CMPE 58N - Lecture 2
Monte Carlo methods

Random Number Generation, Rejection, Importance Sampling

Department of Computer Engineering,

Boğaziçi University, Istanbul, Turkey

Instructor: A. Taylan Cemgil

Fall 2009

Outline

- Random Number Generation
  - Pseudo-random numbers
  - Source of random bits: Generating Bernoulli Random Variables
- Inversion
- Transformation
- Rejection Sampling
- Importance Sampling

Random Number Generation

- Physical methods
  - throw dice, flip coins, shuffle playing cards, roulette wheel
  - thermal noise in Zener diodes or other analog circuits
  - Listen to atmospheric noise (www.Random.org)
  - Run a hash function against a frame of a video stream
  - . . .
- A random number generator (deterministic computation) to obtain numbers that “look” random
  - Efficient
  - Repeatable (seeds) – good for debugging

Pseudo-random number generator

- Linear congruential generator

$$\begin{aligned} Z_i &= (aZ_{i-1} + b) \mod M \\ X_i &= Z_i / M \end{aligned}$$

main flaw: the crystalline nature

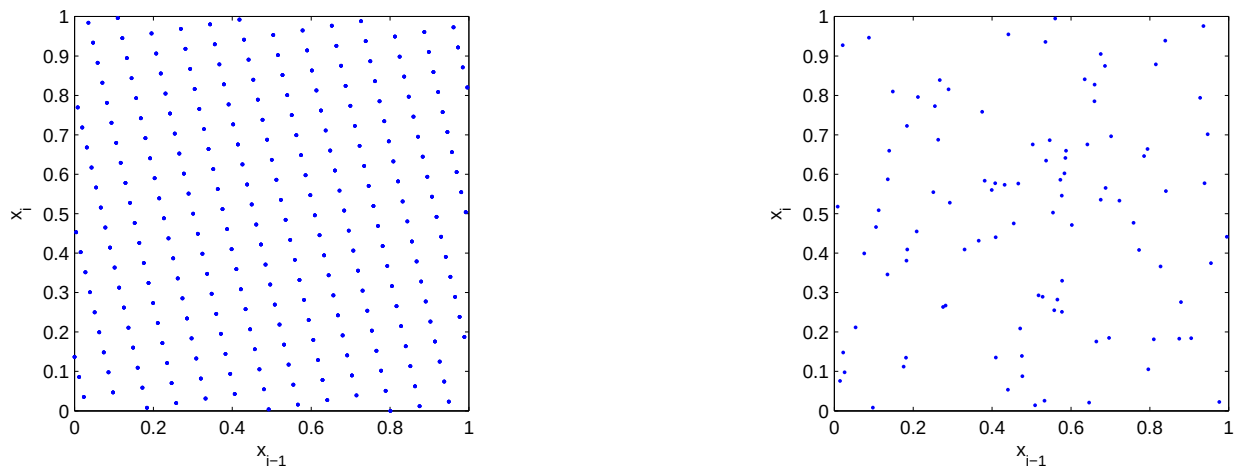


Figure: (left) $a = 81$; $b = 35$; $M = 256$;, (right) Matlab's `rand`

Remarks

A poor design which was popular during 1970's

$a = 2^{16} + 3; b = 0; M = 2^{31}$

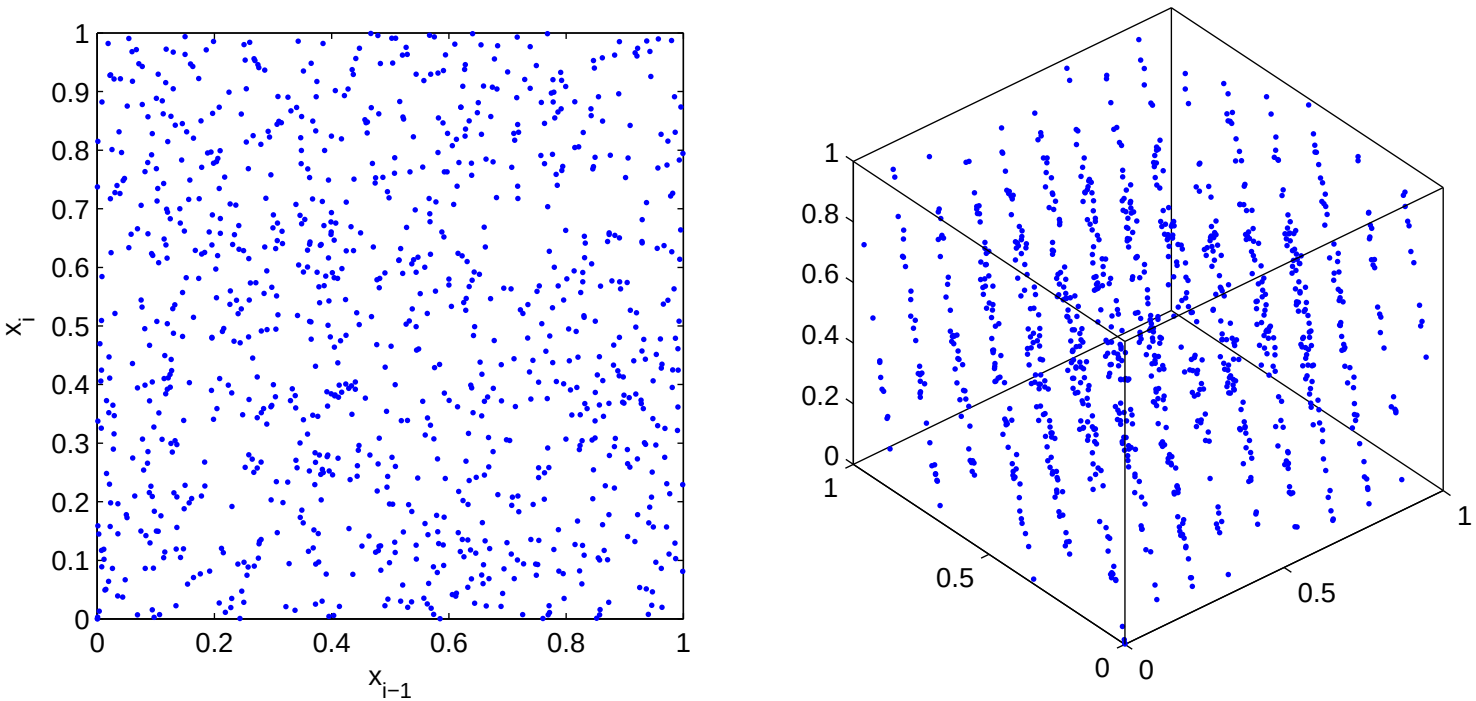


Figure: (left) (X_{i-1}, X_i), (right) (X_{i-2}, X_{i-1}, X_i)

Bernoulli Random Variables

- $x \in \{0, 1\}$

$$x \sim \mathcal{BE}(x; p) = p^x (1 - p)^{(1-x)},$$

- How to sample from a Bernoulli distribution on a computer given $p \in \mathbb{R}$ using samples from the uniform distribution $u \sim \mathcal{U}(u; 0, 1)$?

Bernoulli Random Variables

- $x \in \{0, 1\}$

$$x \sim \mathcal{BE}(x; p) = p^x (1-p)^{(1-x)},$$

$$u \sim \mathcal{U}(u; 0, 1)$$

$$x = u < p$$

Note that this is an idealisation as we can not represent irrational numbers on a computer.

The Knuth-Yao algorithm

- How to sample exactly from a Bernoulli distribution $x \sim \mathcal{BE}(x; p)$, $x \in \{0, 1\}$ on a computer given $p \in \mathbb{R}$ using a random bit source $\omega \sim \mathcal{BE}(\omega; 1/2)$?

