Introduction to Monte Carlo Methods
Lecture Notes

Chapter 2
Importance Sampling and Sequential Monte Carlo

Qing Zhou[*]

Contents

[*] UCLA Department of Statistics (email: zhou@stat.ucla.edu).

Goal: To estimate integrals $\begin{cases} \text{1. Expectations: } \mu_h = \mathbb{E}h(X) = \int_D h(x)f(x)dx, \\ \quad X \sim f, x \in D; \\ \\ \text{2. } I = \int_D h(x)dx. \end{cases}$

1. Importance Sampling

1.1. Given probability densities

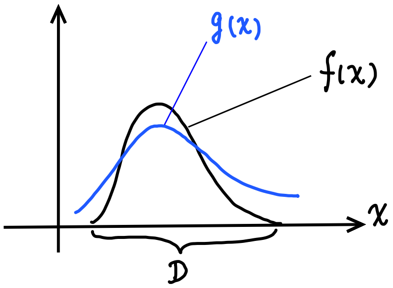
Suppose $X \sim f,\ x \in D,$ such that $\int_D f(x)dx = 1$.

Want to compute $\mathbb{E}_f[h(X)] = \int_D h(x)f(x)dx$.

But: we cannot sample from f directly!!

Find a trial distribution $g(x), \int g(x)dx = 1$, such that

$$\begin{cases} \text{1. } g(x) > 0 \text{ for all } x \in D, \\ \text{2. we can sample from } g. \end{cases}$$



Key Idea: Suppose that $D \subset S = \text{supp}(g) := \{x : g(x) > 0\}$.

$$\begin{aligned}
\mathbb{E}_f h(X) &= \int_D h(x)f(x)dx \\
&= \int_S h(x)f(x)dx && (\because f(x) = 0 \text{ for } x \notin D) \\
&= \int_S h(x)\frac{f(x)}{g(x)}g(x)dx && (g(x) > 0 \text{ for } x \in S) \\
&= \mathbb{E}_g\left[h(X)\frac{f(X)}{g(X)}\right] && (X \sim g) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})\frac{f(x^{(i)})}{g(x^{(i)})} && (x^{(1)}, x^{(2)}, \cdots, x^{(n)} \overset{\text{iid}}{\sim} g).
\end{aligned}$$

Because by the strong law of large numbers, if $x^{(i)} \overset{\text{iid}}{\sim} g$ for $i = 1, \dots, n$, then

$$\frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})\frac{f(x^{(i)})}{g(x^{(i)})} \xrightarrow{a.s.} \mathbb{E}_g\left[h(X)\frac{f(X)}{g(X)}\right] = \mathbb{E}_f[h(X)].$$

The precise statement of the strong law of large numbers:

Theorem 1. *Let $X_i, i = 1, 2, ...$ be a sequence of independent and identically distributed random variables, each having a finite mean $\mu = \mathbb{E}(X_i)$. Then*

$$\mathbb{P}\Big[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i = \mu\Big] = 1,$$

that is, $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu$ as $n \to \infty$.

Definition 1. The importance weight of $x^{(i)}$ is $w(x^{(i)}) = \frac{f(x^{(i)})}{g(x^{(i)})}$.

Algorithm 1. Importance sampling (IS):

1. Draw $x^{(1)}, \cdots, x^{(n)}$ from g independently, and calculate importance weight

$$w(x^{(i)}) = \frac{f(x^{(i)})}{g(x^{(i)})} \quad \text{for } i = 1, 2, \cdots, n; \tag{1}$$

2. Estimate $\mathbb{E}_f(h)$ by

$$\widehat{\mu}_h = \frac{1}{n}\sum_{i=1}^{n} w(x^{(i)})h(x^{(i)}).$$

Then $\widehat{\mu}_h \xrightarrow{a.s.} \mathbb{E}_f[h(X)]$ as $n \to \infty$.

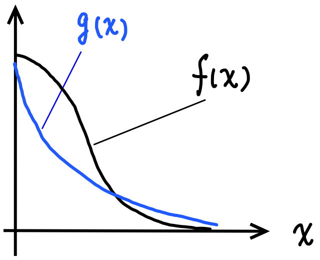
Example 1. Use IS to estimate $E_f(X)$, where f is absolute normal, i.e.,

$$f(x) = \sqrt{\frac{2}{\pi}}e^{-\frac{x^2}{2}}, \ (x \geq 0).$$

In other words, $X = |Z|$, where $Z \sim \mathcal{N}(0, 1)$.

Choose $g(x) = 2e^{-2x}$ (Exp($\lambda = 2$)).
$\int_0^\infty g(x)dx = 1$, $g(x) > 0$ for $x \geq 0$.

Importance weight $w(x) = \frac{f(x)}{g(x)} = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}+2x}$.
$\therefore$ Draw $x^{(1)}, \cdots, x^{(n)} \overset{\text{iid}}{\sim}$ Exp($\lambda = 2$).
Then $\mathbb{E}_f(X) \approx \widehat{\mu}_X = \frac{1}{n}\sum\limits_{i=1}^{n} w(x^{(i)})x^{(i)}$.



Example 2. The mean & variance of importance weights.

$$\frac{1}{n}\sum_{i=1}^{n} w(x^{(i)}) \xrightarrow{a.s.} \mathbb{E}_g[w(x^{(i)})] = \int \frac{f(x)}{g(x)} g(x)dx = 1.$$

Efficiency of Algorithm 1:

$$\text{Eff} = \frac{1}{\text{Var}_g(w(X)) + 1} : \quad \text{the closer } g(x) \text{ is to } f(x)\text{, the more efficient.}$$

Quantify the accuracy of $\widehat{\mu}_h$ by estimating $\text{Var}(\widehat{\mu}_h) = \frac{1}{n}\text{Var}_g[w(X)h(X)]$.

