Bayesian Statistics, Monte Carlo Methods (Importance Sampling)

May 31, 2017

- Let us introduce the Dirac-delta function δ_{θ_0} for $\theta_0 \in \Theta$ defined for any $f : \Theta \to R^{n_f}$ as follows:

$$\int_{\Theta} f(\theta)\delta_{\theta_0}(\theta)d\theta = f(\theta_0)$$

- Note that this implies in particular that for $A \subset \Theta$,

$$\int_{\Theta} \mathcal{I}_A(\theta)\delta_{\theta_0}(\theta)d\theta = \int_A \delta_{\theta_0}(\theta)d\theta = \mathcal{I}_A(\theta_0)$$

- Now, for $\theta^{(i)} \sim \pi, i = 1, 2 \cdots, N$, we can introduce the following mixture of Dirac-delta functions

$$\hat{\pi}_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta^{(i)}}(\theta),$$

which is the empirical measure.

- Now consider the problem of estimating $E_{\pi}(f)$. We simply replace π with its sample representation $\hat{\pi}_N$ and obtain

$$E_{\pi}(f) \simeq \int_{\Theta} f(\theta) \sum_{i=1}^{N} \frac{1}{N} \delta_{\theta^{(i)}}(\theta) d\theta$$
$$= \sum_{i=1}^{N} \int_{\Theta} f(\theta) \frac{1}{N} \delta_{\theta^{(i)}}(\theta) d\theta = \frac{1}{N} \sum_{i=1}^{N} f(\theta^{(i)})$$

  which is precisely $S_N(f)$, the Monte Carlo estimator suggested earlier.
- Clearly based on $\hat{\pi}_N$, we can easily estimate $E_{\pi}(f)$ for any f.
- More precisely,

$$E_X[E_{\hat{\pi}_N}(f(X))] = E_{\pi}(f(X)), \text{ and } var_X(E_{\hat{\pi}_N}(f(X))) = \frac{var_{\pi}(f(X))}{N}.$$

- Direct methods feasible for standard distributions: inverse method, composition, etc.
- In case where $\pi \propto \pi^*$ does not admit any standard form, we can use a proposal distribution q on $\mathcal{X}$ where $q \propto q^*$.
- We need q to demoniate π,

$$C = sup_{x \in X} \frac{\pi^*(x)}{q^*(x)} < +\infty.$$

Generating Continuous Random Variables (The Rejection Method)

Suppose we have a method for generating a random variable Y having density function $\pi(x)$. We can use this as basis for generating a random variable X having density function $q(x)$.
Let C be a constant such that

$$\frac{\pi(y)}{q(y)} \leq C \text{ for all } y$$

The Rejection Method
Step 1: Generate Y having density q.
Step 2: Generate a random number U.
Step 3: If $U \leq \frac{\pi(Y)}{Cq(Y)}$, set $X = Y$. Otherwise, return to Step 1.

Generating Continuous Random Variables (The Rejection Method)

- This is a simple generic algorithm but it requires coming up with a bound C.
- Its performance typically degrade exponentially fast with the dimension of X.
- It seems you are wasting some information by rejecting samples.
- You need to wait a random time to obtain some samples from π.
- Is it possible to "recycle" these samples?

Importance Sampling

- Consider again the target distribution π and the proposal distribution q. We only require $\pi(x) > 0 \Rightarrow q(x) > 0$.
- In this case, the Importance Sampling (IS) identity is

$$E_\pi(\phi(X)) = \int_\mathcal{X} \phi(x)\pi(x)dx = \int_\mathcal{X} \phi(x)\frac{\pi(x)}{q(x)}q(x)dx = E_q(w(X)\phi(X))$$

  where the so-called Importance Weight is given by $w(x) = \pi(x)/q(x)$
- This is a simple yet very flexible identity.

- Monte Carlo approximation of q is

$$\hat{q}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(x), \text{ where } X^{(i)} \sim q, i.i.d$$

- It follows that an estimate of $E\pi(\phi(X)) = E_q(w(X)\phi(X))$ is

$$E_{\hat{q}_N}(w(X)\phi(X)) = \frac{1}{N}\sum_{i=1}^{N} w(X^{(i)})\phi(X^{(i)}),, X^{(i)} \sim q, i.i.d$$

- It corresponds to the following approximation

$$\hat{\pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N} w(X^{(i)})\delta_{X^{(i)}}(x), \text{ where } X^{(i)} \sim q, i.i.d$$

- We have

$$E_X[E_{\hat{q}_N}(w(X)\phi(X))] = E_\pi(\phi(X)),$$

and

$$\begin{aligned} var_X(E_{\hat{q}_N}(w(X)\phi(X))) &= \frac{var_q(w(X)\phi(X))}{N} \\ &= \frac{E_\pi(w(X)\phi^2(X)) - E_\pi^2(\phi(X))}{N}. \end{aligned}$$

