STAT 4010 Bayesian Learning

Tutorial 8

Spring 2022

Cheuk Hin (Andy) CHENG (Email | Homepage)

Di SU (Email | Homepage)

1 Theoretical Justification

This section shows that the Bayesian methods studied in previous chapters are theoretically sensible.

Definition 1. *Given any DGP $f_\star(x)$ and model $\mathscr{F} = \{f(x \mid \theta) : \theta \in \Theta\}$. Denote the expectation and variance under the DPG $f_\star(x)$ by $\mathrm{E}_\star$ and $\mathrm{Var}_\star$. Define*

$$\theta_\star = \underset{\theta\in\Theta}{\arg\max} \mathrm{E}_\star \{\log f(x_1 \mid \theta)\},$$

and

$$I_\star = \left[\mathrm{Var}_\star\left\{\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(x_1 \mid \theta)\right\}\right]_{\theta=\theta_\star} \qquad J_\star = \left[-\mathrm{E}_\star\left\{\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(x_1 \mid \theta)\right\}\right]_{\theta=\theta_\star},$$

provided that the expectations exist. The quantities $I_\star$ and $J_\star$ are called Fisher information. If $\mathscr{F}$ well specifies $f_\star$, then $\theta_\star = \theta_0$ and $I_\star = J_\star$, where θ_0 is the true DGP parameter.

Theorem 1.1. *(Consistency of posterior). Assume regularity conditions (RCs). If n is large enough, then*

$$\widehat{\theta}_{\mathrm{MLE}} \approx \theta_\star \quad \text{and} \quad [\theta \mid x_{1:n}] \approx \theta_\star.$$

Theorem 1.2. *(Asymptotic distributions of posterior). Assume RCs. If n is large enough, then*

$$\widehat{\theta}_{\mathrm{MLE}} \approx \mathrm{N}\left(\theta_\star, \frac{J_\star^{-1}I_\star J_\star^{-1}}{n}\right) \quad \text{and} \quad [\theta \mid x_{1:n}] \approx \mathrm{N}\left(\widehat{\theta}_{\mathrm{MLE}}, \frac{J_\star^{-1}}{n}\right).$$

If the model is well-specified, the precision of Bayesian framework and frequentist framework are consistent.

Theorem 1.3. *(Asymptotic representation of posterior mean). Assume RCs. If n is large enough, then*

$$\mathrm{E}(\theta \mid x_{1:n}) \approx \widehat{\theta}_{\mathrm{MLE}}.$$

Remark 1.1. Some remark on the sign " $\approx$ ".

- We have different modes of convergence for random variables (rvs). Let A_n and B be two rvs. Consider when n goes to infinity.

  1. (Convergence in distribution) $A_n \xrightarrow{\mathrm{d}} B \Leftrightarrow F_{A_n} \to F_B$ for all continuity points of F_B, where F is the cdf.
  2. (Convergence in probability) $A_n \xrightarrow{\mathrm{pr}} B \Leftrightarrow \Pr(|A_n - B| > \epsilon) \to 0$ for some $\epsilon > 0$.
  3. (Convergence in L^p) $A_n \xrightarrow{L^p} B \Leftrightarrow (\mathsf{E}A_n^p)^{1/p} \to (\mathsf{E}B^p)^{1/p}$.

4. (Convergence almost surely/with probability one) $A_n \overset{a.s.}{\to} B \Leftrightarrow$ for any $\omega \in \Omega$ the Sigma-field, $\Pr(\lim_{n\to\infty} A_n(\omega) \to B(\omega)) = 1$.