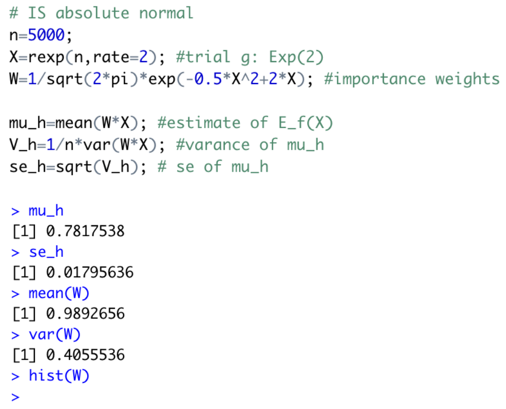
Since $\{x^{(i)}, i = 1, \dots, n\}$ is an iid sample from g, we can estimate $\text{Var}_g[w(X)h(X)]$ by the sample variance of $\{w(x^{(i)})h(x^{(i)}), i = 1, \dots, n\}$. Therefore, an estimate of $\text{Var}(\widehat{\mu}_h)$ is

$$\begin{aligned}\widehat{V} &= \frac{1}{n} \cdot \text{sample variance}\Big\{w(x^{(1)})h(x^{(1)}), \dots, w(x^{(n)})h(x^{(n)})\Big\} \\ &= \frac{1}{n}\Big[\frac{1}{n-1}\sum_{i=1}^{n}\Big\{w(x^{(i)})h(x^{(i)}) - \widehat{\mu}_h\Big\}^2\Big]. \qquad (2)\end{aligned}$$

The following code implements Example 1 with $n = 5,000$ iterations. The vector `W` stores all importance weights. The estimate $\widehat{\mu}_X$ (`mu_h`) $= 0.782$, which is close the true expectation

$$\mathbb{E}_f(X) = \int_0^\infty x f(x)dx = \sqrt{2/\pi} \approx 0.798.$$

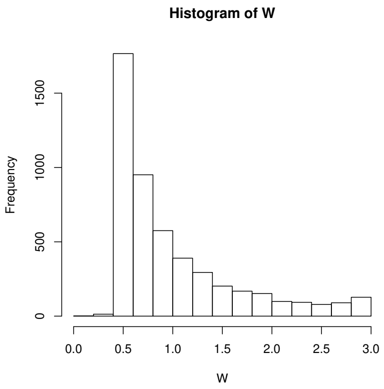
The estimated variance $\widehat{V}$ of $\widehat{\mu}_X$ is `V_h`, from which we get an estimated standard error `se_h`=0.018. The histogram and the variance (`var(W)=0.406`) of the importance weights show that this is a quite efficient design.



```
# IS absolute normal
n=5000;
X=rexp(n,rate=2); #trial g: Exp(2)
W=1/sqrt(2*pi)*exp(-0.5*X^2+2*X); #importance weights

mu_h=mean(W*X); #estimate of E_f(X)
V_h=1/n*var(W*X); #varance of mu_h
se_h=sqrt(V_h); # se of mu_h

> mu_h
[1] 0.7817538
> se_h
[1] 0.01795636
> mean(W)
[1] 0.9892656
> var(W)
[1] 0.4055536
> hist(W)
>
```



Example 3. Use $N(2,1)$ as the trial distribution g to estimate the integral

$$\mu = \int_0^5 \exp\left[-0.5(x-2)^2 - 0.1|\sin(2x)|\right] dx$$

by importance sampling. Generate $n = 1000$ samples to calculate $\widehat{\mu}$ and estimate its standard deviation.

To solve this problem, the first step is to write μ as an expectation with respect to a target distribution f. Note that the integrand is

$$\exp\left[-0.5(x-2)^2 - 0.1|\sin(2x)|\right] = \exp\left[-\frac{1}{2}(x-2)^2\right] \exp\left[-0.1|\sin(2x)|\right],$$

for $x \in (0,5)$, where the first factor can be regarded as an unnormalized density for $N(2,1)$ truncated to $(0,5)$. Thus we can write

$$\begin{aligned}\mu &= \int_0^5 \left\{\frac{1}{Z}\exp\left[-\frac{1}{2}(x-2)^2\right]\right\}\left\{Z\exp\left[-0.1|\sin(2x)|\right]\right\} dx \\ &= \mathbb{E}_f\left\{Z\exp\left[-0.1|\sin(2x)|\right]\right\},\end{aligned}$$

where $Z = \int_0^5 \exp\left[-(x-2)^2/2\right] dx$ is the normalizing constant and the target distribution

$$f(x) = \frac{1}{Z}\exp\left[-\frac{1}{2}(x-2)^2\right] I(0 < x < 5)$$

is $N(2,1)$ truncated to $(0,5)$. Now we may apply Algorithm 1 with $g = N(2,1)$ and $n = 1,000$ to compute $\widehat{\mu}$ and use (2) to estimate its standard deviation. In this process, the normalizing constant Z cancels when we compute $w(x)h(x)$.

1.2. With unknown normalizing constant

In Example 1:

$f(x) = \sqrt{\frac{2}{\pi}}e^{-\frac{x^2}{2}}\ (x \geq 0)$: normalized density, $\int_0^\infty f(x)dx = 1$.

Let $q(x) = e^{-\frac{x^2}{2}}\ (x \geq 0)$: unnormalized density, $\int_0^\infty q(x)dx = \sqrt{\frac{\pi}{2}} \neq 1$.

Then $Z_q := \int_0^\infty e^{-\frac{x^2}{2}}dx = \sqrt{\frac{\pi}{2}}$ is the normalizing constant of $q(x)$.

Definition 2. Suppose that $q(x) > 0$, for $x \in D$, and $\int_D q(x)dx = Z_q < \infty$, then we call $q(x)$ an unnormalized density on D. The corresponding (normalized) probability density is $f(x) = \frac{1}{Z_q}q(x)$.

Want to estimate

$$\mathbb{E}_f[h(X)] = \int_D h(x)f(x)dx = \int_D h(x)\frac{q(x)}{Z_q}dx,$$

where $X \sim f(x) \propto q(x)$, but Z_q is unknown, and we **cannot** simulate from $f(x)$.

Use IS, let $g(x) = \frac{1}{Z_r}r(x)$, $Z_r = \int r(x)dx < \infty$. So $r(x)$ is an unnormlized density for the trial distribution g, where Z_r may be unknown.

Algorithm 2. Importance sampling without normalizing constants:

1. Draw $x^{(1)}, \cdots, x^{(n)}$ from g independently, and calculate importance weight $w(x^{(i)}) = q(x^{(i)})/r(x^{(i)})$ for $i = 1, 2, \cdots, n$;
2. Estimate $E_f[h(X)]$ by

$$\widehat{\mu}_h = \frac{\sum\limits_{i=1}^{n} w(x^{(i)})h(x^{(i)})}{\sum\limits_{i=1}^{n} w(x^{(i)})}. \tag{3}$$

Then, $\widehat{\mu}_h \xrightarrow{a.s.} \mathbb{E}_f[h(X)]$ as $n \to \infty$.