- In practice, it is recommended to ensure

$$E_\pi(w(X)) = \int \frac{\pi^2(x)}{q(x)} dx < \infty$$

- Even if it is not necessary, it is actually even better to ensure that:

$$sup w(x) < \infty . x \in X$$

Example

- Consider the function $h(x) = 10exp(-2|x-5|)$. Suppose that we want to calculate $E(h(X))$, where $X \sim Uniform(0,1)$. That is, we want to calculate the integral

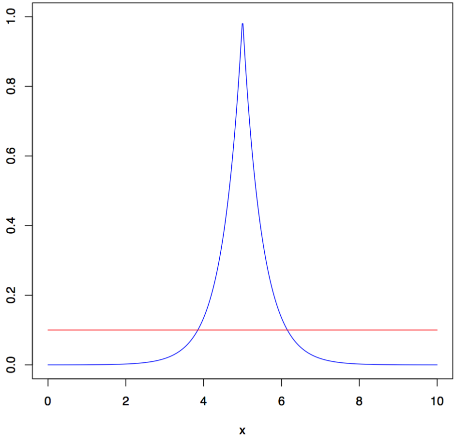
$$\int_0^{10} exp(-2|x-5|)dx.$$

```
X <- runif(100000,0,10)
Y <- 10*exp(-2*abs(X-5))
c( mean(Y), var(Y) )
[1] 0.9919611 3.9529963
```

- The function h in this case is peaked at 5, and decays quickly elsewhere, therefore, under the uniform distribution, many of the points are contributing very little to this expectation.

Example

Figure : The integrand (blue) and the density being integrated against (red) approach 1



- Rewrite the integral as:

$$\int_0^{10} 10exp(-2|x-5|)\frac{1/10}{\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}}\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2}/2dx$$

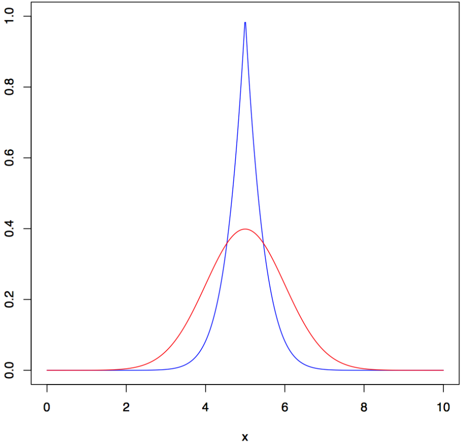
- That is, $E(h(X)w(X))$, where $X \sim N(5, 1)$.
- The integral is:

$$\int_0^{10} exp(-2|x-5|)\sqrt{2\pi}e^{(x-5)^2/2}\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}dx$$

```
w <- function(x) dunif(x, 0, 10)/dnorm(x, mean=5, sd=1)
f <- function(x) 10*exp(-2*abs(x-5))
X=rnorm(1e5,mean=5,sd=1)
Y=w(X)*f(X)
c( mean(Y), var(Y) )
[1] 0.9999271 0.3577506
```

Example

Figure : The integrand (blue) and the density being integrated against (red) approach 2



Example

- Double exponential density: $\pi(x) = \frac{1}{2}e^{-|x|}$. The CDF is

$$F(x) = \frac{1}{2}e^{x}I(x \leq 0) + (1 - e^{-x}/2)I(x > 0)$$

- Estimate $E(X^2)$. That is, calculate the integral $\int_{\infty}^{\infty} x^2 \frac{1}{2}e^{-|x|}dx$
- Rewrite the integral as:

$$\int_{-\infty}^{\infty} x^2 \frac{\frac{1}{2}e^{-|x|^2}}{\frac{1}{\sqrt{8\pi}}e^{-\frac{x^2}{8}}} \frac{1}{\sqrt{8\pi}}e^{-\frac{x^2}{8}}dx$$

```
X <- rnorm(1e5, sd=2)
Y <- (X^2) * .5 * exp(-abs(X))/dnorm(X, sd=2)
mean(Y)
[1] 1.998898
```

Optimal IS Distribution

- For a given test function, one can minimize the IS variance using:

$$q^{opt}(x) = \frac{|\phi(x)|\pi(x)}{\int_{\mathcal{X}} |\phi(x)|\pi(x)dx}$$

- Proof.

$$Var_q(w(x)\phi(x)) = \int q(x)\frac{\pi^2(x)}{q^2(x)}\phi^2(x)dx - (\int q(x)\frac{\pi(x)}{q(x)}\phi(x)dx)^2$$

and $\int q(x)\frac{\pi^2(x)}{q^2(x)}\phi^2(x)dx \geq (\int q(x)\frac{\pi(x)}{q(x)}|\phi(x)|dx)^2 =$
$(\int q(x)\frac{\pi(x)}{q(x)}|\phi(x)|dx)^2$.
This lower bound is attained for $q^{opt}(x)$.

