MATHEMATICS & STATISTICS
THE UNIVERSITY OF NEW SOUTH WALES

Lecture 4: Inference, Asymptotics & Monte Carlo

August 11, 2018

Outline

1. Posterior Inference
    - Loss functions, predictive inference
2. Posterior Asymptotics
3. Monte Carlo methods
    - Importance sampling.

Outline

1. Posterior Inference
   - Loss functions, predictive inference
2. Posterior Asymptotics
3. Monte Carlo methods
   - Importance sampling.

Summarising Posterior Information

The posterior $\pi(\theta \,|\, x)$ is a complete summary of the inference about θ. In some sense $\pi(\theta \,|\, x)$ *is* the inference.

However, for many applications we wish to summarise this information to make a decision.

1. What is the "best" point estimate $\hat{\theta}$ of θ?
   E.g. Posterior mean, median, mode etc.

2. What decision $d \in \{d_1, d_2, \ldots\} = \mathcal{D}$ is the optimal choice to make, given knowledge of θ? E.g.
   - How many loaves of bread to bake per day to maximise profit?
     – balancing baking costs and number of loaves sold/wasted.
   - How high to build sea walls to minimise cost?
     – balancing construction cost, chance of breach and resulting damages.
   - How much local rail infrastructure to build?
     – balancing construction costs versus improvement to economy and other benefits.

Loss Functions

How do we define “best” or optimal?

Loss Function

For a prediction $d \in \mathcal{D}$, a loss function

$$\ell(\theta, d)$$

defines the penalty in taking decision d given (fixed) parameter value θ.

- Negative loss is a gain, and is beneficial.
- Sometimes expressed as maximising utility $-\ell(\theta, d)$.

The premise is then:

- Choose the decision $d^* = \operatorname{argmin}_d \ell(\theta, d)$ that minimises the loss.

However, θ is not known, but rather $\theta \sim \pi(\theta \,|\, x)$. So alternatively

$$d^* = \underset{d}{\operatorname{argmin}} \, \mathbb{E}_\pi \left[\ell(\theta, d)\right].$$

i.e. choose d which minimises the expected posterior loss.

Loss functions

A full (decision theoretic) Bayesian setup consists of specifying:

- Prior distribution $\pi(\theta)$
- Model $f(x \mid \theta)$ (leading to the likelihood $L(x \mid \theta)$)
- Loss function $\ell(\theta, d)$.

Although most people only consider the first two of these.

Loss functions for estimating θ

We first consider loss functions for parameter estimation:

Given $\pi(\theta \,|\, x)$ what is the optimal point estimate of θ? $(d = \hat{\theta})$

Consider 4 standard loss functions:

- Quadratic loss:
$$\ell(\theta, d) = (\theta - d)^2$$
- Absolute error loss:
$$\ell(\theta, d) = |\theta - d|$$
- 0-1 loss:
$$\ell(\theta, d) = \begin{cases} 0 & \text{if } |d - \theta| \leq \epsilon \\ 1 & \text{if } |d - \theta| > \epsilon \end{cases}$$
- Linear loss:
$$\ell(\theta, d) = \begin{cases} \alpha(d - \theta) & \text{if } d > \theta \\ \beta(\theta - d) & \text{if } d < \theta \end{cases}$$
for given $\alpha, \beta > 0$.

Quadratic Loss

$$\begin{aligned}
\mathbb{E}_\pi[\ell(\theta,d)] &= \int \ell(\theta,d)\pi(\theta\,|\,x)\mathrm{d}\theta \\
&= \int (\theta-d)^2\pi(\theta\,|\,x)\mathrm{d}\theta \\
&= \int (\theta-\mathbb{E}(\theta\,|\,x)+\mathbb{E}(\theta\,|\,x)-d)^2\pi(\theta\,|\,x)\mathrm{d}\theta \\
&= \int (\theta-\mathbb{E}(\theta\,|\,x))^2\pi(\theta\,|\,x)\mathrm{d}\theta + \int (\mathbb{E}(\theta\,|\,x)-d)^2\pi(\theta\,|\,x)\mathrm{d}\theta \\
&\qquad +2\int (\theta-\mathbb{E}(\theta\,|\,x))(\mathbb{E}(\theta\,|\,x)-d)\pi(\theta\,|\,x)\mathrm{d}\theta \\
&= \mathbb{V}\mathrm{ar}(\theta\,|\,x) + (\mathbb{E}(\theta\,|\,x)-d)^2 + 0
\end{aligned}$$

- This is minimised when $d=\mathbb{E}(\theta\,|\,x)$.
- The posterior mean minimises quadratic loss.
- The expected loss is the posterior variance.

