Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

Department of mathematics
KTH Royal Institute of Technology
johawes@kth.se

Lecture 3
Importance sampling
24 March 2017

Plan of today's lecture

1. Last time
2. Rejection sampling
3. Importance sampling (IS)
4. Self-normalized IS

Outline

Last time: the delta method

- For a given estimand τ one is often interested in estimating $\varphi(\tau)$ for some function $\varphi : \mathbb{R} \to \mathbb{R}$.
- For this purpose, we simply used the plug-in estimator $\varphi(\tau_N)$ of $\varphi(\tau)$.
- The estimator $\varphi(\tau_N)$ is generally biased for finite N; indeed, under suitable assumptions on φ it holds that

$$\mathbb{E}\left(\varphi(\tau_N) - \varphi(\tau)\right) = \frac{\varphi''(\tau)\sigma^2(\phi)}{2N} + O(N^{-3/2}).$$

- In addition, one may establish the CLT

$$\sqrt{N}\left(\varphi(\tau_N) - \varphi(\tau)\right) \stackrel{\text{d.}}{\longrightarrow} \mathrm{N}(0, \varphi'(\tau)^2\sigma^2(\phi)), \quad \text{as } N \to \infty.$$

Last time: MC output analysis (Ch. 4)

- We used the CLT

$$\sqrt{N}\,(\tau_N - \tau) \xrightarrow{\text{d.}} \mathsf{N}(0, \sigma^2(\phi))$$

to target τ by the approximate confidence interval

$$\mathsf{I}_\alpha = \left(\tau_N \pm \lambda_{\alpha/2}\frac{\sigma(\phi)}{\sqrt{N}}\right).$$

- Moreover, the delta method provides the approximate confidence interval

$$\mathsf{I}_\alpha = \left(\varphi(\tau_N) \pm \lambda_{\alpha/2}|\varphi'(\tau_N)|\frac{\sigma(\phi)}{\sqrt{N}}\right)$$

for $\varphi(\tau)$.

Last time: pseudo-random number generation (Ch. 3)

- We discussed (briefly) how to generate pseudo-random uniformly distributed numbers (U_n) using the linear congruential generator

$$U_n = (a \cdot U_{n-1} + c) \mod m.$$

- Having at hand such $\mathrm{U}(0,1)$-distributed numbers U, we also looked at how to generate pseudo-random numbers X from an arbitrary distribution F by means of the inversion method, i.e., by letting

$$X = F^{\leftarrow}(U) = \inf\{x \in \mathbb{R} : F(x) \geq U\}.$$

Last time: the inversion method

- Then the following holds true:

Theorem (the inverse method)

The output X of the algorithm above has distribution function F.

- If F is continuous and strictly increasing, then $F^{\leftarrow} = F^{-1}$.
- The method is limited to cases where
  - we want to generate univariate random numbers and
  - the generalized inverse $F^{\leftarrow}$ is easy to evaluate (which is far from always the case).

Outline

Rejection sampling

- The inversion method looks promising, but what do we do if, e.g., $f(x) \propto \exp(\cos^2(x))$, $x \in (-\pi/2, \pi/2)$? Here we cannot find an inverse and do not even know the normalizing constant. ☹
- This is a very common situation in statistics.
- The following (somewhat magic!) algorithm saves the day. Let g be a density or probability function on the same state space X ($\subseteq \mathbb{R}^d$) as f and assume that there exists a constant $K < \infty$ such that

$$f(x) \leq Kg(x) \quad \forall x \in \mathsf{X}.$$

Rejection sampling

- The inversion method looks promising, but what do we do if, e.g., $f(x) \propto \exp(\cos^2(x))$, $x \in (-\pi/2, \pi/2)$? Here we cannot find an inverse and do not even know the normalizing constant. ☹
- This is a very common situation in statistics.
- The following (somewhat magic!) algorithm saves the day. Let g be a density or probability function on the same state space X ($\subseteq \mathbb{R}^d$) as f and assume that there exists a constant $K < \infty$ such that

$$f(x) \leq Kg(x) \quad \forall x \in \mathsf{X}.$$

Rejection sampling (cont.)