- Strength of the mode of convergences is different. We have $\overset{L^p}{\to}, \overset{a.s.}{\to} \Rightarrow \overset{\text{pr}}{\to} \Rightarrow \overset{\text{d}}{\to}$ for $p \geq 1$. However $\overset{L^p}{\to}$ and $\overset{a.s.}{\to}$ do not imply each other.
- For Theorem 1.1, $\widehat{\theta}_{\text{MLE}} \overset{\text{pr}}{\to} \theta_\star$ and $\theta \overset{\text{pr}}{\to} \theta^*$ (given x).
- Let $Z \sim \text{N}(0,1)$. Theorem 1.2 means that $\widehat{\theta}_{\text{MLE}} - \theta_\star - \frac{J_\star^{-1} I_\star J_\star^{-1}}{n} Z \overset{\text{d}}{\to} 0$ and $\theta - \widehat{\theta}_{\text{MLE}} - \frac{J_\star^{-1}}{n} Z \overset{\text{d}}{\to} 0$ (given x).

Theorem 1.4. *We have the following bi-directional relation*

$$\begin{aligned} & x_{1:n} \textit{ are exchangeable with joint density } f(x_{1:n}) \\ \iff \quad & \exists \theta \in \Theta, f(x \mid \theta), \pi(\theta) \textit{ s.t. } \begin{cases} [x_{1:n} \mid \theta] \overset{IID}{\sim} f(x_{1:n} \mid \theta) \\ \theta \sim \pi(\theta). \end{cases} \end{aligned}$$

The direction "$\Longrightarrow$" is stated in theorem 6.5. De Finiti Theorem, and the direction "$\Longleftarrow$" is given in proposition 6.4.

Example 1.1. Consider the true DGP, $x_{1:n} \overset{\text{IID}}{\sim} Ga(a)/b$ where $a = 4$ and $b = 2$. We consider the model, $x_{1:n} \overset{\text{IID}}{\sim} \theta\text{Exp}(1)$ where $\theta > 0$.

1. Compute the MLE. Discuss its asymptotic behaviour.
2. Propose a prior and compute its posterior. Discuss its asymptotic behaviour.
3. Produce a plot of the exact and asymptotic distributions of the MLE and the posterior.

Solution:

1. Let $S_n = \sum_{i=1}^n x_i$. We can compute directly,

$$\begin{aligned} f(x_{1:n}, \theta) &= \frac{1}{\theta^n} e^{-S_n/\theta}, \\ \ell_{1:n}(\theta) &:= \log f(x_{1:n}, \theta) = -n \log \theta - \frac{S_n}{\theta}, \\ \ell'_{1:n}(\theta) &:= \frac{\partial \log f(x_{1:n}, \theta)}{\partial \theta} = \frac{-n}{\theta} + \frac{S_n}{\theta^2} = 0, \\ \ell''_{1:n}(\theta) &= \frac{n}{\theta^2} - \frac{2S_n}{\theta^3}. \end{aligned}$$

By setting $\ell'_{1:n}(\theta) = 0$, we can see that $\widehat{\theta}_{MLE} = S_n/n$ and $\ell''_{1:n}(\widehat{\theta}_{MLE}) < 0$. Next, we want to compute $\theta_\star$, $I_\star$ and $J_\star$. For simplicity, let $\ell_{1:1}(\theta) = \ell(\theta)$. Firstly,

$$\theta_\star = \arg\max_\theta \mathsf{E}_\star \ell(\theta) = \arg\max_\theta \left[-log(\theta) - \frac{\mathsf{E}_\star x_1}{\theta} \right] = \left[-log(\theta) - \frac{2}{\theta} \right].$$

Similarly to the derivation of the MLE (take $n=1$ and $S_n=2$), we have $\theta_\star=2$. Next by some computation,

$$I_\star = [\mathsf{Var}_\star \ell'(\theta)]_{\theta=\theta_\star} = \left[\mathsf{Var}_\star\left(-\frac{1}{\theta}+\frac{x_1}{\theta^2}\right)\right]_{\theta=\theta_\star} = \frac{1}{\theta_\star^4}\mathsf{Var}_\star(x_1) = \frac{a}{b^2\theta_\star^4} = \frac{1}{2^4},$$

$$J_\star = [-\mathsf{E}\ell''(\theta)]_{\theta=\theta_\star} = \left[-\mathsf{E}\left(\frac{1}{\theta^2}-\frac{2x_1}{\theta^3}\right)\right]_{\theta=\theta_\star} = -\frac{1}{\theta_\star^2}+\frac{2\mathsf{E}_\star x_1}{\theta_\star^3} = -\frac{1}{\theta_\star^2}+\frac{2a}{b\theta_\star^3} = \frac{1}{4}.$$

