STA 602 - Intro to Bayesian Statistics

Lecture 7

Li Ma

Duke University

The need for evaluating expectations

$u_1, u_2, \dots, u_n \overset{i.i.d}{\sim} p$

$\frac{1}{n}(g(u_1) + \dots + g(u_n))$

by strong law of large numbers.

- Consider the following expectation

converge.

$$\mathrm{E}_p g = \int g(u)p(u)du$$

where $g(u)$ is some general integrable function and $p(u)$ a probability density function.

- The notation $\mathrm{E}_p g$ indicates that its the expectation of the function g under the distribution p for its argument.
- In carrying out Bayesian inference, we commonly need to evaluate integrals of the above form.
- Very often, u is the unknown parameter θ, and p is the posterior density of θ given the data.

Some examples (identify what "g" and "p" are)

- For computing posterior mean of some function $h(\theta)$,

$$\mathrm{E}(h(\theta)\,|\,\mathbf{x}) = \int h(\theta)p(\theta|\mathbf{x})d\theta.$$

- For computing posterior quantiles and credible intervals for $h(\theta)$,

$$\mathrm{P}(h(\theta) \le c\,|\,\mathbf{x}) = \mathrm{E}(\mathbf{1}_{\{h(\theta)\le c\}}|\mathbf{x}) = \int \mathbf{1}_{\{h(\theta)\le c\}}p(\theta|\mathbf{x})d\theta$$

- For computing predictive probabilities,

$$\begin{aligned}
\mathrm{P}(h(x_{n+1}) \in A\,|\,\mathbf{x}_n) &= \mathrm{E}(\mathbf{1}_{\{h(x_{n+1})\in A\}}|\mathbf{x}_n) \\
&= \int \mathbf{1}_{\{h(x_{n+1})\in A\}}p(x_{n+1}\,|\,\mathbf{x}_n)dx_{n+1} \\
&= \int\int \mathbf{1}_{\{h(x_{n+1})\in A\}}p(x_{n+1},\theta\,|\,\mathbf{x}_n)d\theta dx_{n+1} \\
&= \int\int \mathbf{1}_{\{h(x_{n+1})\in A\}}p(x_{n+1}|\theta,\mathbf{x}_n)p(\theta|\mathbf{x}_n)d\theta dx_{n+1}.
\end{aligned}$$

$\int p(x_{n+1},\theta\,|\,x_n)d\theta$

"u" $= (\theta, x_{n+1})$
$g(u) = \mathbb{1}(h(x_{n+1})\in A)$
θ can be anything

Approaches to evaluate the integral

- We can try to evaluate it analytically, such as in the case of exponential families.
- We can carry out numerical integration, Laplace approximation, numerical quadrature, etc.
  - The difficulty and complexity of numerical integration grows quickly with the dimensionality of θ. For example, if one adopt K grid points in each dimension, then one need a total of K^d grid points.
  - The numerical integration becomes impractical when θ is more than a few dimensions.
- *Monte Carlo* simulation.

The Monte Carlo (MC) idea

- Suppose we are able to generate independent draw from the density p:
$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)} \overset{\text{iid}}{\sim} p.$$
- Then by the law of large number (LLN), we have
$$\bar{g} = \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to_{a.s.} \mathrm{E}_p g = \int g(\theta)p(\theta)d\theta \quad \text{when } S \to \infty.$$
when the integral is finite.
- Central limit theorem implies that if in addition, $g(\theta)$ has finite variance under $\theta \sim p$, then
$SE_{mc} = \sqrt{\frac{Var_p g}{S}}.$
$$\sqrt{S}(\bar{g} - \mathrm{E}_p g) \to_d \mathrm{N}(0, \mathrm{Var}_p g)$$
where $\mathrm{Var}_p g = \int (g(\theta) - \mathrm{E}_p g)^2 p(\theta)d\theta < \infty$. $\mathbb{E}_p (g - \mathbb{E}_p g)^2.$
- *Regardless of the dimensionality of θ*, the error of the Monte Carlo (MC) estimator for the integral is $O_p(1/\sqrt{S})$. Caveat: The constant can be large sometimes!