Proof.

$$\frac{1}{n}\sum_{i=1}^{n} w(x^{(i)}) \xrightarrow{a.s.} \mathbb{E}_g\left[\frac{q(X)}{r(X)}\right] = \int \frac{q(x)}{r(x)}g(x)dx = \frac{Z_q}{Z_r}.$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} w(x^{(i)})h(x^{(i)}) \xrightarrow{a.s.} \mathbb{E}_g\left[\frac{q(X)}{r(X)}h(X)\right] &= \int \frac{q(x)}{r(x)}h(x)g(x)dx \\ &= \frac{1}{Z_r}\int q(x)h(x)dx.\end{aligned}$$

By (3):

$$\widehat{\mu}_h \xrightarrow{a.s.} \frac{1}{Z_q}\int q(x)h(x)dx = \int \frac{q(x)}{Z_q}h(x)dx = \int h(x)f(x)dx = \mathbb{E}_f[h(X)].$$

□

Example 4. Repeat Example 1 using unnormalized densities:
$f(x) \propto q(x) = e^{-\frac{x^2}{2}}$ $(x \geq 0)$, $g(x) \propto r(x) = e^{-2x}$ $(x \geq 0)$ [Exp($\lambda = 2$)],
importance weight $w(x) = \frac{q(x)}{r(x)} = e^{-\frac{x^2}{2}+2x}$.

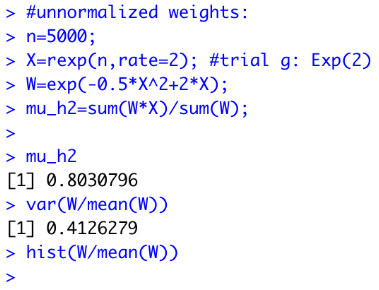
Draw $x^{(1)}, \cdots, x^{(n)} \overset{\text{iid}}{\sim} \text{Exp}(\lambda = 2)$.
Then

$$\mathbb{E}_f(X) \approx \widehat{\mu}_X = \frac{\sum_{i=1}^{n} w(x^{(i)}) x^{(i)}}{\sum_{i=1}^{n} w(x^{(i)})}. \tag{4}$$

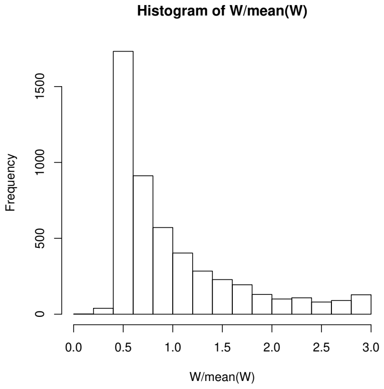
Efficiency:

$$\text{Eff} = \frac{1}{\text{Var}_g[\frac{w(X)}{\mathbb{E}_g(w(X))}] + 1} = \frac{[\mathbb{E}_g(w(X))]^2}{\text{Var}_g(w(X)) + [\mathbb{E}_g(w(X))]^2}.$$

The following code estimates $\mathbb{E}_f(X)$ using (4) (`mu_h2=0.803`). The variance $\text{Var}_g[\frac{w(X)}{\mathbb{E}_g(w(X))}] = 0.413$ is very comparable to the variance of W (`var(W)=0.406`) in the implementation of Example 1.



```
> #unnormalized weights:
> n=5000;
> X=rexp(n,rate=2); #trial g: Exp(2)
> W=exp(-0.5*X^2+2*X);
> mu_h2=sum(W*X)/sum(W);
>
> mu_h2
[1] 0.8030796
> var(W/mean(W))
[1] 0.4126279
> hist(W/mean(W))
>
```



Example 5 (Truncated normal)**.** Suppose our target distribution has density

$$f(x) \propto q(x) = \phi(x) I(x > c) = \begin{cases} \phi(x), & \text{if } x > c, \\ 0, & \text{otherwise,} \end{cases}$$

where $\phi(x)$ is the density of $N(0,1)$, and we want to estimate

$$\mathbb{E}_f[X^k] = \int_c^\infty x^k f(x) dx = \int_c^\infty x^k \frac{q(x)}{Z_q} dx,$$

without calculating the normalizing constant Z_q.

Use shifted Exp(λ) as the trial distribution g in Algorithm 2:

$$g(x) = \lambda e^{-\lambda(x-c)} I(x > c) = \begin{cases} \lambda e^{-\lambda(x-c)}, & \text{for } x > c; \\ 0, & \text{otherwise.} \end{cases}$$

If $X' \sim \mathrm{Exp}(\lambda)$, then $X = X' + c \sim g$.

Then the importance weight for any $x > c$ is

$$w(x) = \frac{q(x)}{g(x)} = \frac{\phi(x)}{\lambda e^{-\lambda(x-c)}} = \frac{\exp(-\frac{x^2}{2} + \lambda x - \lambda c)}{\sqrt{2\pi}\lambda} \propto \exp\Big(-\frac{x^2}{2} + \lambda x\Big),$$

after ignoring all constant factors that do not depend on x.

Therefore, our estimate of $\mathbb{E}_f[X^k]$ is

$$\widehat{\mu} = \frac{\sum\limits_{i=1}^{n} w(x_i)(x_i)^k}{\sum\limits_{i=1}^{n} w(x_i)},$$

where $x_i \sim \mathrm{Exp}(\lambda) + c$ and $w(x_i) = \exp(-x_i^2/2 + \lambda x_i)$ for $i = 1, \ldots, n$.

Remark 1. The estimate (3) is invariant to rescaling the importance weights $w(x^{(i)})$. That is why we can simply use $w(x_i) = \exp(-x_i^2/2 + \lambda x_i)$ in the above example, ignoring the constant $\exp(-\lambda c)/(\sqrt{2\pi}\lambda)$ in the importance weight.

1.3. Importance resampling

Note the importance sampling does not generate samples from the target distribution f. If we want to get samples from f, we may apply a resampling approach according to the importance weights: Given samples with importance weights $\{(x^{(i)}, w^{(i)}) : i = 1, \ldots, n\}$, where $w^{(i)} = w(x^{(i)})$, if we resample with replacement $x^{(*i)}$ from $\{x^{(1)}, \ldots, x^{(n)}\}$ with probabilities proportional to the importance weights, i.e.

$$\mathbb{P}\big[x^{(*i)} = x^{(k)} \big| \{x^{(1)}, \ldots, x^{(n)}\}\big] = \frac{w^{(k)}}{\sum_{j=1}^{n} w^{(j)}}, \tag{5}$$

then the distribution of $\{x^{(*1)}, \ldots, x^{(*n)}\}$ is approximately the target distribution f when n is large. This method is called sampling-importance-resampling.