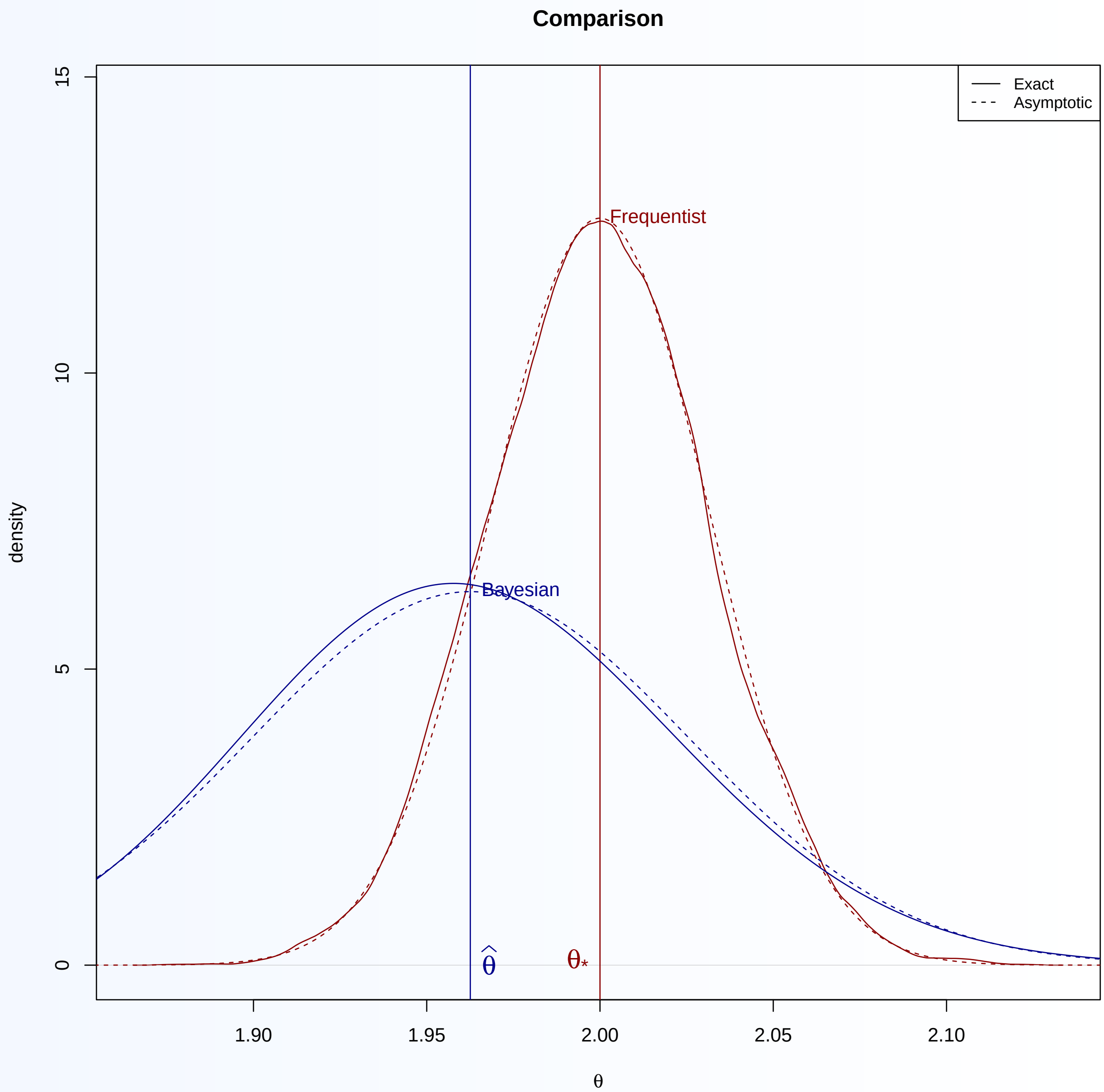
Note that $J_\star^{-1}I_\star J_\star^{-1}=1$. By theorem 1.1 and 1.2, we have