Linear Loss

Linear loss:

$$\ell(\theta, d) = \begin{cases} \alpha(d-\theta) & \text{if } d > \theta \\ \beta(\theta - d) & \text{if } d < \theta \end{cases}$$

for given $\alpha, \beta > 0$.

For any d, we have $\mathbb{E}\ell(X, d) =$

$$\begin{aligned} &= \alpha\mathbb{E}(d-X)^+ + \beta\mathbb{E}(X-d)^+ \\ &= \alpha\mathbb{E}[d-X; X<d] + \beta\mathbb{E}[X-d; X>d] \\ &= d(\alpha\mathbb{P}[X<d] - \beta\mathbb{P}[X>d]) + \beta\mathbb{E}[X; X>d] - \alpha\mathbb{E}[X; X<d] \\ &= d((\alpha+\beta)\mathbb{P}[X<d] - \beta) - (\alpha+\beta)\mathbb{E}[X; X<d] + \beta\mathbb{E}X \end{aligned}$$

Let d^* be the $\beta/(\alpha+\beta)$ quantile, that is,

$$\mathbb{P}[X < d^*] = \beta/(\alpha+\beta)\ .$$

Linear Loss

Then, for any other d, we have $\mathbb{E}\ell(X,d) - \mathbb{E}\ell(X,d^*) =$

$$= d((\alpha+\beta)\mathbb{P}[X<d]-\beta) - (\alpha+\beta)(\mathbb{E}[X;X<d]-\mathbb{E}[X;X<d^*])$$
$$= (\alpha+\beta)\Big\{d(\mathbb{P}[X<d]-\mathbb{P}[X<d^*]) + \mathbb{E}[X;X<d^*] - \mathbb{E}[X;X<d]\Big\}$$

Hence, if $d^* > d$, then

$$\frac{\mathbb{E}\ell(X,d)-\mathbb{E}\ell(X,d^*)}{(\alpha+\beta)} = \mathbb{E}[X;d<X<d^*] - d\mathbb{P}[d<X<d^*]$$
$$= \mathbb{P}[d<X<d^*]\underbrace{(\mathbb{E}[X\,|\,d<X<d^*]-d)}_{\geq 0}$$

The case for $d^* < d$ is dealt with similarly, so we obtain the following.

So linear loss is minimised at the $\frac{\beta}{\alpha+\beta}$ posterior quantile.

Absolute Error Loss

Absolute error loss:

$$\ell(\theta, d) = |\theta - d|$$

Absolute error loss = Linear loss with $\alpha = \beta = 1$.

When $\alpha = \beta = 1$, d^* is the median of the posterior distribution.

$\Rightarrow$ The posterior median minimises absolute error loss.

0-1 Loss

0-1 loss:

$$\ell(\theta, d) = \begin{cases} 0 & \text{if } |d-\theta| \le \epsilon \\ 1 & \text{if } |d-\theta| > \epsilon \end{cases}$$

Here

$$\begin{aligned} \mathbb{E}[\ell(\theta, d)] &= \mathbb{P}(|\theta - d| > \epsilon) \\ &= 1 - \mathbb{P}(|\theta - d| \le \epsilon). \end{aligned}$$

- $|\theta - d| \le \epsilon$ defines an interval $[\theta - \epsilon, \theta + \epsilon]$ of length 2ϵ
- To minimise, choose interval with highest probability i.e. high density region
- Then θ is mid-point of interval with highest probability
- Choosing ϵ arbitrarily small will select the posterior mode for d.

$\Rightarrow$ The posterior mode minimises 0-1 loss.

Loss functions for making other decisions

Example: Baking loaves of bread

- $c =$ the cost of baking a loaf of bread
- $s > c$ is price loaf sells for
- $\pi(d \,|\, x)$ is the (posterior) distribution of demand for bread
- $b \in \{0, 1, \ldots\}$ is the decision (i.e. number of loaves to bake).

Want to maximise expected profit.

