|z^{(t)})$, then accepts it to be $z^{(t+1)}$ with probability $p(z^{(t)} \to z')$. If z' is not accepted, $z^{(t+1)} = z^{(t)}$.

It can be proved that with the above acceptance rate (and other technical requirements such as ergodicity), stationary distribution of the Markov chain is the same as the target distribution $\pi(z) = p(z)$.

The advantage of Metropolis-Hastings algorithm is that it is very generic. In fact, it is regarded as the single most important algorithm in the 20th century. In order to use it, one only needs to know unnormalized probability density, because only the ratio $p(z')/p(z)$ appears in (4). This is extremely useful in the context of probabilistic models because the posteriors or marginals are intractable exactly due to the normalization factor. According to Bayes theorem, for fixed x, the posterior is proportional to the joint likelihood $p(z|x) = p(x|z)p(z)/p(x) \propto p(x|z)p(z)$, which can be easily sampled with Metropolis-Hastings algorithm.

The disadvantage of Metropolis-Hastings is that the convergence of the Markov chain can be very slow—a price we always need to pay for a generic algorithm. Even worse, it is also hard to assess convergence. In many scenarios, we don't know in prior about the complicated, high-dimensional distribution $p(z)$, so the best choice of the proposal distribution is simply the uniform distribution. In these cases, (4) simplifies to $p(z \to z') = \min(1, p(z')/p(z))$. When the density ratio is small, the acceptance rate is also small. Now consider a multimodal distribution. It is difficult to make a transition from one mode to another because the density ratio from a mode to the region that connects different modes is very small. It is thus very easy for the Markov chain to get stuck in one of the modes, and we would never know whether there is a single mode or more than one mode in the target distribution.

2.2.2 Gibbs Sampling

Gibbs sampling is a special case of Metropolis-Hastings algorithm when the dimensional of z is high. It samples each component of z in turn while fixing remaining components. For example, consider $z \in \mathbb{R}^3$. An iteration of Gibbs sampling includes the following three samplings:

- $z_1^{(t+1)} \sim p(z_1|z_2^{(t)}, z_3^{(t)})$;
- $z_2^{(t+1)} \sim p(z_2|z_1^{(t+1)}, z_3^{(t)})$;
- $z_3^{(t+1)} \sim p(z_3|z_1^{(t+1)}, z_2^{(t+1)})$.

To proof its correctness, first denote $z_{\neg k}$ as all components of z except for k-th component z_k. Gibbs sampling corresponds to Metropolis-Hastings with proposal distribution $q(z|z') \equiv p(z_k|z'_{\neg k})$. It follows

$$\frac{q(z|z')}{q(z'|z)}\frac{p(z')}{p(z)} = \frac{p(z_k|z'_{\neg k})}{p(z'_k|z_{\neg k})}\frac{p(z'_k|z'_{\neg k})p(z'_{\neg k})}{p(z_k|z_{\neg k})p(z_{\neg k})} = 1, \tag{5}$$

where we have used the important fact that $z'_{\neg k} = z_{\neg k}$. Therefore, the acceptance rate is always one, justifying Gibbs sampling.

The advantage of Gibbs sampling is that it potentially converges faster than a generic Metropolis-Hastings algorithm, at the cost that one needs to know the conditional probability $p(z_k|z_{\neg k})$ exactly (NOT only up to a normalization factor). One should also be cautious about ergodicity breaking. A single example of a bimodal distribution with $z \in \mathbb{R}^2$ is constructed in Exercise 10, Problem 4.