The Knuth-Yao algorithm

Represent p in binary $p = 0.p_1 p_2 p_3 \ldots$

$i \leftarrow 0$

Repeat

$i \leftarrow i + 1$

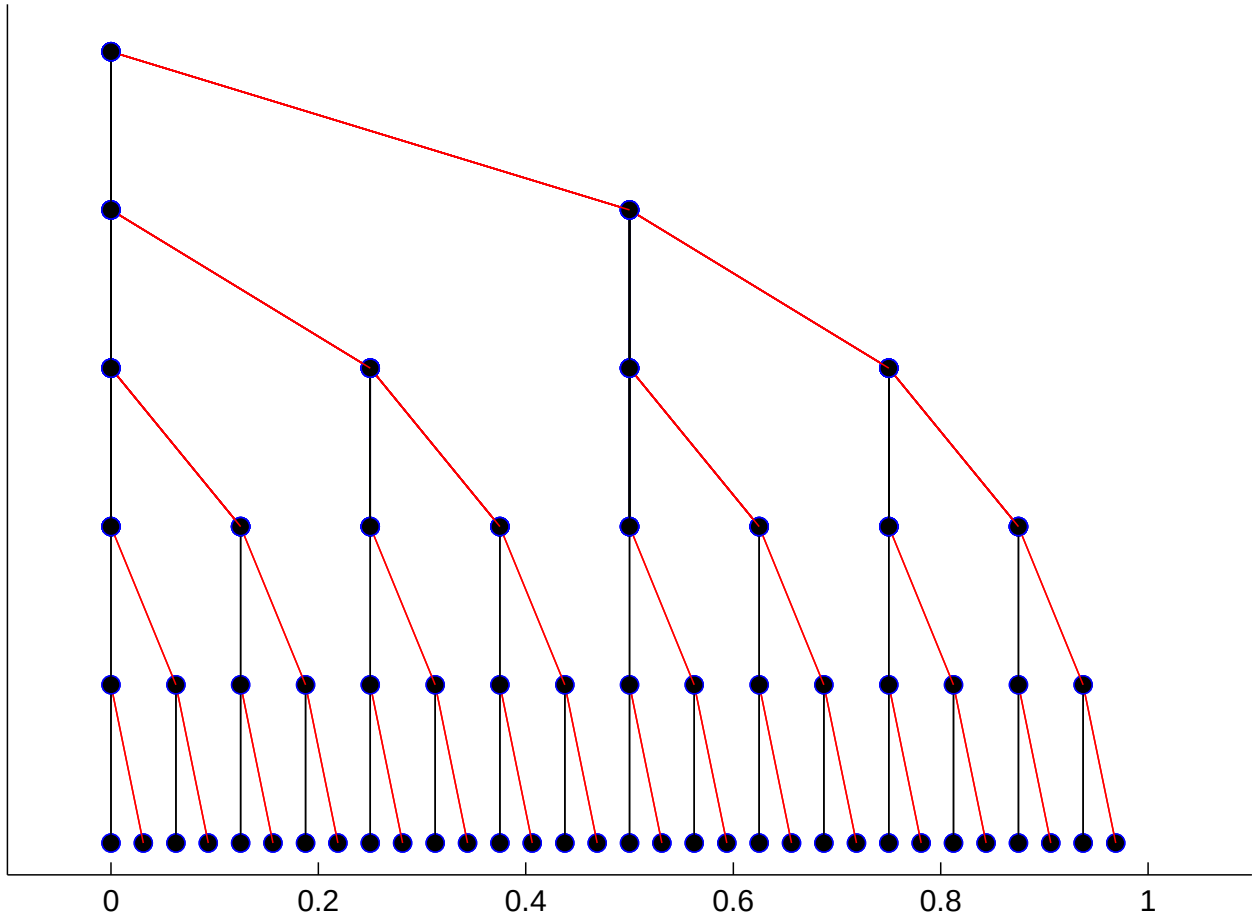
$\omega \sim \mathcal{BE}(\omega; 1/2)$

Until $\omega \neq p_i$