Set up (obvious) loss function:

- If baker bakes b loaves with demand d then profit is:

$$\text{Profit} = \begin{cases} (s-c)d - c(b-d) & \text{for } b > d \\ (s-c)d & \text{for } b = d \\ (s-c)b & \text{for } b < d \end{cases}$$

Loss functions for making other decisions

- Make Profit relative to selling $b = d$ loaves
  (i.e. remove $(s-c)d$ from all terms)

$$\text{Profit}' = \begin{cases} -c(b-d) & \text{for } b > d \\ 0 & \text{for } b = d \\ (s-c)b - (s-c)d = (s-c)(b-d) & \text{for } b < d \end{cases}$$

- Loss = -Profit

$$\ell(b,d) = \begin{cases} c(b-d) & \text{for } b \geq d \quad \text{(cost of baking surplus loaves)} \\ (s-c)(d-b) & \text{for } b < d \quad \text{(lost profit from not baking enough loaves)} \end{cases}$$

- This is linear loss with $\alpha = c$ and $\beta = s - c$
- Therefore the optimal decision minimising expected posterior loss is the $\frac{\beta}{\alpha+\beta} = \frac{s-c}{s}$ quantile of $\pi(d \,|\, x)$.

Outline

1. Posterior Inference
    - Loss functions, predictive inference
2. Posterior Asymptotics
3. Monte Carlo methods
    - Importance sampling.

Predictive Inference

Previous focus on parameter estimation (e.g via loss functions).

Common interest in predictions about future observations.

In predictions there are two forms of uncertainty:

1. Uncertainty over the parameter values, which have been estimated based on data x.
2. Uncertainty due to the fact that any future value is itself a random event.

In classical statistics, typically predict with $f(y \mid \hat{\theta})$

- $\hat{\theta}$ is the MLE - fixed.
- Only accounts for second source of uncertainty
- So predictions are more precise than they should be
- Problem stems from classical assumption of a single true value of θ

Predictive Inference

Bayesian framework allows for both sources of uncertainty by averaging over uncertainty of parameter estimates.

The predictive density function of a future observation, Y, is

$$f(y \mid x) = \int f(y \mid \theta, x)\pi(\theta \mid x)\mathrm{d}\theta.$$

While simple to express, can sometimes be difficult to compute.

Can use standard conjugate families to give tractable forms for predictive distribution in some cases.

Is also simple to estimate via Monte Carlo methods.

Predictive Inference

Example: Binomial model.

Suppose we have $X \sim \text{Binomial}(n, \theta)$ with conjugate prior $\theta \sim \text{Beta}(a, b)$. Then we know that

$$\theta \,|\, X = x \sim \text{Beta}(a + x, b + n - x).$$

Now suppose we intend to make N further observations.
Let Y be the number of successes, so $Y \,|\, \theta \sim \text{Binomial}(N, \theta)$. Hence

$$f(y \,|\, \theta) = \binom{N}{y} \theta^y (1 - \theta)^{N-y}.$$

So for $y = 0, 1, \ldots, N$

$$\begin{aligned} f(y \,|\, x) &= \int_0^1 \binom{N}{y} \theta^y (1 - \theta)^{N-y} \times \frac{\theta^{a+x-1} (1 - \theta)^{b+n-x-1}}{\mathrm{B}(a + x, b + n - x)} \mathrm{d}\theta \\ &= \binom{N}{y} \frac{\mathrm{B}(y + a + x, N - y + b + n - x)}{\mathrm{B}(a + x, b + n - x)}. \end{aligned}$$

This is known as the Beta-Binomial distribution.

Predictive Inference

Example: Poisson model

We have $X_1, \dots, X_n \sim \text{Poisson}(\theta)$ with conjugate $\theta \sim \text{Gamma}(a, b)$.
Then we know that

$$\theta \,|\, X = x \sim \text{Gamma}(a + n\bar{x}_n, b + n).$$

Now $Y \sim \text{Poisson}(\theta)$ so that

$$f(y \,|\, \theta) = \exp(-\theta)\frac{\theta^y}{y!}.$$

Hence the predictive distribution is

$$\begin{aligned} f(y \,|\, x) &= \int \exp(-\theta)\frac{\theta^y}{y!}\frac{(b+n)^{a+n\bar{x}_n}}{\Gamma(a+n\bar{x}_n)}\theta^{a+n\bar{x}_n-1}\exp(-(b+n)\theta)\mathrm{d}\theta \\ \text{(Tute 1)} &= \binom{y+a+n\bar{x}_n-1}{y}\left(\frac{1}{b+n+1}\right)^y\left(1-\frac{1}{b+n+1}\right)^{a+n\bar{x}_n}. \end{aligned}$$

which is the pdf of a $\text{NegBin}(a + n\bar{x}_n, 1/(b + n + 1))$ distribution.