The following derivation justifies this resampling method: When n is large,

$$\frac{1}{n}\sum_{j=1}^{n} w^{(j)} \approx \mathbb{E}_g[w(X)],$$

which is a constant. Thus, (5) is approximately

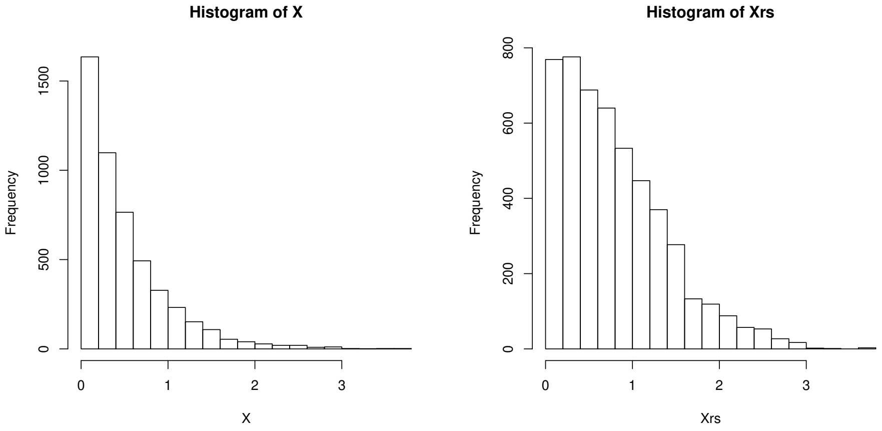
$$\mathbb{P}\big[x^{(*i)} = x^{(k)} \big| \{x^{(1)}, \ldots, x^{(n)}\}\big] = \frac{w^{(k)}/n}{\frac{1}{n}\sum_{j=1}^{n} w^{(j)}} \approx \frac{w^{(k)}/n}{\mathbb{E}_g[w(X)]} \propto w^{(k)}.$$

Let p^* be the density of $x^{(*i)}$ (the samples after resampling). Then we have

$$p^*(x) \propto g(x)w(x) \propto g(x)\frac{f(x)}{g(x)} = f(x).$$

The following code implements the resampling approach for Example 4. The histogram of `X` is $\text{Exp}(\lambda = 2)$, while after resampling the histogram of `Xrs` is f (the absolute normal).

```
n=5000;
X=rexp(n,rate=2); #trial g: Exp(2)
W=exp(-0.5*X^2+2*X);
Xrs=sample(X,n,replace=TRUE,prob=W) #importance resampling
```



2. Estimating Volume and Normalizing Constant

2.1. *Estimating V_D by IS*

Let $D \subset \mathbb{R}^p$ and $h(x) = I(x \in D) = \begin{cases} 1, & \text{if } x \in D; \\ 0, & \text{otherwise.} \end{cases}$

Then the volume of D: $V_D := \int_D dx = \int_{\mathbb{R}^p} h(x)dx$ (Denote by $|D|$).

1. Find a region A s.t. :
   a) $D \subset A$,
   b) we can generate $x \sim \text{Unif}(A) := g$,

$$g(x) = \begin{cases} \frac{1}{|A|}, & \text{if } x \in A; \\ 0, & \text{otherwise.} \end{cases}$$

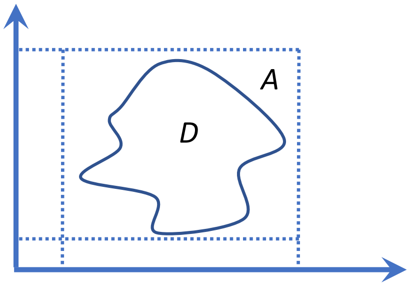
2. Generate $x^{(1)}, x^{(2)}, \cdots, x^{(n)} \sim g(x)$, calculate $w^{(i)} = \frac{I(x^{(i)} \in D)}{g(x^{(i)})} = |A| I(x^{(i)} \in D)$;
3. Approximate $|D|$ by $\hat{V}_D = \frac{1}{n} \sum_{i=1}^{n} w^{(i)}$.

Then $\hat{V}_D \xrightarrow{a.s.} |D|$.

Proof. By the strong law of large numbers (SLLN),

$$\hat{V}_D = \frac{1}{n} \sum_i w^{(i)} \xrightarrow{a.s.} \mathbb{E}_g(W) = |A| \cdot \mathbb{E}_g[I(X \in D)] = |A| \cdot \frac{|D|}{|A|} = |D|.$$

□



Now we quantify the accuracy of $\hat{V}_D$. The importance weight

$$W = \frac{I(X \in D)}{g(X)} = |A| \cdot I(X \in D), \; X \sim \text{Unif}(A) = g.$$

Note that $I(X \in D) \sim \text{Bern}(p = |D|/|A|)$. It is easy to see that $\hat{V}_D$ is unbiased for $|D|$:

$$\mathbb{E}(\hat{V}_D) = |A|\mathbb{E}[I(X \in D)] = |A| \cdot \mathbb{P}(X \in D) = |A|(|D|/|A|) = |D|.$$

Find the variance of $\hat{V}_D$:

$$\text{Var}[W] = |A|^2 \cdot \text{Var}(I(X \in D)) = |A|^2 \frac{|D|}{|A|}\left(1 - \frac{|D|}{|A|}\right) = |D|(|A| - |D|).$$

$$\Rightarrow \text{Var}(\hat{V}_D) = \frac{1}{n}\,\text{Var}(W) = \frac{1}{n}|D|(|A| - |D|).$$

- Choose A close to D: reduce variance of $\hat{V}_D$.
- In practice, $\widehat{\text{Var}}(\hat{V}_D) = \frac{1}{n}\hat{V}_D(|A| - \hat{V}_D)$.

Example 6. Estimate the area of the unit disk $D = \{(x, y) : x^2 + y^2 \leq 1\}$.

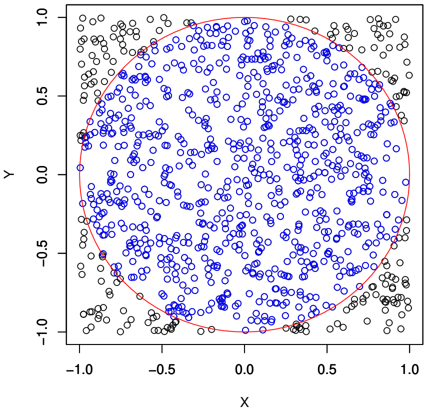
Choose $A : [-1, 1] \times [-1, 1] \Rightarrow |A| = 4$

(1) For $i = 1, 2, \cdots, n$, generate $x^{(i)} \sim$ Unif $(-1, 1)$ and $y^{(i)} \sim$ Unif $(-1, 1)$ independently; compute importance weights

$$w^{(i)}(x^{(i)}, y^{(i)}) = |A| \cdot I((x^{(i)}, y^{(i)}) \in D) = \begin{cases} 4, & \text{if } (x^{(i)})^2 + (y^{(i)})^2 \leq 1; \\ 0, & \text{otherwise.} \end{cases}$$

(2) Estimated area $\hat{V}_D = \frac{1}{n}\sum_{i=1}^n w^{(i)}$;

$$\text{Var}(\hat{V}_D) = \frac{1}{n}|D|(|A| - |D|) = \frac{1}{n} \cdot \pi(4 - \pi) = \frac{\pi(4 - \pi)}{n}.$$



In the plot, the importance weights $w^{(i)}$ for blue and black points are 4 and 0, respectively.

