CSE 788.04: Topics in Machine Learning **Lecture Date: April 16, 2012**

Monte Carlo Sampling Methods

Lecturer: Brian Kulis *Scribe: Arash Ashari*

1 Monte Carlo Sampling

- Sample from posterior distribution; $p(x)$
- Computing expectation; $\int f(x)p(x)dx$

Monte Carlo is used for both problems.

1.1 Sampling from $p(x)$

Sample $x_1, x_2, ..., x_n$ iid from $p(x)$ and then take the average: $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ For large n we have that the average converges almost surely to $\int f(x)p(x)dx$. Convergence depends on the dimensionality, which is one problem we will have to deal with. Another problem is that sampling from complex distributions is not as easy as uniform.

1.1.1 Change of variable

$$p(y) = p(z)\left|\frac{dz}{dy}\right| \tag{1}$$

In some cases, a simple change of variable can be used to convert a uniform sample to the desired distribution (example: the Cauchy distribution). We will not discuss this in detail.

1.1.2 Rejection sampling

Consider an envelope distribution around the underlying $p(x)$. In the simplest case it can be a uniform (rectangle) around it. Then choose (x, y) such that:

- $x \sim u(A, B)$
- $y \sim u(0, Max(f(x)))$

If y is under the $p(x)$ curve accept x, otherwise reject it. Figure 1 shows an example.

We can generalize the envelope distribution. For example, we may have a Gaussian instead of a uniform. If the Gaussian (or any other envelope distribution), $q(x)$, couldn't envelope the underlying distribution, $p(x)$, we can simply scale it with a scale k, to have $kq(x)$ that envelops $p(x)$; $p(x) < kq(x)$ $\forall x$. Furthermore, to be more in the area of interest, we can sample x from the Gaussian. That means we choose (x, y) such that:

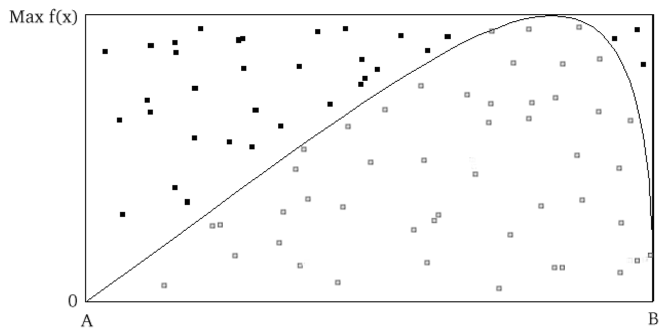

Figure 1: Rejection sampling example, uniform envelope

- $x \sim q(x)$
- $y \sim u(0, kq(x))$

If y is under the $p(x)$ curve accept x, otherwise reject it. Figure 2 shows an example.

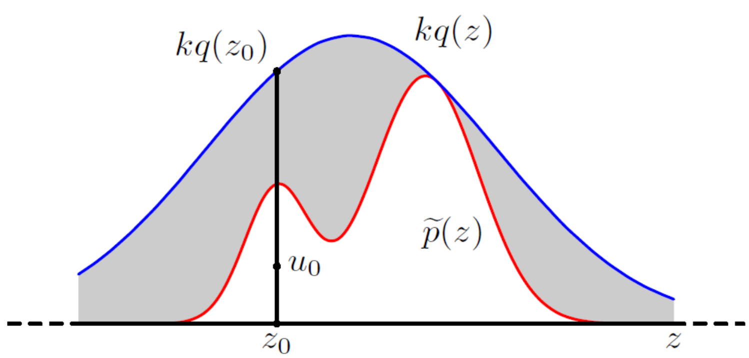

Figure 2: Rejection sampling example, Scaled Gaussian envelope

In general, rejection sampling has a very simple algorithm as follows:

$i = 0$
while $i < n$ **do**
  $x_i \sim q(x)$
  $u \sim u(0, 1)$
  if $u < \frac{p(x_i)}{kq(x_i)}$ **then**
    accept x_i
    $i \leftarrow i + 1$
  else
    reject x_i
  end if
end while

Example: when we sample from Gamma distribution, we usually use the Cauchy distribution as the envelope.

Now let us try to sample from posterior: $p(\theta \mid x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} = \frac{p(x|\theta)p(\theta)}{z}$ in which z is the normalizer (a scalar). So, this time instead of $p(x_i)$ we have $p(x_i \mid \theta)p(\theta)$. Then, $p(accept) \propto \frac{1}{M}$ and the execution time $\propto O(Mn)$.

1.1.3 Adaptive rejection sampling

Here we first choose some random points and using the tangent to distribution $p(x)$ at that points we create a piecewise linear envelope, $q(x)$, around $p(x)$. Then we follow the same algorithm we had for rejection sampling as before, except this time whenever a sample is rejected we update the piecewise envelope $q(x)$ by getting its intersection with the tangent at this point. In Figure 3 x_1, x_2, x_3 are initial random points that make $q(x)$; the bold dashed line. Now the area between $q(x)$ and $p(x)$, the curve, is the rejection area that is updated whenever a sample is rejected.

