CSE6243: Advanced Machine Learning **Fall 2023**

Lecture 6: Acceptance-Rejection & Importance Sampling

Lecturer: Bo Dai *Scribes: Kartik Narang, Shiva Ramaswami*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

6.1 Recap

6.1.1 Project Deadlines

1. Team Formation : 9/18
2. Project Proposal : 10/04
3. Midterm Report : 11/06
4. Presentation : 11/29 - 12/04
5. Final Report: 12/11

6.1.2 Last Class

- $\mathbb{E}_{p(x)}[f(x)]$: it can be difficult and time consuming to calculate this function, but it holds important application.
- **Monte-Carlo Approximation**
  Instead of directly calculating the expectation, we use this approximation below, where the left side is the expected value of the function with respect to probability function $p(x)$. We approximate it by taking N random samples (x_i) from the distribution $p(x)$, evaluating the function for each of these samples, and taking the average of these function values, i.e., $\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{N}\sum_{i=1}^{n} f(x_i)$, where $\{x_i\}_{i=1}^{n} \sim p(x)$.
- **Inverse Transformation**
  We learned that the application of this is limited because we need to calculate the CDF, and finding the inverse of the CDF or CDF itself can be a difficult problem.
- **Accept-Reject Sampling**

6.2 New Content

6.2.1 Acceptance-Rejection Sampling

Target: We want to sample $x \sim p(x)$.

Algorithm:

- Find M such that $Mq(x) \geq p(x)$ for every x.
- Generate random $y \sim q(y)$.
- Generate random $u \sim \mathcal{U}[0,1]$. If $u \leq \frac{p(y)}{Mq(y)}$ accept, otherwise re-select sample from step 2.

Proof: In this proof, we will show that in Acceptance-Rejection Sampling, the samples that we "accept" will always be a part of our true distribution. That is,

$$P(Y = i|accepted) = P(X = i)$$

We will use Bayes rule to show this.

$$P(Y = i|accepted) = \frac{P(Y = i, accepted)}{P(accepted)} \Rightarrow P(X = i)$$

For the numerator part,

$$P(Y = i, accepted) = q_i \frac{P_i}{Mq_i} = \frac{P_i}{M}$$

where q_i is our sample $P(Y = i)$, and $\frac{P_i}{Mq_i}$ is our sample acceptance criteria.

For the denominator part,

$$P(accepted) = \sum_{i=1}^{i} P(Y = i, accepted) = \sum_{i=1}^{k} \frac{P_i}{M} = \frac{1}{M}$$

Therefore,

$$P(Y = i|accepted) = \frac{P(Y = i, accepted)}{P(accepted)} = \frac{\frac{P_i}{M}}{\frac{1}{M}} = P(X = i)$$

■

Dilemma. We need M to be above $p(x)$ at all points – this can force M to be very large. If M is large, you waste a lot of time sampling, because a majority of your points may be rejected during the sampling process. To use this sampling method, M should not be large – this is difficult to satisfy.

Let's apply this algorithm to a problem involving posterior Bayesian inference and see what issues arise.

Example. We have the prior: $\pi(\theta)$ and the likelihood: $p(y|x,\theta) \propto \exp\left(-\frac{||y-\theta^T x||^2}{2\sigma^2}\right)$. This is the Bayes version of Least Squares Regression.
For $\mathcal{D} = \{X_i, Y_i\}_{i=1}^n$, our eventual goal we want to calculate is:

$$p(\theta|\mathcal{D}) = \prod_{i=1}^{n} \frac{p(Y_i|X_i,\theta)\pi(\theta)}{Z(\mathcal{D})}$$

where

$$Z(\mathcal{D}) = \int \prod_{i=1}^{n} p(Y_i|X_i,\theta)\pi(\theta)d\theta$$

Our proposal is the prior:

$$q(\theta) = \pi(\theta)$$

Can we calculate optimal M?

$$M \geq \frac{p(\theta|\mathcal{D})}{q(\theta)} = \frac{\prod_{i=1}^{n} p(Y_i|X_i,\theta)\pi(\theta)}{Z(\mathcal{D})\pi(\theta)} = \frac{\prod_{i=1}^{n} p(Y_i|X_i,\theta)}{Z(\mathcal{D})}$$

Z contains an integral which is usually intractable. We can modify the proposal slightly to get rid of this.

$$q(\theta) = \frac{\pi(\theta)}{Z(\mathcal{D})}$$

Now,

$$M \geq \frac{p(\theta|\mathcal{D})}{q(\theta)} = \frac{\prod_{i=1}^{n} p(Y_i|X_i,\theta)\pi(\theta)}{Z(\mathcal{D})\frac{\pi(\theta)}{Z(\mathcal{D})}} = \prod_{i=1}^{n} p(Y_i|X_i,\theta)$$

Therefore,

$$M = \max_{\theta} \prod_{i=1}^{n} p(Y_i|X_i,\theta)$$

Note that this doesn't completely solve the problem. This probability value can be greater than 1, and the optimal M value calculated here can be very large.

6.2.2 Importance Sampling

This sampling is even simpler than Acceptance-Rejection sampling. The motivation of it is that it is difficult to sample from $p(x)$, so we try to use $q(x)$.

$$\mathbb{E}_{p(x)}[f(x)] \tag{6.1}$$

$$=\mathbb{E}_{q(x)}[\frac{p(x)}{q(x)}f(x)] \quad \text{(We directly sample from some distribution)} \tag{6.2}$$

$$\approx\frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)}f(x_i) \quad \text{(Apply Monte Carlo approximation here)} \tag{6.3}$$

where $x \sim q(x)$.

Now let's look at the variance of importance sampling. Using the linearity of variance, we can pull in the variance term in the summation.

$$Var[\frac{1}{n}\sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)}f(x_i)] \tag{6.4}$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}Var[\frac{p(x)}{q(x)}f(x)] \tag{6.5}$$

$$=Var[\frac{p(x)}{q(x)}f(x)] \tag{6.6}$$

$$=\mathbb{E}[\frac{p^2(x)}{q^2(x)}f(x)] - \mathbb{E}^2[\frac{p(x)}{q(x)}f(x)] \tag{6.7}$$

We can now use the fact that this is a continuous random variable to integrate

$$(6.7) = \int f(x)\frac{p^2(x)}{q(x)}\,dx - \mathbb{E}_p^2[f(x)] \tag{6.8}$$

Finally, let's investigate how to find the best q. We want to find q to minimize the variance term, i.e.,

$$\min_q Var[g(x)] \tag{6.9}$$

where $g(x) = var[x] = E[x^2] - E^2[x]$.

We then use Cauchy-Schwarz inequality to finish this derivation.

$$(6.9) = \min \int f(x) \frac{p^2(x)}{q(x)}\, dx \tag{6.10}$$

$$\geq \int f(x) \frac{p(x)}{\sqrt{q(x)}} \sqrt{q(x)}\, dx \tag{6.11}$$

We then rewrite $\sqrt{q(x)} = \dfrac{f(x)p(x)}{\sqrt{q(x)}}$ and plug it in to get $\int f(x)p(x)\, dx$

$$q_{OPT}(x) = \frac{f(x)p(x)}{\int f(x)p(x)\, dx}$$

We note that we almost never use it because it is harder to find $q(x)$ than to do the original problem, and sometimes we can't calculate $p(x)$, so it's not even possible to run this.

In next class, we will cover MCMC (Markov Chain Monte Carlo), where the fundamental difference between that and Importance sampling is that every sample is dependent on the previous sample, whereas previously we had each sample being independent.