- We proceed as follows:

set $\texttt{accepted} \leftarrow \texttt{false}$;
while $\texttt{accepted} = \texttt{false}$ **do**
    draw $X^* \sim g$;
    draw $U \sim \mathrm{U}(0,1)$;
    if $U \leq \frac{f(X^*)}{Kg(X^*)}$ **then**
        $X \leftarrow X^*$;
        $\texttt{accepted} \leftarrow \texttt{true}$;
    end
end
return X

Rejection sampling (cont.)

- The following holds true:

Theorem (rejection sampling)

The output X of the rejection sampling algorithm has density function f.

- Moreover:

Theorem

The expected number of trials needed before acceptance is K.

Consequently, the upper bound K should be chosen as small as possible.

Rejection sampling (cont.)

- The following holds true:

Theorem (rejection sampling)

The output X of the rejection sampling algorithm has density function f.

- Moreover:

Theorem

The expected number of trials needed before acceptance is K.

Consequently, the upper bound K should be chosen as small as possible.

Example

- We wish to simulate from $f(x) = \exp(\cos^2(x))/c$, $x \in (-\pi/2, \pi/2)$, where $c = \int_{-\pi/2}^{\pi/2} \exp(\cos^2(z))\, dz$ is the unknown normalizing constant.
- However, since for all $x \in (-\pi/2, \pi/2)$,

$$f(x) = \frac{\exp(\cos^2(x))}{c} \leq \frac{e}{c} = \underbrace{\frac{e\pi}{c}}_{K} \times \underbrace{\frac{1}{\pi}}_{g},$$

where g is the density of $U(-\pi/2, \pi/2)$, we may use rejection sampling where a candidate $X^* \sim U(-\pi/2, \pi/2)$ is accepted if

$$U \leq \frac{f(X^*)}{Kg(X^*)} = \frac{\exp(\cos^2(X^*))/c}{e/c} = \exp(\cos^2(X^*) - 1).$$

Example

- We wish to simulate from $f(x) = \exp(\cos^2(x))/c$, $x \in (-\pi/2, \pi/2)$, where $c = \int_{-\pi/2}^{\pi/2} \exp(\cos^2(z))\,dz$ is the unknown normalizing constant.
- However, since for all $x \in (-\pi/2, \pi/2)$,

$$f(x) = \frac{\exp(\cos^2(x))}{c} \leq \frac{e}{c} = \underbrace{\frac{e\pi}{c}}_{K} \times \underbrace{\frac{1}{\pi}}_{g},$$

where g is the density of $\mathrm{U}(-\pi/2, \pi/2)$, we may use rejection sampling where a candidate $X^* \sim \mathrm{U}(-\pi/2, \pi/2)$ is accepted if

$$U \leq \frac{f(X^*)}{Kg(X^*)} = \frac{\exp(\cos^2(X^*))/c}{e/c} = \exp(\cos^2(X^*) - 1).$$

Example

- In MATLAB:

```
prob = @(x) exp((cos(x))^2 - 1);
trial = 1;
accepted = false;
while ~accepted,
    Xcand = - pi/2 + pi*rand;
    if rand < prob(Xcand),
        accepted = true;
        X = Xcand;
    else
        trial = trial + 1;
    end
end
```

Example

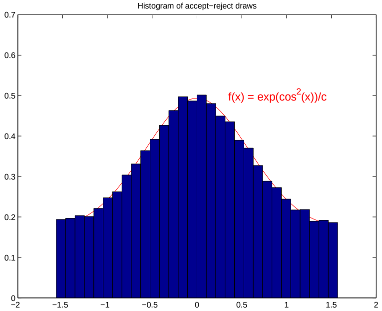
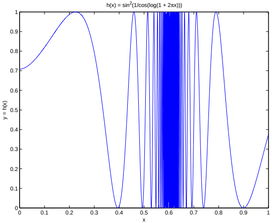


Figure: Plot of a histogram of 20,000 accept-reject draws together with the true density. The average number of trials was 1.5555. In this case the expected number is $\pi e/c = 1.5503$.

Outline

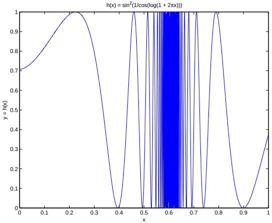
Advantages of the MC method

- The MC method
  - is more efficient than deterministic methods in high dimensions,
  - does generally not require knowledge of the normalizing constant of a density f for computing expectations, and
  - handles efficiently "strange" integrands ϕ that may cause problems for deterministic methods.