An example

- Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are independent random variables where $\theta_1 \sim \mathrm{N}(0,1)$ and $\theta_2 \sim \mathrm{Beta}(10,20)$.
- What is the expection of $g(\boldsymbol{\theta}) = (\sqrt{\theta_2} + \theta_1^2)^{1/3}$. That is

$$\mathrm{E}_p g = \int (\sqrt{\theta_2} + \theta_1^2)^{1/3} p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

```
options(digits=4)
S=10000
theta1 <- rnorm(S,mean=0,sd=1)
theta2 <- rbeta(S,10,20)
mc.samples <- (sqrt(theta2)+theta1^2)^(1/3)
Eg <- mean(mc.samples)
print(Eg)
```

```
## [1] 1.095
```

Quantifying Monte Carlo error

- The variance of the MC estimator for the integral *under repeated MC runs* (differentiate this from sampling error!) is

$$\mathrm{Var}_p(\bar{g}) = \frac{1}{S}\mathrm{Var}_p g = \frac{1}{S}\int \left(g(\theta) - \mathrm{E}_p g\right)^2 p(\theta)d\theta.$$

- One can estimate it using the sample variance of the MC sample,

$$\widehat{\mathrm{Var}}_p(\bar{g}) = \frac{1}{S}\widehat{\mathrm{Var}}_p g = \frac{1}{S(S-1)}\sum_{s=1}^{S}(g(\theta^{(s)}) - \bar{g})^2.$$

- Its square root is the *Monte Carlo standard error*

$$s.e._{MC} = \sqrt{\frac{1}{S(S-1)}\sum_{s=1}^{S}(g(\theta^{(s)}) - \bar{g})^2}$$

P $\to \mathrm{Var}_p g.$

which quantifies the random error induced by the MC simulation in the estimate for the integral. It converges to 0 at $1/\sqrt{S}$ rate. not depend on dim.

In the previous example

- The sample variance for the MC estimates is

```
var.g.hat <- var(mc.samples)
print(var.g.hat)
```

```
## [1] 0.07409
```

```
se.mc <- sd(mc.samples)/sqrt(S)
print(se.mc)
```

```
## [1] 0.002722
```

- Exercise: Estimate $\mathrm{P}(\theta_1 > \sqrt{\theta_2})$ and find the MC standard error.

$\mathbb{E}\left[\mathbb{1}(\theta_1 > \sqrt{\theta_2})\right]$ so here $g(\theta) = \mathbb{1}(\theta_1 > \sqrt{\theta_2})$.

Example: Bayesian inference

- Estimate posterior expectation of $h(\theta)$,

$$\mathrm{E}(h(\theta) \,|\, \mathbf{x}) \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}).$$

- Estimate posterior tail probability $\mathrm{P}(h(\theta) \leq c)$,

$$\mathrm{P}(h(\theta) \leq c) \approx \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}_{\{h(\theta) \leq c\}}.$$

- The above implies that the αth quantile of the sample $h(\theta^{(1)}), \ldots, h(\theta^{(S)})$ converges to the αth quantile of $h(\theta)$.

$$F_{h(\theta)}^{-1}(\alpha) \approx \hat{F}_{h(\theta)}^{-1}(\alpha).$$

- Hence credible interval based on the empirical quantiles of the sample $h(\theta^{(1)}), \ldots, h(\theta^{(S)})$ give an estimate for the corresponding credible interval on $h(\theta)$.