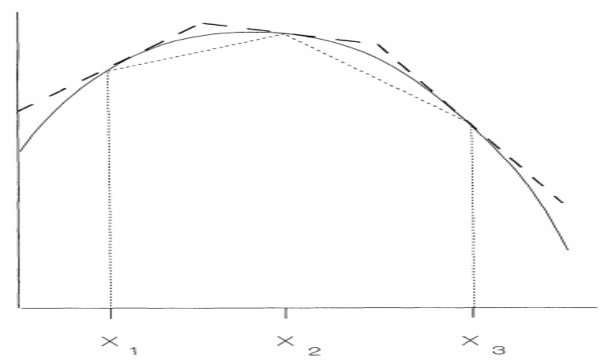

Figure 3: Adaptive rejection sampling example, piecewise linear envelope

The problem with adaptive rejection sampling is that it is simple only when the probability distribution is log convex. For example it is really hard (not applicable) to apply adaptive rejection sampling on distributions like $p(x)$ in Figure 2.

1.2 Computing expectation: $\int f(x)p(x)dx$

Here focus is on how we compute expectation.

1.2.1 Importance sampling

We have:

$$\int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx \approx 1/n\sum_{i=1}^{n}\frac{p(x_i)}{q(x_i)}f(x_i). \tag{2}$$

By sampling using $q(x)$ we have $p(x_i)/q(x_i)$s as the *importance weights.* These correct the bias introduced by sampling from the wrong distribution. Let us imagine:

- $p(x) = \frac{\tilde{p}(x)}{z_p}$

- $q(x) = \frac{\tilde{q}(x)}{z_q}$

We can approximate z_p and z_q using importance sampling.

$$\int f(x)p(x)dx = \frac{z_q}{z_p}\int f(x)\frac{\tilde{p}(x)}{\tilde{q}(x)}q(x)dx \approx \frac{z_q}{z_p}1/n\sum_{i=1}^{n}\frac{\tilde{p}(x)}{\tilde{q}(x)}f(x_i) \tag{3}$$

Letting $r_i = \frac{\tilde{p}(x)}{\tilde{q}(x)}$ then we have:

$$\int f(x)p(x)dx \approx \frac{z_q}{z_p}1/n\sum_{i=1}^{n}\tilde{r}_i f(x_i) \tag{4}$$

Then we have:

$$\frac{z_p}{z_q} = \frac{1}{z_p}\int \tilde{p}(x)dx = \int \frac{\tilde{p}(x)}{\tilde{q}(x)}q(x)dx \approx 1/n\sum_{i=1}^{n}\tilde{r}_i \tag{5}$$

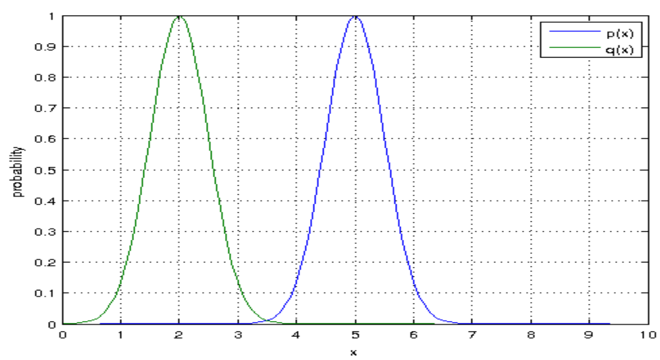

Figure 4: An example of bad match up between p and q in importance sampling

Putting (1) and (2) together we have:

$$\int f(x)p(x)dx \approx \sum_i w_i f(x_i)$$

$$\text{where } w_i = \frac{\frac{\tilde{p}(x)}{\tilde{q}(x)}}{\sum_i \frac{\tilde{p}(x)}{\tilde{q}(x)}}$$

and x_is are sampled from $q(x)$. If q and p don't match up very well, we need lots of samples; see Figure 4. So we need a good match between q and p. Another point is that we sample from q. But if we want to sample from p it needs one additional step. It is called Sampling Importance-Resampling (SIR). It works in two steps:

- Run importance sampling to obtain weight, sample pairs (w_i, x_i) for $i = (1...n)$
- Then select with replacement x_i with probability w_i

1.2.2 Markov Chain Monte Carlo (MCMC) sampling

It is different from reject or importance sampling. We will discuss this more next time, but here is a preview.

1.2.3 Metropolis-Hastings sampling

Here we have:
$q(x^* \mid x_i)$
Then generate:
$x^* \propto q(x^* \mid x_i)$
And accept x^* with probability:
$min(1, \frac{p(x^*)q(x_i|x^*)}{p(x_i)q(x^*|x_i)})$
If we reject, we keep an additional copy of the previous sample as our accepted sample.