Normalized Importance Sampling

- In most if not all applications we are interested in, standard IS cannot be used as the importance weights $w(x) = \pi(x)/q(x)$ cannot be evaluated in closed-form. In practice, we typically only know $\pi(x) \propto \pi^*(x)$ and $q(x) \propto q^*(x)$.
- Normalized IS identity is based on

$$\pi(x) = \frac{\pi^*(x)}{\int \pi^*(x)dx} = \frac{w^*(x)q^*(x)}{\int w^*(x)q^*(x)dx} = \frac{w^*(x)q(x)}{\int w^*(x)q(x)dx}$$

Normalized Importance Sampling

- For any test function $\phi(x)$,we can also write

$$E_\pi(\phi(x)) = \frac{E_q(w^*(X)\phi(X))}{E_q(w^*(X))} = \frac{E_q(w(X)\phi(X))}{E_q(w(X))}$$

- Given a Monte Carlo approximation of q:

$$\hat{q}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X^{(i)}}(x), \text{ where } X^{(i)} \sim q, i.i.d$$

- Then,

$$\hat{\pi}_N(x) = \frac{1}{N}\sum_{i=1}^{N} W(X^{(i)})\delta_{X^{(i)}}(x),$$

where $W(X^{(i)}) = \frac{w^*(X^{(i)})}{\sum_{j=1}^{N} w^*(X^{(j)})}$

$$E_{\hat{\pi}_N}(\phi(X)) = \frac{1}{N}\sum_{i=1}^{N} W(X^{(i)})\phi(X^{(i)}),, X^{(i)} \sim q, i.i.d$$

- The estimates are a ratio of estimates.

Example

- Suppose $X_1, \cdots, X_n \sim \textit{Binomial}(10, \theta)$ where $\theta \in (0,1)$ has a *Beta*$(5,3)$ prior density: $p(\theta) = \frac{\Gamma(8)}{\Gamma(5)\Gamma(3)}\theta^4(1-\theta)^2$. We want to estimate the mean of the posterior distribution: $\int_0^1 \theta p(\theta|x_1, \cdots, x_n)d\theta$.
- Take q to be the *Beta*(α, β) density, where $\alpha = c\bar{X}, \beta = c(10 - \bar{X})$, where $\bar{X}$ is the sample mean.This will ensure that q is peaked near $\bar{X}/10$, which is where the posterior distribution should have a lot of mass.

Example

- The joint distribution of the data, given θ:

$$
\begin{aligned}
p(x_1, x_2, \cdots, x_n|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) \\
&\propto \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{10n-\sum_{i=1}^{n} x_i} \\
&= \theta^{n\bar{X}}(1-\theta)^{n(10-\bar{X})}
\end{aligned}
$$

- The posterior density:

$$
\begin{aligned}
p(\theta|x_1, x_2 \cdots, x_n) &\propto p(x_1, x_2, \cdots, x_n|\theta)p(\theta) \\
&\propto \theta^{n\bar{X}}(1-\theta)^{n(10-\bar{X})}p(\theta) \\
&\propto \theta^{n\bar{X}}(1-\theta)^{n(10-\bar{X})}\theta^4(1-\theta)^2 \\
&= \theta^{n\bar{X}+4}(1-\theta)^{n(10-\bar{X})+2}
\end{aligned}
$$

- The log of this quantity is:

$$
(n\bar{X}+4)\log\theta + (n(10-\bar{X})+2)\log(1-\theta)
$$

Example

- Suppose $X_1, \cdots, X_n \sim N(0, \theta)$ and we specify a *Gamma*$(3, .5)$ distribution for the prior of θ. We will use a trial density q which is Gamma distributed with $\alpha = cs^2$, and $\beta = c$, where c is a positive constant, and s^2 is the sample variance. So the mean of the trial distribution will be s^2. Choose c to optimize estimation precision.

- The joint distribution of the data, given θ,

$$\begin{aligned} & p(x_1, x_2, \cdots, x_n|\theta) \\ = & \prod_{i=1}^{n} \sqrt{\frac{1}{2\pi\theta}} e^{-x_i^2/2\theta} \\ \propto & \theta^{-n/2} \exp(-\frac{1}{2\theta} \sum_{i=1}^{n} x_i^2) \end{aligned}$$

- The posterior distribution is proportional to

$$\begin{aligned} p(\theta|x_1, \cdots, x_n) &= p(x_1, \cdots, x_n|\theta)p(\theta) \\ &\propto \theta^{-n/2} \exp(-\frac{1}{2\theta} \sum_{i=1}^{n} x_i^2) 2^{-3}\Gamma(3)^{-1}\theta^2 e^{-\theta/2} \\ &\propto \theta^{-n/2} \exp(-\frac{1}{2\theta} \sum_{i=1}^{n} x_i^2)\theta^2 e^{-\theta/2} \\ &= \theta^{(4-n)/2} \exp(-\frac{1}{2\theta} \sum_{i=1}^{n} (\theta^2 + x_i^2)) \end{aligned}$$

Normalized Importance Sampling

- Contrary to standard IS, this estimate is biased but asymptotically unbiased by the LLN it is asymptotically consistent.
- Derivation of the asymptotic bias and variance based on the delta method.