2.2. Normalization Constant

Consider $q(x) = I(x \in D)$ as an unnormalized density for the uniform distribution on D.

Normalizing constant of q: $Z_q = \int q(x)dx = \int I(x \in D)dx = \int_D dx = |D|$.

$\Rightarrow \frac{q(x)}{|D|} = \begin{cases} \frac{1}{|D|}, & x \in D; \\ 0, & \text{otherwise.} \end{cases}$ (normalized) pdf for Unif(D).

Example 7.
(1) $q(x) = e^{-\frac{x^2}{2}}$, $x \in (-\infty, \infty)$: unnormalized density for $N(0, 1)$.

$Z_q \triangleq \int_{-\infty}^{+\infty} q(x)dx = \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$

$\Rightarrow f(x) = \frac{q(x)}{Z_q} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \sim N(0, 1)$

(2) $q(x) = e^{-5x}, \ \ x \geq 0 : \exp(\lambda = 5)$.

$Z_q = \int_0^{\infty} e^{-5x} dx = \frac{1}{5}$

$\Rightarrow f(x) = \frac{q(x)}{Z_q} = 5e^{-5x}$

(3) $q(x) = x^3 e^{-\frac{1}{2}x}$, $x \geq 0$: Gamma (α, β), pdf $= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

$\Rightarrow \alpha = 4, \beta = \frac{1}{2}$.

$\therefore q(x)$ is unnormalized Gamma (4, $\frac{1}{2}$).

$$\Rightarrow \int_0^\infty \underbrace{\frac{(\frac{1}{2})^4}{\Gamma(4)} x^3 e^{-\frac{1}{2}x}}_{f(x):normalized} \, dx = 1 \ \Rightarrow \ Z_q = \int_0^\infty x^3 e^{-\frac{1}{2}x} dx = \frac{\Gamma(4)}{(\frac{1}{2})^4}.$$

(4) $q(x) = x^3(1-x)^2, \quad x \in [0, 1]$: Beta (α, β), pdf $= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$.

$\Rightarrow \alpha = 4, \beta = 3$.

$\therefore f$ is Beta(4, 3) $\ \Rightarrow \ Z_q = \frac{\Gamma(4)\Gamma(3)}{\Gamma(7)}$.

2.3. Estimating Z_q by IS

Given $q(x) > 0,\ x \in D$ an unnormalized density, want to compute normalizing constant

$$Z_q = \int_D q(x)dx = \int q(x)I(x \in D)dx.$$

1. Find a trial distribution $g(x)$ with known normalization constant (i.e. $g(x)$ is the pdf) and its domain fully covers D.
2. Generate $x^{(1)}, \cdots, x^{(n)} \sim g$, and compute importance weights

$$w^{(i)} = \frac{q(x^{(i)})}{g(x^{(i)})} I(x^{(i)} \in D),\ \forall i = 1, \cdots, n;$$

3. Estimate $\hat{Z}_q = \frac{1}{n} \sum\limits_{i=1}^{n} w^{(i)}$.

Then $\hat{Z}_q \xrightarrow{a.s.} Z_q$ as $n \to \infty$.

Proof.

$$\hat{Z}_q \xrightarrow{a.s.} \mathbb{E}_g(w(X)) = \int \frac{q(x)I(x \in D)}{g(x)} \cdot g(x)dx = \int_D q(x)dx = Z_q.$$

□

Let $W = q(X)I(X \in D)/g(X)$. Then $\mathrm{Var}(\hat{Z}_q) = \frac{1}{n}\mathrm{Var}_g(W)$. An estimate of $\mathrm{Var}(\hat{Z}_q)$:

$$\widehat{V} = \frac{1}{n} \cdot \text{sample variance} \left\{ w^{(1)}, \ldots, w^{(n)} \right\}.$$

2.4. Estimating integrals

The above two problems are special cases of estimating integrals by IS. Suppose we want to estimate

$$\mu = \int_D h(x)dx,$$

where $|\mu| < \infty$. Choose a distribution g such that $g(x) > 0$ for $x \in D$. i.e. $D \subset S = \mathrm{supp}(g)$. Define $h(x) = 0$ for $x \notin D$. Then

$$\begin{aligned}\mu &= \int_D h(x)dx = \int_S h(x)I(x \in D)dx \\ &= \int_S h(x)I(x \in D)\frac{g(x)}{g(x)}dx = \mathbb{E}_g\Big[\frac{h(X)I(X \in D)}{g(X)}\Big].\end{aligned}$$

Draw $x^{(i)} \sim g$ independently for $i = 1, \dots, n$, and estimate μ by

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \frac{h(x^{(i)})}{g(x^{(i)})} I(x^{(i)} \in D) \xrightarrow{a.s.} \mu, \qquad \text{as } n \to \infty. \tag{6}$$

And the variance of $\widehat{\mu}$, $\mathrm{Var}(\widehat{\mu}) = \frac{1}{n} \mathrm{Var}_g \Big[h(X) I(X \in D)/g(X) \Big]$, can be estimated by

$$\widehat{V} = \frac{1}{n} \cdot \text{sample variance} \left\{ \frac{h(x^{(1)})}{g(x^{(1)})} I(x^{(i)} \in D), \dots, \frac{h(x^{(n)})}{g(x^{(n)})} I(x^{(i)} \in D) \right\}. \tag{7}$$

In this general setting, $h(x^{(i)})$ is not necessarily positive. In fact, Algorithm 1 in Section 1.1 can be regarded as a special case of this general method as well.

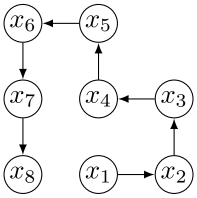
3. Self-Avoid Walk

(not required for exams)

A simple model for (bio)polymers. Consider a simple 2-D lattice model.

Definition 3. A vector $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ is a self-avoid walk (SAW) on the 2-D lattice if

1. $x_t = (a, b)$, where a and b are integers;
2. $\text{distance}(x_t, x_{t+1}) = 1$;
3. $x_{t+1} \neq x_k$ for all $k < t$.