Example: Bayesian inference (cont'ed)

- To estimate predictive probability $\mathrm{P}(h(x_{n+1}) \in A \mid \mathbf{x}_n)$, draw samples

$$(\theta^{(1)}, x_{n+1}^{(1)}), (\theta^{(2)}, x_{n+1}^{(2)}), \ldots, (\theta^{(S)}, x_{n+1}^{(S)}) \overset{\text{iid}}{\sim} p(\theta, x_{n+1} \mid \mathbf{x}_n).$$

- This can be done in two steps, first draw

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)} \overset{\text{iid}}{\sim} p(\theta \mid \mathbf{x}_n).$$

  Then for each $s = 1, 2, \ldots, S$, draw

x_n

$$x_{n+1}^{(s)} \mid \theta^{(s)}, \overset{\text{ind}}{\sim} p(x_{n+1} \mid \theta^{(s)}, \mathbf{x}_n).$$

- Now, the MC estimate is given by

$$\mathrm{P}(h(x_{n+1}) \in A \mid \mathbf{x}_n) \approx \frac{1}{S} \mathbf{1}_{\{h(x_{n+1}^{(s)}) \in A\}}.$$

- So marginalizing out θ boils down to drawing samples from the joint distribution and simply "ignore" the θ values.

Example: Bayesian inference (cont'ed)

- Posterior predictive checks. Again proceed in two steps
  - Draw samples of θ from the posterior given the original data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)} \stackrel{\text{iid}}{\sim} p(\theta \,|\, \mathbf{x}).$$

  - Draw replicate data sets for each θ draw

$$\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}) \,|\, \theta^{(i)} \sim p(\mathbf{x}|\theta^{(i)})$$

- Compare the replicate data sets with the original data $\mathbf{x}$. For example,
  - Compute some summary statistic $h(\mathbf{x}^{(i)})$ for each replicate and compare the resulting histogram with $h(\mathbf{x})$.

Remarks

- The key to applying MC in practice is the ability to draw *independent* samples from p.
- For common one-dimensional distributions, we can directly use the corresponding R functions such as rnorm, rbeta, rpois, ...
- We need general sampling strategies when p is outside of familiar parametric families, and situations when p is known only up to a constant.

Remarks (cont'ed)

- If the target p is one-dimensional and analytically simple (have evaluable inverse CDF), this can be done exactly with inversde-CDF sampling
- If the target p is low-dimensional and known up to a constant (e.g., rejection sampling and importance sampling).
- Depending on the nature of the integrand g and the distribution p, the variance $\mathrm{Var}_p g$ can be huge. In that case, MC error can be very large (in practice, S is never infinite!) There are some techniques for reducing the Monte Carlo standard errors. (e.g., importance sampling)
- In Bayesian inference, the posterior distribution $p(\theta|\mathbf{x})$ is often known only up to a normalizing constant. It turns out one can get away this difficulty by drawing *correlated* samples from p using Markov chains (e.g., MCMC)

Inverse CDF sampling

- For one-dimensional distributions, let F be the corresponding CDF for p.
- Then one can draw i.i.d. samples from p (or F) by
  - Draw independent samples $U^{(1)}_{\cancel{1}}, U^{(2)}_{\cancel{2}}, \ldots, U^{(S)}_{\cancel{S}} \overset{\text{iid}}{\sim} \text{Unif}(0,1)$.
  - Let $\theta_i = F^{-1}(U_i)$.
- Then $\theta^{(1)}_{\cancel{1}}, \theta^{(2)}_{\cancel{2}}, \ldots, \theta^{(S)}_{\cancel{S}}$ are i.i.d. samples from p.

$\theta_i = F^{-1}(u_i) \sim p$ (target distribution)

$\uparrow$ Unif(0,1)

- To see this,

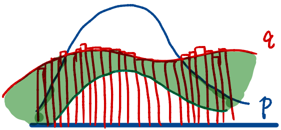
$F(\theta_i) = F(F^{-1}(u_i)) = u_i$

$$P(\theta_i \le c) = P(F(\theta_i) \le F(c)) = P(U_i \le F(c)) = F(c).$$

So F is indeed the CDF for θ_i.

- For multivariate distribution or for one-dimensional distributions where F is not available in closed form, we need some alternatives.