Monte Carlo posterior predictive distributions

How to generate samples from $f(\boldsymbol{y} \mid \boldsymbol{x})$?

$$f(\boldsymbol{y} \mid \boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{x}) \pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \mathrm{d}\boldsymbol{\theta}$$

or in machine learning notation with training data τ:

How to generate samples from $f(\boldsymbol{x} \mid \tau)$?

$$f(\boldsymbol{x} \mid \tau) = \int_{\Theta} f(\boldsymbol{x} \mid \boldsymbol{\theta}, \tau) \pi(\boldsymbol{\theta} \mid \tau) \mathrm{d}\boldsymbol{\theta}$$

By inspection of formula:

- Obtain posterior samples $\boldsymbol{\theta}_i \sim \pi(\theta \mid \tau)$
- For each $\boldsymbol{\theta}_i$ we can generate $\boldsymbol{X}_i \sim f(\boldsymbol{x} \mid \boldsymbol{\theta}_i)$
- This gives us joint samples $(\boldsymbol{\theta}_i, \boldsymbol{X}_i) \mid x \sim f(\boldsymbol{x} \mid \theta, \tau) \pi(\boldsymbol{\theta} \mid \tau)$
- To obtain samples from $f(\boldsymbol{y} \mid \boldsymbol{x})$, "integrate out" θ (i.e. discard the $\boldsymbol{\theta}_i$ values) to leave $\boldsymbol{X}_i \sim f(\boldsymbol{y} \mid \boldsymbol{x})$ only

Hugely simpler than calculating exact algebraic expression!

Outline

1. Posterior Inference
    - Loss functions, predictive inference
2. Posterior Asymptotics
3. Monte Carlo methods
    - Importance sampling.

Posterior Asymptotics

What happens to the posterior distribution as $n \to \infty$?

1. Consistency:

   If the "true" value of $\theta = \theta_0$,
   and if $\pi(\theta_0) \neq 0$ (or is non-zero in a neighbourhood of θ_0),
   then with increasing amounts of data ($n \longrightarrow \infty$) the posterior probability that θ equals (or lies in a neighbourhood of) $\theta_0 \longrightarrow 1$.

   Property akin to classical notion of 'consistency'.

2. Asymptotic Normality for $\boldsymbol{\theta} \in \mathbb{R}^d$:

   As $n \longrightarrow \infty$ then

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \longrightarrow \mathsf{N}(\boldsymbol{\theta}_0, [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}/n).$$

All the following arguments are heuristic/outline proofs.

Posterior Asymptotics

1. Consistency:
Let $X_1, \ldots, X_n \sim f(x \mid \theta_0)$ be *iid* observations and suppose the prior is such that $\pi(\theta_0) \neq 0$.

Then the posterior is

$$\begin{aligned}
\pi(\theta \mid \boldsymbol{X}_n) &\propto \pi(\theta) \prod_{i=1}^{n} f(X_i \mid \theta) \\
&= \pi(\theta) \exp(\textstyle\sum_{i=1}^{n} \ln f(X_i \mid \theta)) \\
&= \pi(\theta) \exp(-n\mathcal{D}_n(\theta)) \displaystyle\prod_{i=1}^{n} f(X_i \mid \theta_0)
\end{aligned}$$

where $\mathcal{D}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln \frac{f(X_i \mid \theta_0)}{f(X_i \mid \theta)}$.