An example SAW of length $N = 8$:

$$x_1 = (0,0), x_2 = (1,0), \ldots, x_8 = (-1,0).$$

Assume every SAW of length N is equally likely ⇒ uniform distribution for $\mathbf{x}$.

$\therefore f(\mathbf{x}) = \frac{1}{Z_N}$, Z_N = total # of different SAWs of length N.

Want to estimate, e.g. Z_N (normalizing constant) and $\mathbb{E}\|x_N - x_1\|^2$ (the mean squared extension).

Naive Simulation (random walk):

Start a random walk at $(0,0) \equiv x_1$, at step i, randomly choose one of the three neighboring positions other than x_{i-1}.

If the chosen position has been occupied by x_t for some $t < i$, restart at $(0,0)$.

Success rate: $r = Z_N/(4 \times 3^{N-2}) \Rightarrow \begin{array}{l} N = 20, \quad r \approx 21.6\%. \\ N = 48, \quad r \approx 0.79\%. \end{array}$

Growing SAW by one-step-look-ahead:

1. $x_1 = (0,0)$;
2. For $t = 1, 2, \cdots, N-1$:
   We have already generated $(x_1, x_2, \cdots, x_t)$ and $x_t = (i,j)$. Let

   $$n_t = \# \text{ of unoccupied neighbors of } x_t.$$

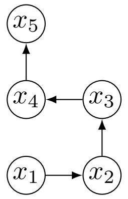
   If $n_t = 0$, restart x_1; otherwise, place x_{t+1} randomly in one of the n_t unoccupied neighbors.

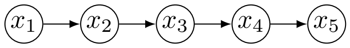
In summary, we draw x_{t+1} from the following conditional distribution:

$$\mathbb{P}(x_{t+1} = (i', j') | x_1, \cdots, x_t) = \frac{1}{n_t}, \qquad t = 1, \ldots, N-1,$$

where (i', j') is one of the n_t unoccupied neighbors of $x_t = (i, j)$.

The one-step-look-ahead "growth" method does *not* produce uniform SAWs. In the following examples, the probability of generating the SAW on the left by the growth method is greater than that of the SAW on the right. However, we may use the growth method as a trial distribution in importance sampling.





$$\mathbb{P} = \frac{1}{4} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{2} \qquad > \qquad \mathbb{P} = \frac{1}{4} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$$

Sequential importance sampling (SIS) for the growth method:

1. Initialize $x_1 = (0, 0)$, $w_1 = 1$;
2. For $t = 1, 2, \cdots, N-1$, use one-step-ahead to draw x_{t+1} and update the weight to
$$w_{t+1} = w_t \cdot n_t. \tag{8}$$

At the end of this algorithm, we will generate $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ with a weight $w(\mathbf{x}) = w_N = 1 \times n_1 \times n_2 \times \cdots \times n_{N-1}$. (If any $n_t = 0, w(\mathbf{x}) = 0$.)

Let $g(\mathbf{x})$ be the probability for the chain $\mathbf{x}$ generated by the growth method, which is used as the trial distribution in importance sampling. The target distribution is $f(\mathbf{x}) = 1/Z_N$, uniform distribution over all SAWs of length N. Then our problem is to estimate the normalizing constant Z_N by IS, so the importance weight is

$$w(\mathbf{x}) = \frac{1}{g(\mathbf{x})} = \frac{1}{\frac{1}{n_1} \times \frac{1}{n_2} \times \cdots \times \frac{1}{n_{N-1}}} = n_1 \times n_2 \times \cdots \times n_{N-1}.$$

Run SIS m times: $(\mathbf{x}^{(1)}, w^{(1)}), \cdots, (\mathbf{x}^{(m)}, w^{(m)})$, where $w^i = w(\mathbf{x}^{(i)})$. Note that this is a special case of Algorithm 2, IS without normalizing constant.

To estimate Z_N:

$$\hat{Z}_N = \frac{1}{m} \sum_{i=1}^{m} w^{(i)} \xrightarrow{a.s.} \mathbb{E}_g(w(\mathbf{x})) = Z_N, \quad \text{as } m \to \infty,$$

since

$$\mathbb{E}_g(w(\mathbf{x})) = \sum_{\mathbf{x}\in\mathcal{S}_N} w(\mathbf{x})g(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{S}_N} \frac{1}{g(\mathbf{x})} \cdot g(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{S}_N} 1 = Z_N,$$

where $\mathcal{S}_N :=$ the set of all SAWs of length N.

Similarly, we can apply (3) to estimate $\mathbb{E}_f[h(\mathbf{x})]$:

$$\widehat{\mu}_h = \frac{\sum_{i=1}^m w^{(i)} h(\mathbf{x}^{(i)})}{\sum_{i=1}^m w^{(i)}}. \tag{9}$$

For example, the squared extension $h(\mathbf{x}) = \|x_N - x_1\|^2$.

4. Sequential Importance Sampling

(not required for exams)

Sequential importance sampling (SIS) and its parallel implementation, sometimes called sequential Monte Carlo.

4.1. The basic idea

High-dimensional problems $\mathbf{x} = (x_1, \dots, x_d)$: difficult to find a good trial distribution g.

Build g sequentially:

$$g(\mathbf{x}) = g_1(x_1) g_2(x_2 \mid x_1) \cdots g_d(x_d \mid x_1, \dots, x_{d-1}).$$

Decompose the target density:

$$\pi(\mathbf{x}) = \pi(x_1)\pi(x_2 \mid x_1) \cdots \pi(x_d \mid x_1, \dots, x_{d-1}).$$

Then importance weight is

$$w(\mathbf{x}) = \frac{\pi(x_1)\pi(x_2 \mid x_1) \cdots \pi(x_d \mid x_1, \dots, x_{d-1})}{g_1(x_1) g_2(x_2 \mid x_1) \cdots g_d(x_d \mid x_1, \dots, x_{d-1})}. \tag{10}$$

Let $\mathbf{x}_t = (x_1, \dots, x_t)$. Then (10) can be calcualated recursively

$$w_t(\mathbf{x}_t) = w_{t-1}(\mathbf{x}_{t-1}) \frac{\pi(x_t \mid \mathbf{x}_{t-1})}{g_t(x_t \mid \mathbf{x}_{t-1})}, \tag{11}$$

so that $w_d(\mathbf{x}_d) = w(\mathbf{x})$.