Rejection sampling



- Suppose we wish to generate samples from a probability density p (up to a constant), but we don't necessarily know the normalizing constant for p that makes it a density.
- On the other hand, we know how to generate samples from q, which is a probability density (up to a constant) that dominates the function p, i.e., q has larger support than p.
- Then we can generate a sample from the desired distribution in two steps.

Rejection sampling

- First, draw
$$\theta \sim q.$$
Then generate
$$U \sim \text{Uniform}(0,1).$$
- Keep the sample θ from q, if $U \leq r(\theta) = p(\theta)/Mq(\theta)$ for some constant $M > 0$ large enough such that $p < Mq$.
  - Discard (or reject) the sample θ otherwise.
- Repeat the above until we have the number of samples S we want.
- In other words, we draw from q instead, and accept a draw θ with probabiliy $r(\theta) = p(\theta)/Mq(\theta)$.

The vailidity of rejection sampling

- We can verify that a sample generated from the above process indeed has the desired distribution. (Draw a figure.)
- In other words, conditional on the event that a draw θ from q (or more generally $q/\int q$) is accepted, its pdf is indeed $p/\int p$.
- To see this, consider two values θ_1 and θ_2 in the support of p.
- What is the "odds" for sampling these two values under the above strategy?

$$\frac{q(\theta_1)\cdot r(\theta_1)}{q(\theta_2)\cdot r(\theta_2)} = \frac{q(\theta_1)\cdot \frac{p(\theta_1)}{Mq(\theta_1)}}{q(\theta_2)\cdot \frac{p(\theta_2)}{Mq(\theta_2)}} = \frac{p(\theta_1)}{p(\theta_2)}.$$

The vailidity of rejection sampling

- A more formal proof:
- By Bayes' theorem,

$$
\begin{aligned}
p(\theta|U < r(\theta)) &= \frac{\frac{q(\theta)}{\int q} \cdot P(U < r(\theta)|\theta)}{\mathrm{P}(U < r(\theta))} \\
&= \frac{q(\theta)/\int q \cdot p(\theta)/Mq(\theta)}{\mathrm{P}(U < r(\theta))} \\
&= \frac{\frac{p(\theta)}{M\int q}}{P(U < r(\theta))}
\end{aligned}
$$

- Now, the denominator (the marginal probability of acceptance) is

$$
\begin{aligned}
\mathrm{P}(U < r(\theta)) &= \mathrm{EP}(U < r(\theta)|\theta) \\
&= \int \frac{p(\theta)}{Mq(\theta)} \frac{q(\theta)}{\int q} d\theta = \frac{\int p}{M\int q}.
\end{aligned}
$$

Example: Political poll

Requirement: q must have larger support than p.

- Suppose for our political poll example, one decides to use a prior

$$p(\theta) \propto e^{-(\theta-0.5)^2} \quad \text{for } \theta \in (0,1).$$

- By Bayes theorem, we know the posterior of θ is

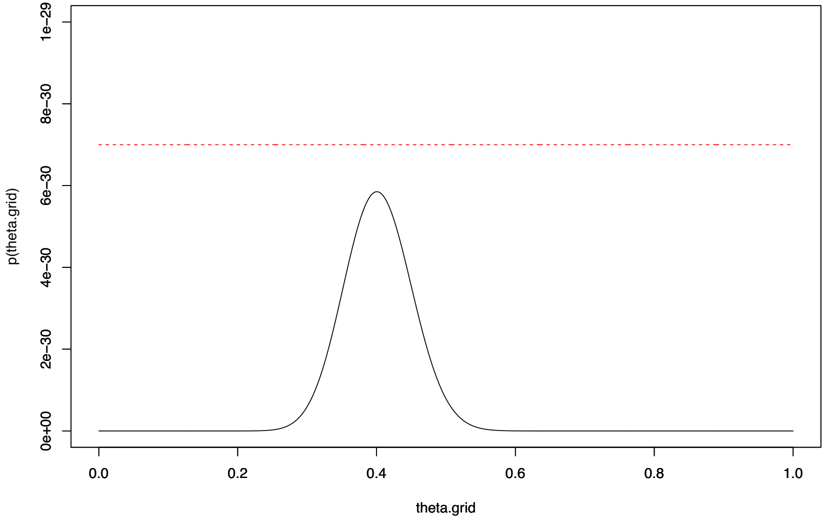
$$p(\theta|x) \propto \theta^x (1-\theta)^{n-x} e^{-(\theta-0.5)^2}.$$