Posterior Asymptotics

For fixed θ, $\mathcal{D}_n(\theta)$ is the average of n *iid* random variables, so converges in probability to its expectation (law of large numbers)

$$\mathbb{E}[\mathcal{D}_n(\theta)] = \int f(x \,|\, \theta_0) \ln \left[\frac{f(x \,|\, \theta_0)}{f(x \,|\, \theta)} \right] \mathrm{d}x := \mathcal{D}(\theta) \; .$$

The RHS is the Kullback-Leibler distance between $f(x \,|\, \theta_0)$ and $f(x \,|\, \theta)$. This distance, $\mathcal{D}(\theta) \geq 0$ for all θ with equality if and only if $f(x \,|\, \theta_0) \equiv f(x \,|\, \theta)$, which here we assume is the same as $\theta_0 = \theta$ (this assumption is called the identifiability assumption). Hence, for $\theta \neq \theta_0$, we obtain $\mathcal{D}(\theta) > 0$ and hence $\exp(-n\mathcal{D}(\theta)) \longrightarrow 0$ as $n \uparrow \infty$. In other words,

$$\exp(-n\mathcal{D}_n(\theta)) \xrightarrow{\mathbb{P}} \begin{cases} 0, & \theta \neq \theta_0 \\ 1, & \theta = \theta_0 \end{cases}, \quad n \uparrow \infty \; .$$

Therefore, the posterior spikes at θ_0:

$$\pi(\theta \,|\, \boldsymbol{X}_n) \xrightarrow{\mathbb{P}} \begin{cases} 0, & \theta \neq \theta_0 \\ \pi(\theta_0 \,|\, \boldsymbol{X}_n), & \theta = \theta_0 \end{cases}, \quad n \uparrow \infty \; .$$

Hence, the posterior mode, say $\hat{\theta}_n$, converges in probability to θ_0.

Posterior Asymptotics (Section 8.4 of Kroese & Chan)

Recall: Taylor series

For a function $f(x)$ that is infinitely differentiable at a, then

$$f(x) = f(a) + \frac{f'(a)(x-a)}{1!} + \frac{f''(a)(x-a)^2}{2!} + \frac{f'''(a)(x-a)^3}{3!} + \dots$$

2. Asymptotic Normality: (univariate and continuous θ). Taylor expanding $\mathcal{D}_n(\theta)$ around θ_0 yields:

$$\mathcal{D}_n(\theta) = \mathcal{D}_n(\theta_0) + (\theta-\theta_0)\tfrac{\mathrm{d}\mathcal{D}_n}{\mathrm{d}\theta}(\theta_0) + \tfrac{(\theta-\theta_0)^2}{2}\tfrac{\mathrm{d}^2\mathcal{D}_n}{\mathrm{d}^2\theta}(\theta_0) + \mathcal{O}((\theta-\theta_0)^3)\ .$$

Now: a) we ignore the negligible residual $\mathcal{O}((\theta-\theta_0)^3)$; b) we note that $\mathcal{D}_n(\theta_0) \xrightarrow{\mathbb{P}} \mathcal{D}(\theta_0) = 0$ and c) we note that (verify!)

$$\frac{\mathrm{d}\mathcal{D}_n}{\mathrm{d}\theta}(\theta_0) \xrightarrow{\mathbb{P}} \frac{\mathrm{d}\mathcal{D}}{\mathrm{d}\theta}(\theta_0) = 0\ .$$

Posterior Asymptotics

Further, we have d):

$$\frac{\mathrm{d}^2\mathcal{D}_n}{\mathrm{d}^2\theta}(\theta_0) \xrightarrow{\mathbb{P}} -\int f(x\,|\,\theta_0)\frac{\mathrm{d}^2\ln f(x\,|\,\theta)}{\mathrm{d}\theta^2}\Big|_{\theta=\theta_0}\mathrm{d}x := I(\theta_0),$$

the last being the definition of Fisher's information (a matrix, in general). Therefore, as $n\uparrow\infty$

$$\mathcal{D}_n(\theta) \simeq \frac{(\theta-\theta_0)^2}{2}I(\theta_0)\ .$$

Similarly, $\ln\pi(\theta) = \ln\pi(\theta_0) + \mathcal{O}(\theta-\theta_0)$. Using all of these results, the posterior is then proportional to (as $n\uparrow\infty$):

$$\pi(\theta\,|\,\boldsymbol{X}_n) \propto \exp(-n\tfrac{(\theta-\theta_0)^2}{2}I(\theta_0))$$

In other words, the posterior converges to the pdf of the

$$\mathsf{N}\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

distribution.