4.2. SIS algorithm

However, the difficulty in (11) is

$$\pi(x_t \mid \mathbf{x}_{t-1}) = \frac{\pi(\mathbf{x}_t)}{\pi(\mathbf{x}_{t-1})}.$$

We need marginal distribution $\pi(\mathbf{x}_t)$ for each t, which is hard to calculate.

Find a sequence of "auxiliary distributions" $\pi_t(\mathbf{x}_t)$, $t = 1, \dots, d$, as an approximation to the marginal distributions $\pi(\mathbf{x}_t)$ so that $\pi_d(\mathbf{x}) = \pi(\mathbf{x})$.

SIS algorithm: For $t = 1, \dots, d$

1. Draw $x_t \sim g_t(x_t \mid \mathbf{x}_{t-1})$ and put $\mathbf{x}_t = (\mathbf{x}_{t-1}, x_t)$.
2. Update importance weight
$$u_t = \frac{\pi_t(\mathbf{x}_t)}{\pi_{t-1}(\mathbf{x}_{t-1}) g_t(x_t \mid \mathbf{x}_{t-1})}, \tag{12}$$
$$w_t = w_{t-1} u_t, \tag{13}$$
where we define $\pi_0(\cdot) \equiv 1$ and $w_0 \equiv 1$.

Step 1 and 2 generate one weighted sample $(\mathbf{x}, w) = (\mathbf{x}_d, w_d)$. Apply SIS algorithm m times independently to obtain $\{(\mathbf{x}^{(i)}, w^{(i)}) : i = 1, \dots, m\}$.

Example 8 (SAW)**.** In one-step ahead method:

$$\pi_t(\mathbf{x}_t) \propto 1, \quad g_t(x_t \mid \mathbf{x}_{t-1}) = 1/n_{t-1}.$$

Thus, $u_t = n_{t-1}$ and $w_t = w_{t-1} n_{t-1}$, identical to (8). Note that $\pi_t(\mathbf{x}_t)$ is *not* the marginal distribution of $\pi(\mathbf{x}) \propto 1$.

Example 9 (Bayesian variable selection)**.** Consider a linear model with d potential predictors:

$$y = \sum_{j=1}^{d} \beta_j X_j + \varepsilon, \tag{14}$$

where $y, X_j, \varepsilon \in \mathbb{R}^n$ and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. To simplify, we assume that $\sigma^2 = 1$ is given. We want to select a subset of the d predictors to build a linear model, i.e. some $\beta_j = 0$ in (14). Let $\mathbf{z} = (z_1, \dots, z_d) \in \{0, 1\}^d$ indicates which predictors are in the linear model, i.e. $z_j = 1$ if and only if $\beta_j \neq 0$. Our goal is to sample from the posterior distribution of $\mathbf{z}$ given the data y, which defines our target distribution $\pi(\mathbf{z})$:

$$\pi(\mathbf{z}) := p(\mathbf{z} \mid y) \propto p(\mathbf{z}) p(y \mid \mathbf{z}), \tag{15}$$

where $p(\mathbf{z})$ is a prior over $\mathbf{z}$ and $p(y \mid \mathbf{z})$ is the likelihood of y given $\mathbf{z}$.

We choose $p(\mathbf{z}) \propto \exp(-\lambda \sum_j z_j)$ for some $\lambda > 0$, i.e. the prior distribution favors simpler models with fewer predictors. The likelihood part can be approximated by the likelihood of y given the least-square estimate with selected predictors $X_{\mathbf{z}} = \{X_j : z_j = 1\}$, that is

$$p(y \mid \mathbf{z}) \propto \exp\{-\text{RSS}(\mathbf{z})/2\},$$

where $\text{RSS}(\mathbf{z}) = \|y - P_{\mathbf{z}} y\|^2$ is the sum of squared residuals after projecting y onto $X_{\mathbf{z}}$ and $P_{\mathbf{z}} = X_{\mathbf{z}}(X_{\mathbf{z}}^{\mathsf{T}} X_{\mathbf{z}})^{-1} X_{\mathbf{z}}^{\mathsf{T}}$. Therefore, we have the following (unnormalized) target density

$$\pi(\mathbf{z}) = \pi(z_1, \dots, z_d) \propto \exp\Big\{ -\frac{1}{2}\text{RSS}(\mathbf{z}) - \lambda \sum_{j=1}^{d} z_j \Big\}. \tag{16}$$

Let $\mathbf{z}_t = (z_1, \dots, z_t) \in \{0,1\}^t$. Note that marginal distribution of $\mathbf{z}_t$ defined by (16) is

$$\pi(\mathbf{z}_t) = \sum_{z_{t+1}} \cdots \sum_{z_d} \pi(z_1, \dots, z_d)$$

which is hard to compute. To apply SIS, we make the following choices:

$$g_t(z_t \mid \mathbf{z}_{t-1}) \propto \exp(-\lambda z_t), \qquad z_t \in \{0,1\}, \tag{17}$$

$$\pi_t(\mathbf{z}_t) = \pi(z_1, \dots, z_t, 0, \dots, 0) \propto \exp\left\{ -\mathrm{RSS}(\mathbf{z}_t)/2 - \lambda \sum_{j=1}^{t} z_j \right\}. \tag{18}$$

Note that $\pi_d = \pi$ and π_t would be the target distribution if there were only the first t predictors to select from.

Plugging these choices into (12) and (13):

$$u_t = \frac{\pi_t(\mathbf{z}_t)}{\pi_{t-1}(\mathbf{z}_{t-1}) \exp(-\lambda z_t)} = \exp\Big[\{\mathrm{RSS}(\mathbf{z}_{t-1}) - \mathrm{RSS}(\mathbf{z}_t)\}/2 \Big], \tag{19}$$

$$w_t = u_1 \times u_2 \times \cdots \times u_t = \exp\Big[-\mathrm{RSS}(\mathbf{z}_t)/2 \Big]. \tag{20}$$

After the last step $t = d$, we have $w(\mathbf{z}) = w_d = \exp\Big[-\mathrm{RSS}(\mathbf{z})/2 \Big]$. Note that the probability of $\mathbf{z}$ in the trial distribution is

$$g(\mathbf{z}) \propto \exp\Big(-\lambda \sum_{j=1}^{d} z_j \Big),$$

and thus indeed the importance weight

$$w(\mathbf{z}) = \frac{\pi(\mathbf{z})}{g(\mathbf{z})} \propto \frac{\exp\Big\{ -\frac{1}{2}\mathrm{RSS}(\mathbf{z}) - \lambda \sum_{j=1}^{d} z_j \Big\}}{\exp\Big(-\lambda \sum_{j=1}^{d} z_j \Big)} = \exp\Big\{ -\frac{1}{2}\mathrm{RSS}(\mathbf{z}) \Big\}.$$

4.3. Parallel implementation

To improve the efficiency of SIS: (a) run m SIS processes in parallel; (b) resample "good" partial samples. At some step t, if $w_t(\mathbf{x}_t^{(i)})$ is much smaller than the weights of other partial samples, we may want to discard the partial sample $\mathbf{x}_t^{(i)}$. How to do this in a correct way?