  This is not a conjagte model so we don't have simple analytic form for the posterior density.
- Apply rejection sampling, to sample from this density.

Example: Political poll (R code)

```r
x=40; n=100

# p is the target distribution to sample from
p = function(theta) {
  theta^x*(1-theta)^(n-x) *exp(-(theta-0.5)^2)
}

# q is something easy to sample
q = function(x) { dunif(x) }

# Choose a constant that satisfies f<M*g,
# but make M as small as possible
# In finding this value I ``cheated''
M=7e-30
```

Plot the function p and the function Mq



Start the sampling

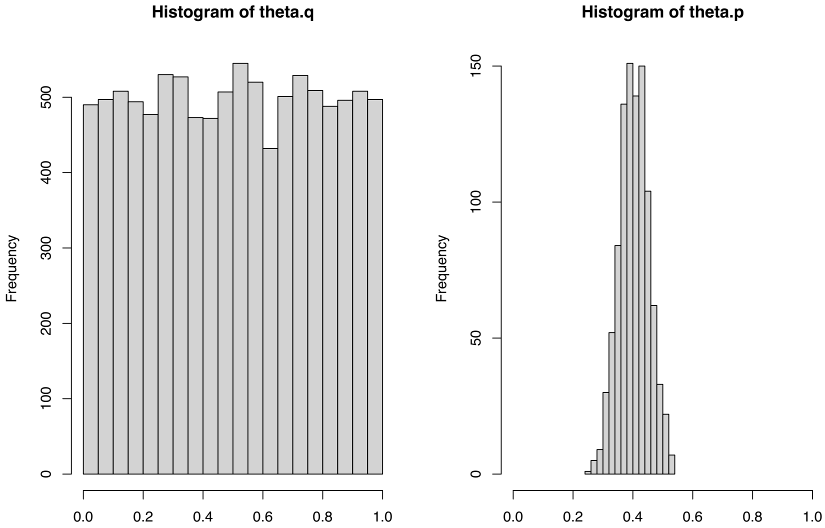
```r
# total number of trial draws
S=10000

# draw from q
theta.q = runif(S)

# compute acceptance probability
acc.prob = p(theta.q)/(M*q(theta.q))

# indicator for acceptance
acc.ind = rbinom(S,size=1,prob=acc.prob)

# proportion of accepted draws
mean(acc.ind)
```

```
## [1] 0.0985
```

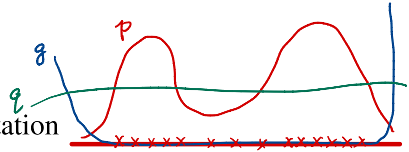
```r
# the accepted draws
theta.p = theta.q[as.logical(acc.ind)]
```

Plot the histogram of the samples

```r
par(mfrow=c(1,2))
hist(theta.q,xlim=xlim)
hist(theta.p,xlim=xlim)
```



Importance sampling (IS)



most samples are useless

sample from q instead.

- Consider again the following expectation

$$\mathrm{E}_p g = \int g(u)p(u)du$$

$$= \int g(u)\frac{p(u)}{q(u)}q(u)du = \int g(u)w(u)q(u)du$$

- Again we might have trouble directly sample from p, but know how to sample from q which dominates p (i.e., has larger support).
- In some cases, even if we can sample from p, the function $g(u)$ is such that its values are determined mostly in low probability regions of p. Draw an example.
  - Similating from p and apply standard Monte Carlo is inefficient, as most draws are not very useful as all.
  - This leads to very low fraction of the MC samples to be of much relevance for evaluating the integral, and hence high MC standard error.