Posterior Asymptotics

Example: Normal model
Let $X_1, \ldots, X_n \sim \mathsf{N}(\theta, \sigma^2)$ where σ^2 is known.
As usual, this gives the log-likelihood

$$\ln f(\boldsymbol{x} \,|\, \theta) = -\frac{\sum_{i=1}^{n}(x_i - \theta)^2}{\sigma^2} + c_1$$

from which we obtain

$$\frac{\mathrm{d} \ln f(\boldsymbol{x} \,|\, \theta)}{\mathrm{d}\theta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \theta)$$

and so

$$\frac{\mathrm{d}^2 \ln f(\boldsymbol{x} \,|\, \theta)}{\mathrm{d}^2\theta} = -n/\sigma^2.$$

The mle is $\hat{\theta} = \bar{X}_n$ and $I_n(\theta) = nI(\theta) = n/\sigma^2$. So asymptotically, as $n \to \infty$,

$$\theta \,|\, \boldsymbol{X}_n \sim \mathsf{N}(\bar{X}_n, \sigma^2/n).$$

This is true for any prior distribution which places non-zero probability around the true value of θ.

Likelihood Asymptotics

Consider again the likelihood model $X \sim \mathsf{Bin}(n, \theta)$.

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, \ldots, n$$

thus

$$\log(f(x \mid \theta)) = x \log \theta + (n-x) \log(1-\theta)$$

so,

$$\frac{\mathrm{d} \ln f(x \mid \theta)}{\mathrm{d}\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

and

$$\frac{\mathrm{d}^2 \ln f(x \mid \theta)}{\mathrm{d}^2\theta} = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}.$$

Consequently

$$I_n(\theta) = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}. \quad (\mathbb{E}X = n\theta)$$

Thus, as $n \to \infty$, we have

$$\theta \mid X \xrightarrow{\mathrm{d}} \mathsf{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right).$$

Outline

1. Posterior Inference
    - Loss functions, predictive inference
2. Posterior Asymptotics
3. Monte Carlo methods
    - Importance sampling.

Importance Sampling

Importance Sampling Algorithm

Algorithm 1 Importance Sampling

```
for $i = 1, \dots, N$ do
    Draw $X^{(i)} \sim g(x)$
    $W^{(i)} \propto \frac{f(X^{(i)})}{g(X^{(i)})}$
```

Notes:

- The samples $(X^{(1)}, W^{(1)}), \dots, (X^{(N)}, W^{(N)})$ are weighted samples from $f(x)$.
- Weight is $\propto f(X)/g(X)$, not $f(X)/Kg(X)$ (as for rejection sampling) as K is lost in proportionality

How does inference work for weighted samples?

Importance Sampling

How does inference work for weighted samples? (Assume $\int f(x)\mathrm{d}x = 1$.)
Unweighted expectation:

$$\mathbb{E}_g[h(X)] = \int h(x)g(x)\mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N} h(X^{(i)})$$

where $X^{(1)}, \ldots, X^{(N)}$ are samples from $g(x)$.

Weighted expectation: Defining weights $w(x) = f(x)/g(x)$, then

$$\begin{aligned}\mathbb{E}_g[w(x)h(x)] &= \int w(x)h(x)g(x)\mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N} W^{(i)}h(X^{(i)}) \\ &= \int h(x)f(x)\mathrm{d}x \\ &= \mathbb{E}_f[h(x)]\end{aligned}$$

where $X^{(1)}, \ldots, X^{(N)}$ are samples from $g(x)$.

i.e. Weighted expectations under $g(x)$ act as expectations under $f(x)$.

Importance Sampling

What if $f(x)$ is unnormalised?

We then have $f(x) = \tilde{f}(x)/\mathcal{Z}$ where $\mathcal{Z} = \int \tilde{f}(x)\mathrm{d}x$ is unknown.

Weighted expectation: Defining weights $\tilde{W}(X) = \tilde{f}(X)/g(X)$, note that:

$$\mathbb{E}_g[\tilde{w}(X)] = \int \tilde{w}(x)g(x)\mathrm{d}x = \int \tilde{f}(x)\mathrm{d}x = \mathcal{Z} \approx \frac{1}{N}\sum_{i=1}^{N} \tilde{W}(X^{(i)}).$$

