Second homework assignment

Professional Master in Economics
Hedibert Freitas Lopes

Bayesian Learning
Due date: 7:30pm, May 23rd, 2021.

Prepare one (and only one) PDF file with your solutions

Send the file to our T.A. Igor Martins (`igorfbm@al.insper.edu.br`)

Assignments will be delivered in pairs (the pairs will be randomly assigned in class)

Solution will be posted on the course webpage at 8pm, Sunday, May 23rd, 2021.

Problem 1: Sampling data from a given distribution via sampling importance resampling (SIR)

In R we can sample from various distributions without ever realizing what is actually going on behind the curtain. Here is a short list: $N(0,1)$, $t_4(0,1)$, $Gamma(3,1)$ and $Binomial(10,0.25)$. Try them out:

```
n = 1000
par(mfrow=c(2,2))
x=rnorm(n);hist(x);mean(x);var(x)
x=rt(n,df=4);hist(x);mean(x);var(x)
x=rgamma(n,3,1);hist(x);mean(x);var(x)
x=rbinom(n,10,0.25);hist(x);mean(x);var(x)
```

Let us now introduce a couple of "new" distributions (not really new!!!):

1. Standard skew-normal distribution with skewness parameter equal to 3:

$$p_1(x) = 2\phi(x)\Phi(3x), \qquad \text{for } x \in \mathbb{R},$$

   where $\phi(x)$ is the density of the standard normal distribution evaluated at x and $\Phi(x)$ is the cumulative distribution function of the standard normal distribution evaluated at x. These values can be obtained in R via `dnorm(x)` and `pnorm(x)`.

2. Standard half-Cauchy distribution

$$p_2(x) = \frac{2}{\pi(1+x^2)} \qquad \text{for } x \in \mathbb{R}^+$$

3. Standard log-normal distribution

$$p_3(x) = \frac{1}{x\sqrt{2\pi}} \exp\{-0.5(\log x)^2\} \qquad \text{for } x \in \mathbb{R}^+$$

4. Beta(2,5) distribution

$$p_4(x) = 30x(1-x)^4 \qquad \text{for } x \in (0,1)$$

Recalling the SIR algorithm to draw from a given distribution. If the goal is to obtain a random sample $x_1, \ldots, x_N$ from the target distribution $p(x)$, but you are only capable of producing a random sample from an auxiliary/candidate/importance distribution $q(x)$, then the SIR algorithm is a promising approach:

Step 1: Sampling from the candidate distribution:

$$\tilde{x}_1, \ldots, \tilde{x}_M \sim q(x).$$

Step 2: Computing unnormalized resampling weights:

$$\tilde{w}_i = \frac{p(x_i)}{q(x_i)}, \qquad i = 1, \ldots, M.$$

Notice that here you need to be able to evaluate $p(x)$ and $q(x)$, despite not knowing how to sample from $p(x)$.

Step 3: Normalize the weights:

$$w_i = \frac{\tilde{w}_i}{\sum_{j=1}^{M} \tilde{w}_j}, \qquad i = 1, \ldots, M.$$

We notice that, when normalizing the weights, we only need to evaluate $p(x)$ and $q(x)$ up to normalizing constants (See Problem 2 for an example).

Step 4: Resampling

$$\text{Sample } x_i \text{ from } \{\tilde{x}_1, \ldots, \tilde{x}_M\} \text{ with weights } \{w_1, \ldots, w_M\}, \qquad i = 1, \ldots, N.$$

Step 5: . For large M, N, the set $\{x_1, \ldots, x_N\}$ is a random sample from $p(x)$.

As an example, if $p(x)$ is the standard normal distribution and $q(x)$ is the standard Student's t with 4 degrees of freedom, here is the R code for the above SIR algorithm:

```
M = 100000
N = 10000
xtilde = rt(M,df=4)
w = dnorm(xtilde)/dt(xtilde,df=4)
x = sample(xtilde,size=N,replace=TRUE,prob=w)
par(mfrow=c(1,2))
hist(xtilde,prob=TRUE,xlab="draws",ylab="Density",main="Proposal",breaks=100)
curve(dt(x,df=4),from=-10,to=10,n=100,col=2,add=TRUE,lwd=2)
hist(x,prob=TRUE,xlab="draws",ylab="Density",main="Target",breaks=40)
curve(dnorm,from=-3,to=3,n=100,col=2,add=TRUE,lwd=2)
```

Your job is to obtain samples of size $N = 10000$ from $p(x)$ for each one of the four "new" distributions introduced above.

```
p1 = function(x){2*dnorm(x)*pnorm(3*x)}
p2 = function(x){2/(pi*(1+x^2))}
p3 = function(x){1/(x*sqrt(2*pi))*exp(-0.5*(log(x))^2)}
p4 = function(x){30*x*(1-x)^4}
par(mfrow=c(2,2))
curve(p1,from=-2,to=10,n=1000,ylab="Density")
curve(p2,from=0,to=10,n=1000,ylab="Density")
curve(p3,from=0,to=10,n=1000,ylab="Density")
curve(p4,from=0,to=1,n=1000,ylab="Density")
```

Use the following proposal distributions:

1. $q_1(x) \equiv N(0, 10)$ (notice that most values below zero will have virtually zero weight!)
2. $q_2(x) \equiv Gamma(1, 1) \equiv Exp(1)$
3. $q_3(x) \equiv Gamma(1, 1)$
4. $q_4(x) \equiv U(0, 1)$

Problem 2: Using SIR for Bayesian inference

For some data $x_1, \ldots, x_n$, model $p(x_1, \ldots, x_n|\theta)$ and prior $p(\theta)$, we have the following posterior for θ:

$$p(\theta|x_1, \ldots, x_n) \propto (1+\theta)^{125}(1-\theta)^{38}\theta^{34} \qquad \text{for} \quad \theta \in (0, 1).$$

Use SIR to obtain the following summaries:

1. $E(\theta|x_1, \ldots, x_n)$
2. $V(\theta|x_1, \ldots, x_n)$
3. $Pr(\theta < 0.6|x_1, \ldots, x_n)$
4. θ_L, where $Pr(\theta < \theta_L|x_1, \ldots, x_n) = 0.025$.
5. θ_U, where $Pr(\theta < \theta_U|x_1, \ldots, x_n) = 0.975$.

As $\theta \in (0, 1)$, the simplest proposal distribution would be the $q_1(\theta) \equiv U(0, 1)$. However, if you draw $p(\theta)$ you will notice that virtually all posterior density of θ lies in the interval $(0.4, 0.9)$, so a "*better*" proposal would be $q_2(\theta) \equiv U(0.4, 0.9)$. Compare both approximations when computing the above 5 summaries. In order to make your life easier, let us assume first that i) $M = 10,000$ and $N = 10,000$. Repeat everything with iii) $M = 100,000$, and then with iii) $M = 1,000,000$.

	$q_1 \equiv U(0,1)$			$q_2 \equiv U(0.35, 0.85)$		
	$M = 10^4$	$M = 10^5$	$M = 10^6$	$M = 10^4$	$M = 10^5$	$M = 10^6$
$E(\theta\|x_1, \ldots, x_n)$						
$V(\theta\|x_1, \ldots, x_n)$						
$Pr(\theta < 0.5\|x_1, \ldots, x_n)$						
θ_L						
θ_U						