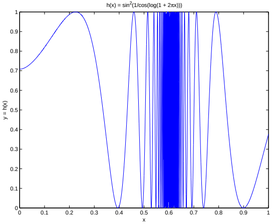
Advantages of the MC method

- The MC method
  - is more efficient than deterministic methods in high dimensions,
  - does generally not require knowledge of the normalizing constant of a density f for computing expectations, and
  - handles efficiently "strange" integrands ϕ that may cause problems for deterministic methods.



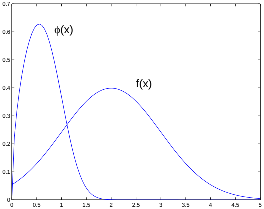
Advantages of the MC method

- The MC method
  - is more efficient than deterministic methods in high dimensions,
  - does generally not require knowledge of the normalizing constant of a density *f* for computing expectations, and
  - handles efficiently “strange” integrands ϕ that may cause problems for deterministic methods.



Problems with MC integration

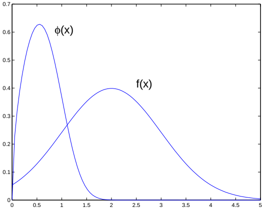
- OK, MC integration looks promising. We may however run into problems if
  - it is hard to sample from f or
  - if the integrand ϕ and the density f are dissimilar; in this case we will end of with a lot of draws where the integrand is small, and consequently only a few draws will contribute to the estimate. This gives a large variance.



- Here importance sampling is useful!

Problems with MC integration

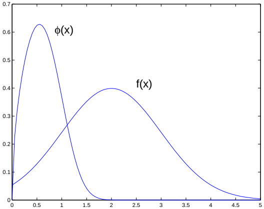
- OK, MC integration looks promising. We may however run into problems if
  - it is hard to sample from f or
  - if the integrand ϕ and the density f are dissimilar; in this case we will end of with a lot of draws where the integrand is small, and consequently only a few draws will contribute to the estimate. This gives a large variance.



- Here importance sampling is useful!

Problems with MC integration

- OK, MC integration looks promising. We may however run into problems if
  - it is hard to sample from f or
  - if the integrand ϕ and the density f are dissimilar; in this case we will end of with a lot of draws where the integrand is small, and consequently only a few draws will contribute to the estimate. This gives a large variance.



- Here importance sampling is useful!

Importance sampling (IS, Ch. 4.1)

- The basis of importance sampling is to take an instrumental density g on X such that $g(x) = 0 \Rightarrow f(x) = 0$ and rewrite the expectation as

$$
\begin{aligned}
\tau = \mathbb{E}_f\left(\phi(X)\right) &= \int_{\mathsf{X}} \phi(x) f(x)\, dx = \int_{f(x)>0} \phi(x) f(x)\, dx \\
&= \int_{g(x)>0} \phi(x) \frac{f(x)}{g(x)} g(x)\, dx = \mathbb{E}_g\left(\phi(X)\frac{f(X)}{g(X)}\right) \\
&= \mathbb{E}_g\left(\phi(X)\omega(X)\right),
\end{aligned}
$$

where we have defined the importance weight function

$$
\omega : \{x \in \mathsf{X} : g(x) > 0\} \ni x \mapsto \frac{f(x)}{g(x)}.
$$

Importance sampling (cont.)

- Now estimate $\tau = \mathbb{E}_g(\phi(X)\omega(X))$ using standard MC:

  for $i = 1 \to N$ **do**
  draw $X^i \sim g$;
  end
  set $\tau_N \leftarrow \sum_{i=1}^{N} \phi(X^i)\omega(X^i)/N$;
  return τ_N

- Here, trivially,

$$\mathbb{V}(\tau_N) = \frac{1}{N}\mathbb{V}_g(\phi(X)\omega(X)),$$

  and we should thus aim at choosing g so that the function $x \mapsto \phi(x)\omega(x)$ is close to constant in the support of g. This gives a minimal variance.

Importance sampling (cont.)