$$\begin{aligned}
\mathbb{E}_f[h(x)] &= \int h(x)f(x)\mathrm{d}x = \frac{1}{\mathcal{Z}}\int h(x)\tilde{f}(x)\mathrm{d}x \\
&= \frac{1}{\mathcal{Z}}\int \tilde{w}(x)h(x)g(x)\mathrm{d}x = \frac{\mathbb{E}_g[\tilde{w}(x)h(x)]}{\mathbb{E}_g[\tilde{w}(x)]} \\
&\approx \frac{\frac{1}{N}\sum_{i=1}^{N}\tilde{W}(X^{(i)})h(X^{(i)})}{\frac{1}{N}\sum_{i=1}^{N}\tilde{W}(X^{(i)})}^{*} = \sum_{i=1}^{N} W(X^{(i)})h(X^{(i)})
\end{aligned}$$

where $W(X) = \tilde{W}(X)/\sum_{i=1}^{N}\tilde{W}(X)$ and $X^{(1)}, \ldots, X^{(N)} \sim g(x)$.

i.e. normalise weights (to sum to one) then take expectations.

* – this is a biased estimator.

Importance Sampling

Example:
Simulate from the density

$$f(x) = \begin{cases} 20x(1-x)^3 & 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Use importance sampling with

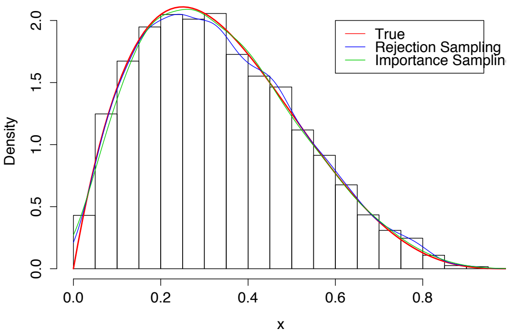
$$g(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(that is, use the uniform density on $[0,1]$).

Previously (Lecture 3) used rejection sampling with

- $K = 135/64$ (if $f(x)$ is normalised as $f(x) = 20x(1-x)^3$)
- $K = 27/256$ (if $f(x) \propto x(1-x)^3$.

Importance Sampling



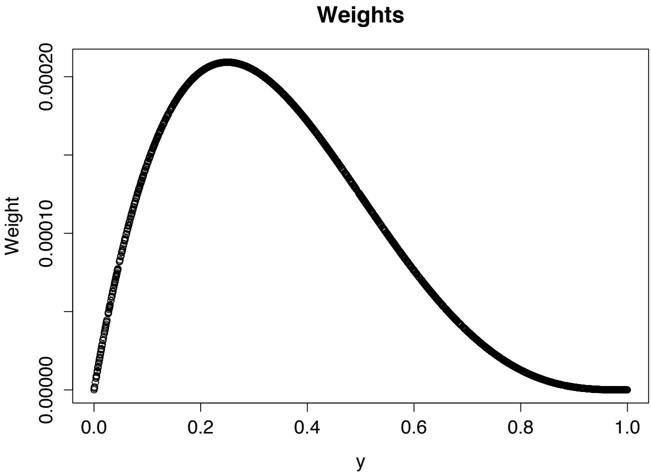
```
L=10000
K=135/64
x=runif(L)
ind=(runif(L)<(20*x*(1-x)^3/K))
hist(x[ind],probability=T,
xlab="x",ylab="Density",main="")
xx=seq(0,1,length=100
lines(xx,20*xx*(1-xx)^3,lwd=2,col=2)
d=density(x[ind],from=0,to=1)
lines(d,col=4)

y=runif(L)
w=20*y*(1-y)^3
wTilde=y*(1-y)^3
W=wTilde/sum(wTilde)
d=density(y,weights=w,from=0,to=1)

lines(d,col=3)
```

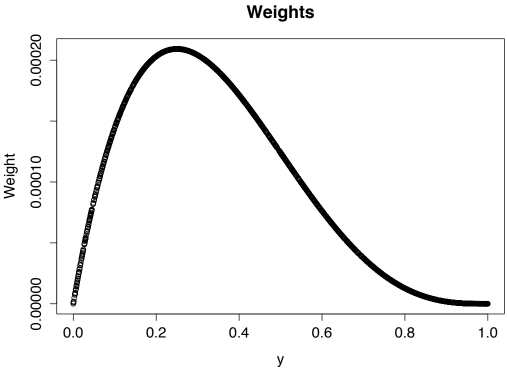
```
> mean(x[ind])
[1] 0.3353401
> mean(y)
[1] 0.5042156
> mean(w*y)
[1] 0.3334709
> sum(W*y)
[1] 0.3363843
```