Importance sampling (IS)

- Idea: Can we sample instead from a distribution q, which oversamples the region that matters most for this integral relative to p, then corrects for the difference in p and q?
- In contrast to rejection sampling, no samples are rejected, and so we end up with a sample from q rather than p, but are weighted differently in computing the MC estimate.

Importance sampling (IS)

The choice of proposal: case by case
depend on g.

- Rewrite the above integral

Practical: sequential & adaptive importance sampling

$$
\begin{aligned}
\mathrm{E}_p g = \int g(u)p(u)du &= \int g(u)\frac{p(u)}{q(u)}q(u)du \\
&= \int g(u)w(u)q(u)du = \mathrm{E}_q gw.
\end{aligned}
$$

  where q is a probability distribution, called the *proposal distribution*, and $w(u) = p(u)/q(u)$ are called the *importance weights*.
- We can sample from the distribution q instead, and use MC to evalute the production of g and w. That is for a sample

$$u^{(1)}, u^{(2)}, \ldots, u^{(S)} \overset{\text{iid}}{\sim} q$$

  compute the MC estimate

$$\frac{1}{S}\sum_{i=1}^{S} g(\theta^{(i)})w(\theta^{(i)}). \xrightarrow{P} \mathbb{E}_p g$$

Importance sampling (IS)

Previous: $\frac{1}{S}\sum_i g(u_i)w(u_i)$, $u_i \overset{\text{i.i.d}}{\sim} \tilde{q}$

here $w(u_i) = \frac{p}{q} = c \cdot \frac{\tilde{p}}{\tilde{q}}$, so summation $\longrightarrow c \cdot \mathbb{E}_{\tilde{p}}\, g$

- Often p and/or q is known only up to a normalizing constant and so the exact weight isn't known. Instead use

$$\frac{\sum_{i=1}^{S} g(\theta^{(i)})w(\theta^{(i)})}{\sum_{i=1}^{S} w(\theta^{(i)})} \longrightarrow \mathbb{E}_{\tilde{p}}\, g, \text{ where } \tilde{p} = \frac{p}{\int p}, \; \tilde{q} = \frac{q}{\int q}$$

  where $w = p/q$ is no longer the actual density ratio but an (unknown) constant c times the actual density ratio $\tilde{p}/\tilde{q}$ where $\tilde{p} = p/\int p$ and $\tilde{q} = q/\int q$ are the underlying densities.

- This is called *self-normalizing* IS and is justified by the fact that

$$\frac{\sum_{i=1}^{S} g(\theta^{(i)})w(\theta^{(i)})}{\sum_{i=1}^{S} w(\theta^{(i)})} = \frac{\frac{1}{S}\sum_{i=1}^{S} g(\theta^{(i)})w(\theta^{(i)})}{\frac{1}{S}\sum_{i=1}^{S} w(\theta^{(i)})}.$$

  The denominator Also practical: adaptive

$$\frac{1}{S}\sum_{i=1}^{S} w(\theta^{(i)}) \rightarrow_{a.s.} \mathrm{E}_{\tilde{q}} w = \mathrm{E}_{\tilde{q}}\, c(\tilde{p}/\tilde{q}) = \int c\tilde{p}(u)/\tilde{q}(u)\tilde{q}(u)du = c\int \tilde{p}(u)du = c,$$

  which is the appropriate normalizing constant for the numerator.

- In practice, this form of IS is used anyway due to its low variance.