- Now estimate $\tau = \mathbb{E}_g(\phi(X)\omega(X))$ using standard MC:

  for $i = 1 \to N$ **do**

  draw $X^i \sim g$;

  end

  set $\tau_N \leftarrow \sum_{i=1}^N \phi(X^i)\omega(X^i)/N$;

  return τ_N

- Here, trivially,

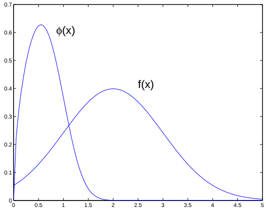
$$\mathbb{V}(\tau_N) = \frac{1}{N}\mathbb{V}_g(\phi(X)\omega(X)),$$

  and we should thus aim at choosing g so that the function $x \mapsto \phi(x)\omega(x)$ is close to constant in the support of g. This gives a minimal variance.

Example: a tricky normal expectation

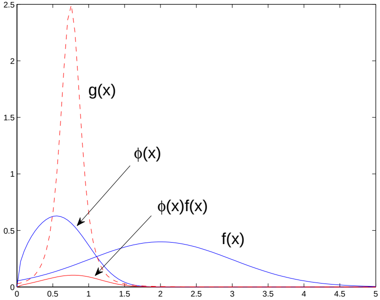
- Let X be N(2, 1)-distributed and consider

$$\tau = \mathbb{E}\left(\mathbb{1}_{X\geq 0}\sqrt{X}\exp(-X^3)\right) = \int \underbrace{\mathbb{1}_{x\geq 0}\sqrt{x}\exp(-x^3)}_{=\phi(x)}\underbrace{\mathrm{N}(x;2,1)}_{=f(x)}\,dx,$$



Example: a tricky normal expectation (cont.)

- Thus, standard MC will lead to a waste of computational power. Better is to use IS with g being a scale-location-transformed student's t-distribution with, say, $\nu = 3$ degrees of freedom:



Example: A tricky normal expectation (cont.)

- The standard deviation is estimated via the full width at half maximum (FWHM) for a Gaussian bell:

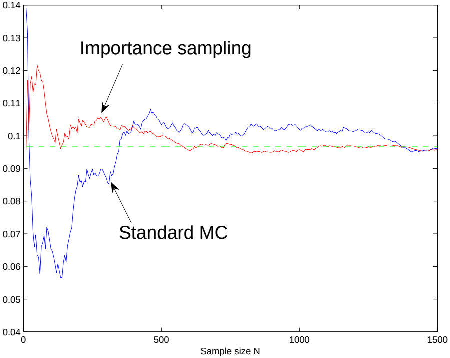
$$\text{FWHM} = \text{standard deviation} \times 2\sqrt{2\log 2}.$$

- In MATLAB:

```
phi = @(x) (x >= 0).*sqrt(x).*exp(- x.^3);
mu = 0.75;
sigma = 1.2/(2*sqrt(2*log(2)));
v = 3;
s = sigma*sqrt((v - 2)/v);
X = s*trnd(v,1,N) + mu;
omega = @(x) normpdf(x,2,1)./(tpdf((x - mu)/s,v)/s);
tau = mean(phi(X).*omega(X));
```

Example: A tricky normal expectation (cont.)

- Executing the IS algorithm and standard MC in parallel yields the following:



Outline

Self-normalized IS (Ch. 4.1.1)

- Often $f(x)$ is known only up to a normalizing constant $c > 0$, i.e. $f(x) = z(x)/c$, where we can evaluate $z(x) = cf(x)$ but not $f(x)$. Then, as before, letting now $\omega(x) = cf(x)/g(x) = z(x)/g(x)$,

$$\tau = \mathbb{E}_f(\phi(X)) = \int_X \phi(x)f(x)\,dx = \frac{c\int_{f(x)>0}\phi(x)f(x)\,dx}{c\int_{f(x)>0}f(x)\,dx}$$

$$= \frac{\int_{g(x)>0}\phi(x)\frac{cf(x)}{g(x)}g(x)\,dx}{\int_{g(x)>0}\frac{cf(x)}{g(x)}g(x)\,dx} = \frac{\int_{g(x)>0}\phi(x)\omega(x)g(x)\,dx}{\int_{g(x)>0}\omega(x)g(x)\,dx}$$

$$= \frac{\mathbb{E}_g(\phi(X)\omega(X))}{\mathbb{E}_g(\omega(X))}.$$

Self-normalized IS (Ch. 4.1.1)