Importance Sampling



- Shows density of $f(x)$ only as $g(x)$ is uniform.
- In general shows $\propto f(x)/g(x)$

Importance Sampling



Note:

- Variability in weights mean some weighted samples contribute more than others in computations
- Samples with low weights have small contribution
- $\Rightarrow$ for efficiency, would like all samples to contribute as equally as possible

Concept: Variability of weights

- If weights are highly variable then $\text{Var}(w_i)$ is high. (=bad)
- If weights have low variability then they are all similar values ($\text{Var}(w_i)$ is small) (=good).
- For best performance, prefer low variability weights

Importance Sampling

Weights variance usually measured through Effective Sample Size (ESS)

$$\text{ESS} = \left[\sum_{i=1}^{n} (W^{(i)})^2\right]^{-1}$$

where $W^{(i)} = W(X^{(i)})$ are normalised weights.
Note that

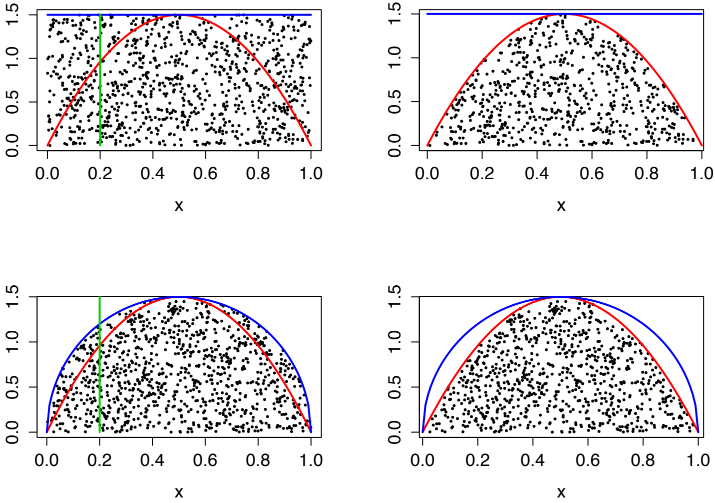
$$1 \leq \text{ESS} \leq n.$$

- ESS=1 when $W^{(1)} = 1, W^{(2)}, \ldots, W^{(n)} = 0$ (sample depletion)
- ESS=n when $W^{(i)} = 1/n$ for all i.
- Loose interpretation: equivalent number of equally weighted independent samples

To maximise ESS, choose $g(x)$ to closely match $f(x)$.
Same idea when improving efficiency of rejection sampling.

Importance Sampling

Example: Obtain N samples from Beta$(2,2)$.



red = f(x), blue = Kg(x)

Top: g(x)=Beta(1,1), eff = 681/1000 ESS=4172.82 (84%)

Bottom: g(x)=Beta(1.5,1.5), eff = 848/1000 ESS=4634.02 (93%)

Importance Sampling

Monte Carlo integration using $g(x)$:

In order to estimate the integral of $\mathcal{Z} = \int \phi(x)\mathrm{d}x$ we can:

- Reparameterise to an integral over $\mathsf{U}(0,1)$ then compute as

$$\int \phi(x)\mathrm{d}x = \int \phi'(u)\mathrm{d}u \approx \frac{1}{N}\sum_i \phi'(U^{(i)})$$

  with $U^{(1)}, \dots, U^{(N)} \sim \mathsf{U}(0,1)$ (as before).

- As above, but with $U^{(1)}, \dots, U^{(N)} \sim q(u)$ on (0,1).

$$\int \phi(x)\mathrm{d}x = \int \phi'(u)\mathrm{d}u = \int \frac{\phi'(u)}{q(u)} q(u)\mathrm{d}u \approx \frac{1}{N}\sum_i \frac{\phi'(U^{(i)})}{q(U^{(i)})}$$

  where $q = \mathsf{U}(0,1)$ recovers the first method.

- Integrate without transformation to (0,1)

$$\int \phi(x)\mathrm{d}x = \int \frac{\phi(x)}{g(x)} g(x)\mathrm{d}x \approx \frac{1}{N}\sum_i \frac{\phi(X^{(i)})}{g(X^{(i)})}$$

  where $X^{(1)}, \dots, X^{(N)} \sim g(x)$.

The aim is to choose q to minimise the variability of $\phi(X)/q(X)$ (etc).