Example: Political poll

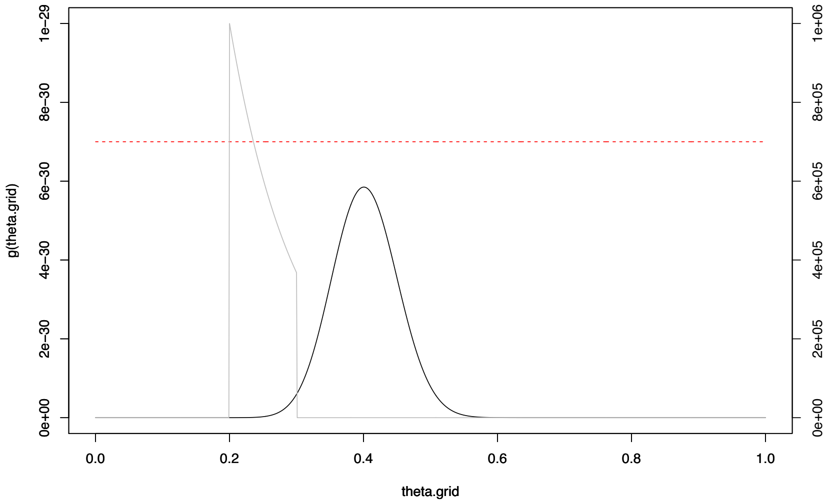
- Again consider our ongoing example.
- Suppose the governor will spend money on ads if $20\% \leq \theta \leq 30\%$, and the amount to be spent is a function the following function

$$g(\theta) = \begin{cases} \$1,000,000 \times e^{-10(\theta - 0.2)} & \text{if } 0.2 \leq \theta \leq 0.3 \\ \$0 & \text{otherwise.} \end{cases}$$

R code

```r
g = function(theta) {
  return(1000000*exp(-10*(theta-0.2))
         * as.integer(theta>=0.2)
         * as.integer(theta<=0.3))
}
```

Plot the function p and the function Mq



Sample from the proposal q, and compute IS estimate

```
# total number of draw from q
S=100000

# draw from q
theta.q = runif(S)

# compute the importance weights
w = p(theta.q)/q(theta.q)

# compute estimate for the integral
# using self-normalizing weights
sum(w*g(theta.q))/sum(w)
```

```
## [1] 6316
```

Remarks

Importance sampling: $\frac{1}{S}\sum_{i=1}^{S} g(\theta^{(i)})\cdot w(\theta^{(i)}),\ \theta^{(i)}\sim q \ \rightsquigarrow \ \frac{\operatorname{Var} \hat{g}_w}{S} = \frac{\operatorname{Var}_q(g(q)\cdot w(q))}{S}$

Vanilla sampling: $\frac{1}{S}\sum_{i=1}^{S} g(\theta^{(i)}),\ \theta^{(i)}\sim p, \ \rightsquigarrow \ \frac{\operatorname{Var} g_p}{S_{eff}}$

- Importance sampling is often used just as a device for sampling from the target p.
  - In this case, the effective sample size can be defined as the sample size of i.i.d. draws from p that gives the same Monte Carlo variance.

    $\frac{1}{S}\operatorname{Var}\hat{g}_w = \frac{1}{S_{eff}}\operatorname{Var} g_p$

    (depend on q and g)
    $$S_{eff} = S \cdot \frac{\operatorname{Var}\hat{g}_p}{\operatorname{Var}\hat{g}_w}.$$
  - Similar to rejection sampling, the S_{eff} is higher when q is close to p/m.
  - When $q = p$ then $S_{eff} = S$.

    In MCMC case, $S_{eff} \ll S$.
    Here generally, $S_{eff} > S$.
- The proposal q can be data-dependent.
  - It's merely a mathematical/computational device.
  - Inference is still under p.
  - Important in high-dimensional problems.

Challenges with multi-dimensional model space

- Vanilla rejection sampling and importance sampling are most helpful for single-parameter or low-dimensional models.
- For moderate to high-dimensional models, it becomes very difficult to design a reasonable effective proposal distribuiton.
- Some strategies to overcome such difficulties exist include
  - Construct proposals adaptively step-by-step. (E.g., sequential importance sampling.)
  - Drawing correlated sample from the target distribution rather than independent samples. (E.g., MCMC.)
- Each encounters their own challenges as well.