$$\widehat{\theta}_{MLE} \approx \theta_\star = 2 \quad \text{and} \quad [\widehat{\theta}_{MLE}] \approx \mathrm{N}(2, 1/n).$$

2. Consider the conjugate prior for θ, $\theta \sim k/\mathrm{Ga}(h)$. The posterior is $\theta \mid x_{1:n} \sim k_n/\mathrm{Ga}(h_n)$ where $h_n = h+n$ and $k_n = k+S_n$. By theorem 1.2, we have

$$[\theta \mid x_{1:n}] \approx N(\widehat{\theta}_{MLE}, 4/n).$$

3. Set $h=2$ and $k=1$. We have the following plot.



```
##Truth
a = 4
b = 2

##Frequentist MLE
theta0 = a/b
I = a/(b^2*theta0^4)
J = -1/theta0^2+2*a/(b*theta0^3)
varF = I/J^2

##Plot setting
set.seed(100)
par(mfrow=c(1,1), mar=c(4.5,5,3,2))
col = c("red4","blue4")
lty = c(1,2)
n = 1000
nRep = 2^12
theta = seq(1, 3, length.out=2000) ##grid of theta for the density plot

# Frequentist
##Step 2: simulate the exact distribution for the MLE
out = rep(NA, nRep)
for(iRep in 1:nRep){
  x = rgamma(n,a,b) ##Simulate data from the DGP
  out[iRep] = mean(x) ##theta_MLE
}
deF = density(out, kernel="epanechnikov")
##Step 1: Compute the asymp. distribution of the MLE
daF = dnorm(theta, theta0, sqrt(varF/n)) #asymptotic


# Bayesian model
h = 2
k = 1
post = function(theta,x,h,k){
  hn = a+n
  kn = b+sum(x)
  logd = (-hn-1)*log(theta)-kn/theta
  d = exp(logd-max(logd))
  d/sum(d)/(theta[2]-theta[1])
}

##theory
set.seed(4010)
x = rgamma(n,h,k) #fix a realization using DGP for the posterior
##Step 3 compute the exact posterior distribution
deB = post(theta,x,alpha,beta)
##Step 4 compute the asymp. posterior distribution
theta_mle = mean(x)
varB = 1/J
daB = dnorm(theta, theta_mle, sqrt(varB/n))

##Plot
plot(deF, type="l",col=col[1], lty=lty[1],
     main="Comparison", ylab="density", xlab=bquote(theta),
     ylim=c(0,max(daF)+2))
points(theta, daF, type="l", col=col[1], lty=lty[2])
legend("topright", c("Exact","Asymptotic"), col="black", lty=lty, cex=.8)
```

```
text(theta[which.max(daF)], max(daF), "Frequentist", pos=4, col=col[1])
abline(v=theta0, col=col[1])
text(theta0, 0, expression(theta["*"]), pos=2, col=col[1],cex=1.4)
points(theta, deB, type="l", col=col[2], lty=lty[1])
points(theta, daB, type="l", col=col[2], lty=lty[2])
text(theta[which.max(daB)], max(daB), "Bayesian", pos=4, col=col[2])
abline(v=theta_mle, col=col[2])
text(theta_mle, 0, expression(widehat(theta)), pos=4, col=col[2],cex=1.4)
```

2 Posterior Computation

We are interested in following tasks.