- Often $f(x)$ is known only up to a normalizing constant $c > 0$, i.e. $f(x) = z(x)/c$, where we can evaluate $z(x) = cf(x)$ but not $f(x)$. Then, as before, letting now $\omega(x) = cf(x)/g(x) = z(x)/g(x)$,

$$\tau = \mathbb{E}_f(\phi(X)) = \int_{\mathsf{X}} \phi(x)f(x)\,dx = \frac{c\int_{f(x)>0} \phi(x)f(x)\,dx}{c\int_{f(x)>0} f(x)\,dx}$$

$$= \frac{\int_{g(x)>0} \phi(x)\frac{cf(x)}{g(x)}g(x)\,dx}{\int_{g(x)>0} \frac{cf(x)}{g(x)}g(x)\,dx} = \frac{\int_{g(x)>0} \phi(x)\omega(x)g(x)\,dx}{\int_{g(x)>0} \omega(x)g(x)\,dx}$$

$$= \frac{\mathbb{E}_g(\phi(X)\omega(X))}{\mathbb{E}_g(\omega(X))}.$$

Self-normalized IS (cont.)

- Since $\omega(x) = z(x)/g(x)$ can be evaluated for each x, we may now estimate the ratio

$$\tau = \frac{\mathbb{E}_g(\phi(X)\omega(X))}{\mathbb{E}_g(\omega(X))}$$

  by solving one MC problem for the numerator and another for the denominator.
- Note that since $c = \mathbb{E}_g(\omega(X))$, this approach provides, as a by-product, an estimate also of the normalizing constant c.

Self-normalized IS (cont.)

- Since $\omega(x) = z(x)/g(x)$ can be evaluated for each x, we may now estimate the ratio

$$\tau = \frac{\mathbb{E}_g(\phi(X)\omega(X))}{\mathbb{E}_g(\omega(X))}$$

by solving one MC problem for the numerator and another for the denominator.

- Note that since $c = \mathbb{E}_g(\omega(X))$, this approach provides, as a by-product, an estimate also of the normalizing constant c.

Example

- We reconsider the density

$$f(x) = \exp(\cos^2(x))/c, \quad x \in (-\pi/2, \pi/2),$$

treated previously and estimate its variance as well as the normalizing constant $c > 0$ using self-normalized IS.

- Let the instrumental distribution g be the uniform distribution $\mathrm{U}(-\pi/2, \pi/2)$.

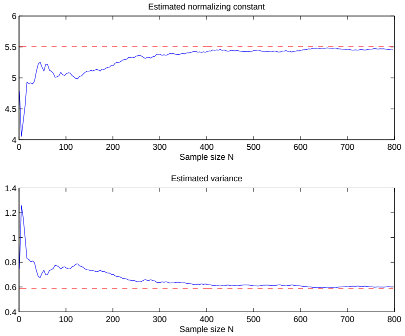
Example (cont.)

- In MATLAB:

```
z = @(x) exp(cos(x).^2);
X = - pi/2 + pi*rand(1,N);
omega = @(x) pi*z(x);
tau = cumsum(X.^2.*omega(X))./cumsum(omega(X));
c = cumsum(omega(X))./(1:N);
subplot(2,1,1);
plot(1:N,c);
subplot(2,1,2);
plot(1:N,tau);
```

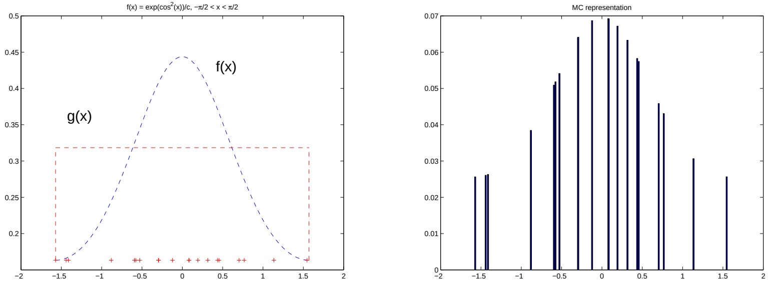
Example (cont.)

- Plotting the outcome:



IS $\Rightarrow$ representation of f

The weighted sample $(X^i, \omega(X^i))$ can be viewed as a discrete MC representation of the target distribution f.



$$f(x) \overset{\text{IS}}{\Longrightarrow} (X^i, \omega(X^i))$$

E1

E1 comprises problems on

- random number generation (transformation-based methods, the inverse method, rejection sampling),
- MC/IS (power production of a wind turbine),
- Plug-in MC estimators and the delta method.