Sampling-importance-resampling (SIR):

Given samples with importance weights $\{(\mathbf{x}^{(i)}, w^{(i)}) : i = 1, \ldots, m\}$, if we resample with replacement $\mathbf{x}^{(*i)}$ from $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ with probabilities proportional to the importance weights, i.e.

$$\mathbb{P}\big[\mathbf{x}^{(*i)} = \mathbf{x}^{(k)} \big| \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}\big] = \frac{w^{(k)}}{\sum_j w^{(j)}},$$

then the distribution of $\{\mathbf{x}^{(*1)}, \ldots, \mathbf{x}^{(*m)}\}$ is approximately the target distribution when m is large.

Proof. Let $p_*(\cdot)$ be pdf of $\mathbf{x}^{(*i)}$. For $\mathbf{x} \in \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\} \equiv \mathbf{X}$, write $\mathbf{x}_o = \mathbf{X} \setminus \{\mathbf{x}\}$. Then by the resampling procedure,

$$\begin{aligned} p_*(\mathbf{x}) &= \binom{m}{1} g(\mathbf{x}) \int g(\mathbf{x}_o) \frac{w(\mathbf{x})}{\sum_j w^{(j)}} d\mathbf{x}_o \\ &= g(\mathbf{x}) w(\mathbf{x}) \int \frac{g(\mathbf{x}_o)}{\sum_j w^{(j)}/m} d\mathbf{x}_o \to \frac{1}{Z_\pi} \pi(\mathbf{x}), \end{aligned}$$

since $\sum_j w^{(j)}/m \xrightarrow{a.s.} Z_\pi = \int \pi(\mathbf{y}) d\mathbf{y}$ (= 1 if π is normalized). □

Sequential Monte Carlo (Parallel implementation with resampling):

Suppose $\mathbf{x}_{t-1}^{(i)}$, $i = 1, \ldots, m$, are partial samples from the auxiliary distribution $\pi_{t-1}(\mathbf{x}_{t-1})$.

1. Draw $x_t^{(i)} \sim g_t(x_t \mid \mathbf{x}_{t-1}^{(i)})$ and put $\mathbf{x}_t^{(i)} = (\mathbf{x}_{t-1}^{(i)}, x_t^{(i)})$.
2. Calculate weights as in (12):

$$w_t^{(i)} = \frac{\pi_t(\mathbf{x}_t^{(i)})}{\pi_{t-1}(\mathbf{x}_{t-1}^{(i)}) g_t(x_t^{(i)} \mid \mathbf{x}_{t-1}^{(i)})}, \quad i = 1, \ldots, m. \tag{21}$$

3. SIR the weighted samples $\{(\mathbf{x}_t^{(i)}, w_t^{(i)})\}_{i=1}^m$ to get $(\mathbf{x}_t^{(*1)}, \ldots, \mathbf{x}_t^{(*m)})$ and replace $\mathbf{x}_t^{(i)} \leftarrow \mathbf{x}_t^{(*i)}$, $i = 1, \ldots, m$.

Due to the resampling step 3, $\mathbf{x}_t^{(i)} = \mathbf{x}_t^{(*i)} \sim \pi_t$ approximately.

Remark 2. There are a few key differences between SIS and sequential Monte Carlo: (i) The distribution of the partial sample $\mathbf{x}_t = (x_1, \ldots, x_t)$ is different. In SIS, $\mathbf{x}_t \sim g(\mathbf{x}_t)$ while in sequential Monte Carlo $\mathbf{x}_t \sim \pi_t$ due to resampling. (ii) Accordingly, their importance weights are also calculated differently: compare (13) and (21). The u_t in (12) is the weight $w_t^{(i)}$ in (21). (iii) In the end, we obtain weighted samples $(\mathbf{x}, w)$ from SIS and need to use weighted average, e.g. (3), to construct estimates. In sequential Monte Carlo, since $\mathbf{x} \sim \pi$ after the last step, we simply use sample averages of $\mathbf{x}^{(i)}$ to estimate expectations, i.e. our estimate of $\mathbb{E}_\pi[h(\mathbf{X})]$ is

$$\frac{1}{m}\sum_{i=1}^{m} h(\mathbf{x}^{(i)}).$$

Continuing Example 9, under the same choices of g_t and π_t as in (17) and (18), the importance weight in sequential Monte Carlo will be the same as u_t in (19), so we have

$$w_t^{(i)} = \exp\left[\{\mathrm{RSS}(\mathbf{z}_{t-1}^{(i)}) - \mathrm{RSS}(\mathbf{z}_t^{(i)})\}/2\right], \quad i = 1, \dots, m.$$

Therefore we have the following implementation for this problem:

For $t = 1, \dots, d$:

1. Draw $z_t^{(i)} \sim \exp(-\lambda z_t^{(i)})$ and put $\mathbf{z}_t^{(i)} = (\mathbf{z}_{t-1}^{(i)}, z_t^{(i)})$ for $i = 1, \dots, m$.
2. Calculate weights:

$$w_t^{(i)} = \exp\left[\{\mathrm{RSS}(\mathbf{z}_{t-1}^{(i)}) - \mathrm{RSS}(\mathbf{z}_t^{(i)})\}/2\right], \quad i = 1, \dots, m.$$

3. Resample the weighted samples $\{(\mathbf{z}_t^{(i)}, w_t^{(i)})\}_{i=1}^m$ to get $(\mathbf{z}_t^{(*1)}, \dots, \mathbf{z}_t^{(*m)})$ and replace $\mathbf{z}_t^{(i)} \leftarrow \mathbf{z}_t^{(*i)}$, $i = 1, \dots, m$.

R code for the resampling step: Suppose `w` is a numerical vector of length m that stores the current importance weights $\{w_t^{(i)} : i = 1, \dots, m\}$ and `x` is an $m \times t$ matrix that stores the current partial samples $\{\mathbf{x}_t^{(i)} : i = 1, \dots, m\}$, i.e. each row of `x` corresponds to a partial sample.

```
s=sample(1:m,size=m,replace=TRUE,prob=w);
y=x[s,];
```

Then `y` stores the partial samples after resampling $\{\mathbf{x}_t^{(*i)} : i = 1, \dots, m\}$.