1. Draw sample $\theta_1, \dots, \theta_J \sim \pi(\theta)$.
2. Compute the integral $\mathsf{E}_\pi g(\theta) = \int_\Theta g(\theta)\pi(\theta)\mathrm{d}\theta = \frac{\int_\Theta g(\theta)\pi_u(\theta)\mathrm{d}\theta}{\int_\Theta \pi_u(\theta)\mathrm{d}\theta}$, where $\pi_u(\theta)$ is the unnormalized density.

2.1 Classic Methods

Algorithm 1: Trapezoidal rule.

Input: (i) knot number J; (ii) bound a, b; (iii) unnormalized target density $\pi_u(\cdot)$; and (iv) function $g(\cdot)$.

begin

(1) Compute the grid points $\theta_j = a + hj$ for $j = 0, \dots, J$ and $h = (b-a)/J$.

(2) Compute $\hat{I}_{Trap} = \hat{U}_{Trap}/\hat{L}_{Trap}$, where

$$\begin{aligned}\hat{U}_{Trap} &:= \sum_{j=1}^{J} \frac{G(\theta_j) + G(\theta_{j-1})}{2} h, \\ \hat{L}_{Trap} &:= \sum_{j=1}^{J} \frac{\pi_u(\theta_j) + \pi_u(\theta_{j-1})}{2} h, \\ G(\theta) &:= g(\theta)\pi_u(\theta).\end{aligned}$$

end

Output: $\hat{I}_{Trap}$

Algorithm 2: Inverse Probability transform.

Input: Inverse function of the CDF, i.e., $F^{-1}(\cdot)$.

begin

(1) Generate $U \sim \mathrm{Unif}(0,1)$. (2) Compute $\theta = F^{-1}(U)$.

end

Output: θ

Remark 2.1. In practice, the bound $[a,b]$ can be infinite. Suppose $\hat{I}_{Trap}$ is monotone with respect to the width of the interval. We can try Trapezoidal rule several time by enlarging the range (at the same time increase J as well) until the absolute change in $\hat{I}_{Trap}$ is less than certain tolerance level.

Theorems below justify the use of the trapezoidal rule and the inverse probability transformation.

Theorem 2.1. *(Justification of trapezoidal rule) Assume $\Theta = [a,b]$ is a bounded interval. If $g(\cdot)$ is twice differentiable on $[a,b]$, then as $J \to \infty$*

$$\hat{I}_{Trap} - I = O\left(\frac{1}{J^2}\right).$$

Theorem 2.2. *(Justification of inverse probability transform) Let $U \sim Unif(0,1)$ and $F(\cdot)$ be the CDF of θ. Assume the inverse function of CDF exists. Then,*

$$\Pr(F^{-1}(U) < c) = \Pr(F(F^{-1}(U)) < F(c)) = \Pr(U < F(c)) = F(c).$$

That is θ and $F^{-1}(U)$ has the same CDF. Thus, they have the same distribution.

Example 2.1. Consider $\theta \sim F(\theta)$ and $f(\theta) \propto \exp\{-|\theta|/3\}\mathbb{1}(\theta \in \mathbb{R})$. Simulate $\mathsf{E}\theta^2$ using Trapezoidal rule and Inverse Probability transform.

Solution: Note that $\theta \sim$ Laplace(3) and $\mathsf{E}\theta^2 = 18$. For Trapezoidal rule,

```
target_den <- function(theta,b=3){
  log_d = -abs(theta)/b
  exp(log_d - max(log_d))
}

target_g <- function(theta){
  theta^2
}

##Trapezoidal rule
trap <- function(J,a,b){
  theta_grid = seq(a,b,length.out = J)
  h = (b-a+1)/J
  pi_u = target_den(theta_grid)
  G = target_g(theta_grid)*pi_u
  L = sum((pi_u[2:J] + pi_u[1:(J-1)])/2*h)
  U = sum((G[2:J] + G[1:(J-1)])/2*h)
  U/L
}

trap(2^10,-10,10)
[1] 12.08103
trap(2^10,-20,20)
[1] 17.33751
trap(2^10,-40,40)
[1] 17.99753
trap(2^10,-80,80)
[1] 18.00204
```

Note that,

$$F(\theta) = \begin{cases} \frac{1}{2}e^{\theta/3}, & \theta \leq 0; \\ 1 - \frac{1}{2}e^{-\theta/3}, & \theta > 0. \end{cases}$$

Therefore,

$$F^{-1}(U) = \begin{cases} 3\log(2U), & U \leq 1/2; \\ -3\log(2\{1-U\}), & U > 1/2. \end{cases}$$

For Inverse Probability transform,

```
##Inverse Probability transform
inv_cdf <- function(U,b=3){
  b*log(2*U)*(U <=0.5) + -b*log(2*(1-U))*(U>0.5)
}

set.seed(100)
nRep = 2^14
theta_sim = inv_cdf(runif(nRep))
mean(theta_sim^2)
[1] 17.88102
```

3 Basic Monte Carlo/Importance Sampling

In importance sampling, we slightly modified our target,

$$I = \int_\Theta g(\theta)\pi_u(\theta)\mathrm{d}\theta = \int_\Theta g(\theta)\frac{\pi_u(\theta)}{p(\theta)}p(\theta)\mathrm{d}\theta = \int_\Theta g(\theta)w(\theta)p(\theta)\mathrm{d}\theta,$$

where $w(\theta) = \pi_u(\theta)/p(\theta)$. Note, the above equation is valid when $p(\theta) = 0$ implies $\pi_u(\theta) = 0$ or $\pi(\cdot)$ is absolutely continuous with respect to $p(\cdot)$.

Algorithm 3: Importance Sampling.

Input: (i) simulation size J; (ii) proposed PDF $p(\cdot)$; (iii) (unnormalized) target density $\pi_u(\cdot)$.

begin

(1) Generate $\tilde{\theta}_1, \ldots, \tilde{\theta}_J \overset{\text{IID}}{\sim} p(\cdot)$.

(2) Compute $w_j = \pi_u(\tilde{\theta}_j)/p(\tilde{\theta}_j)$, for $j = 1, \ldots, J$.

(3) Compute $\hat{I}_{IS} = \hat{U}_{IS}/\hat{L}_{IS}$, where $\hat{U}_{IS} = J^{-1}\sum_{j=1}^J g(\tilde{\theta}_j)w_j$ and $\hat{L}_{IS} = J^{-1}\sum_{j=1}^J w_j$.

end

Output: $\hat{I}_{IS}$

Theorem 3.1. *(Justification of the importance sampling) If* $\mathsf{Var}_p(g(\theta)w(\theta)) < \infty$ *and* $\mathsf{Var}_p(w(\theta)) < \infty$*, then as* $J \to \infty$*,*

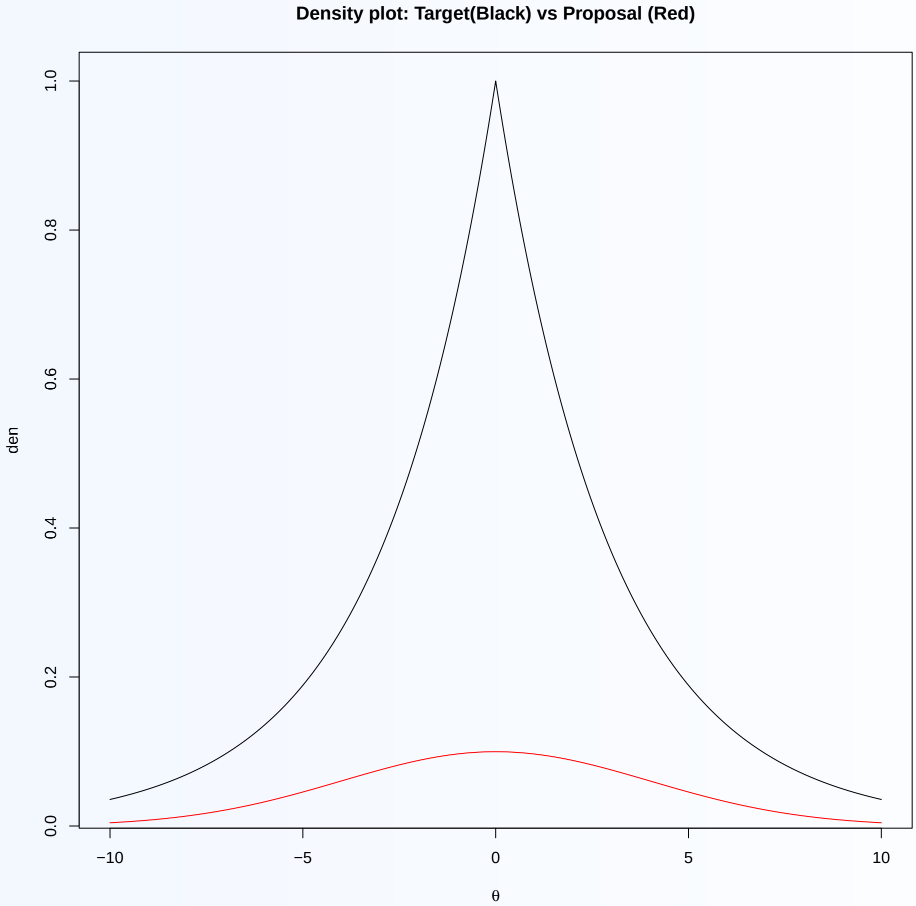
$$\sqrt{J}\left(\hat{I}_{IS} - I\right) \xrightarrow{\mathrm{d}} N(0, \sigma^2_{IS}) \quad \textit{and} \quad \sigma^2_{IS} = \int_\Theta \{g(\theta) - I\}\frac{\pi^2(\theta)}{p(\theta)}\mathrm{d}\theta.$$

Remark 3.1. For the choice of $p(\cdot)$, there is one restriction and two criteria:

1. The support of $p(\cdot)$ covers $\pi(\cdot)$.
2. The shape of $p(\cdot)$ is similar to $\pi(\cdot)$ or $\pi_u(\cdot)$.
3. It is easy to draw samples from $p(\cdot)$.

Example 3.1. Continue from example 2.1. Use importance sampling to simulate $\mathsf{E}\theta^2$.

Solution: The density plot suggests that we should use $N(0, \sigma^2)$ as the proposal.




```r
##Step 1 visualize the target density and proposal density
theta_grid = seq(-10,10,length.out = 2^10+1)
tar_den = target_den(theta_grid)
plot(theta_grid,tar_den,type = 'l')
proposal_den = dnorm(theta_grid,0,4)
points(theta_grid,proposal_den,type = 'l',col='red')

##step 2
importance_sampling <- function(J,sd = 4,b=3){
  tilde_theta = rnorm(J,0,sd)
  w = target_den(tilde_theta,b)/dnorm(tilde_theta,0,sd)
  U = sum(tilde_theta^2*w)/J
  L = sum(w)/J
  U/L
```

```
}

set.seed(4010)
importance_sampling(2^14,sd = 2)
[1] 10.6472
importance_sampling(2^14,sd = 3)
[1] 12.942
importance_sampling(2^14,sd = 4)
[1] 17.31651
importance_sampling(2^14,sd = 5)
[1] 17.0142
importance_sampling(2^14,sd = 6)
[1] 17.75909
importance_sampling(2^14,sd = 7)
[1] 18.26192
importance_sampling(2^14,sd = 8)
[1] 18.03691
```