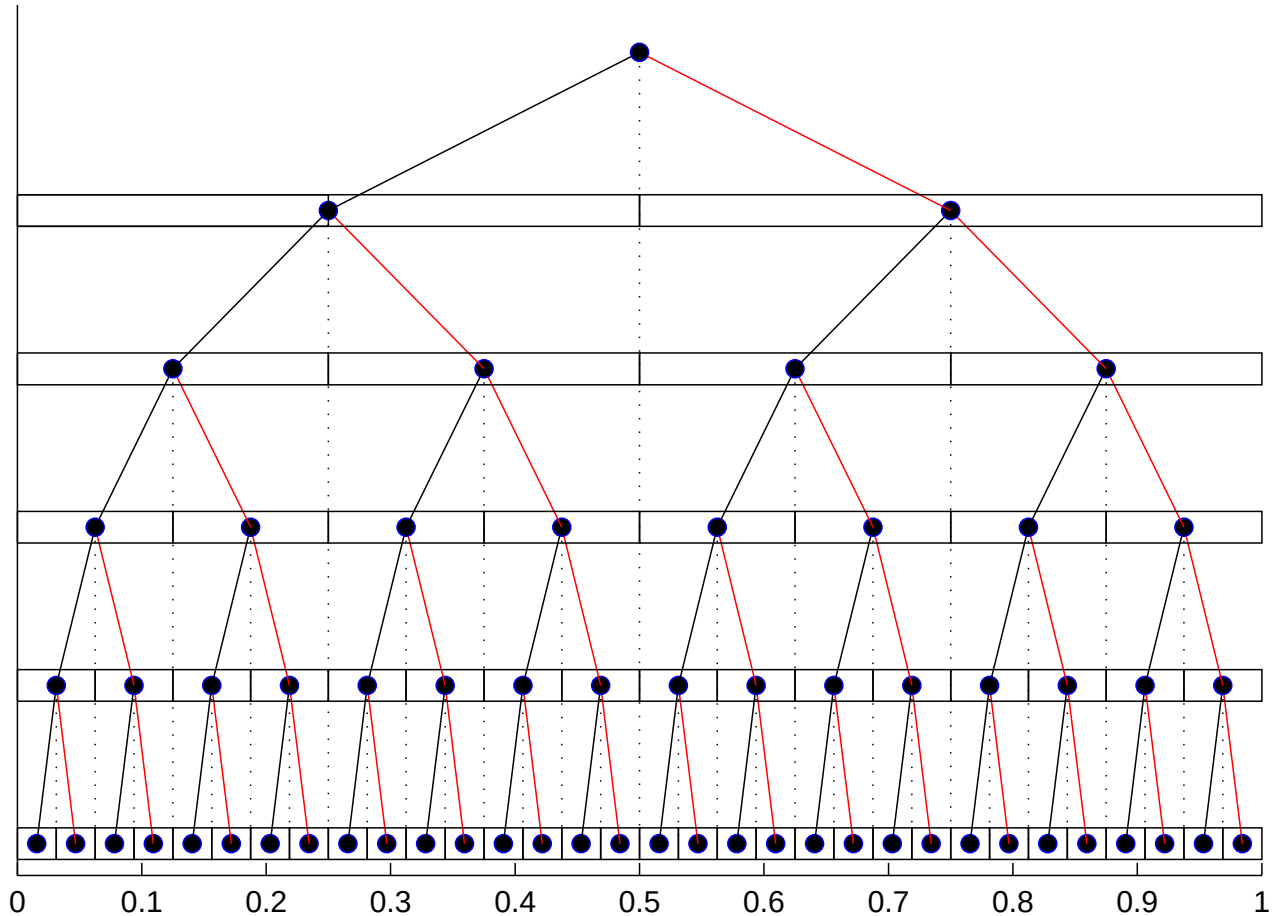
$x \leftarrow \omega < p_i$

See Luc Devroye, Non uniform random variate generation, available online, Ch. 15

The Knuth-Yao algorithm (cont.)



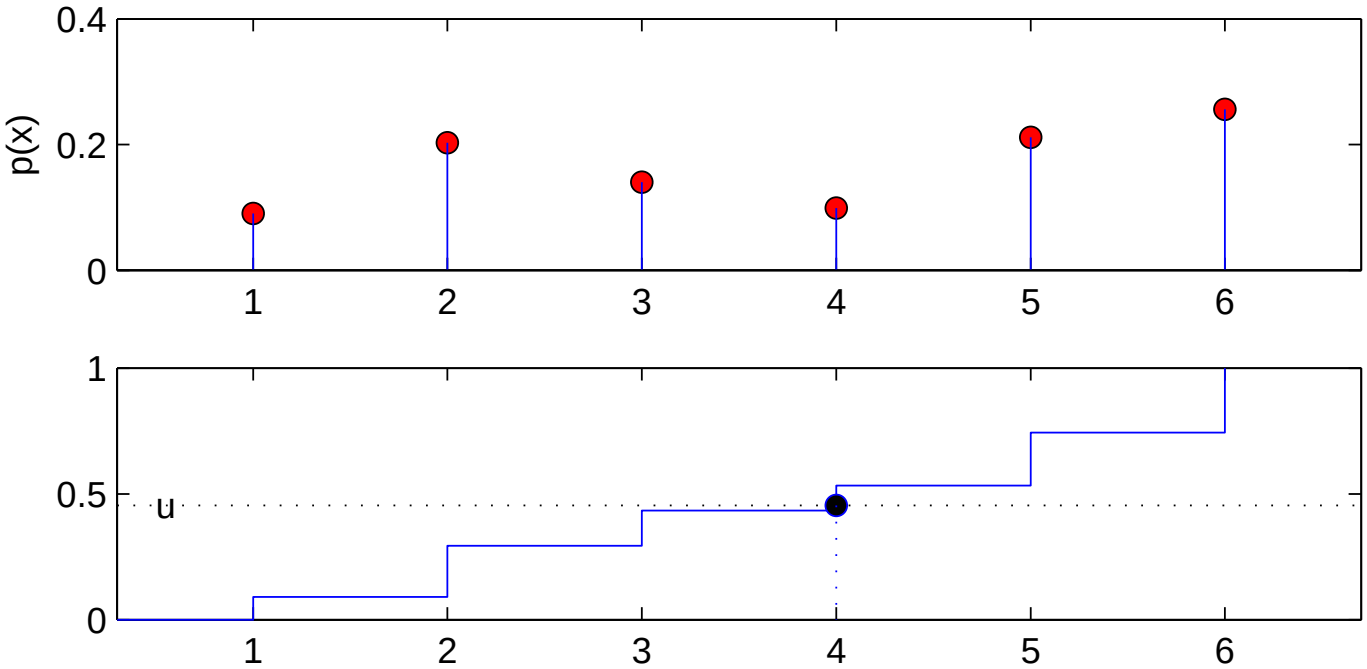
Alternative view



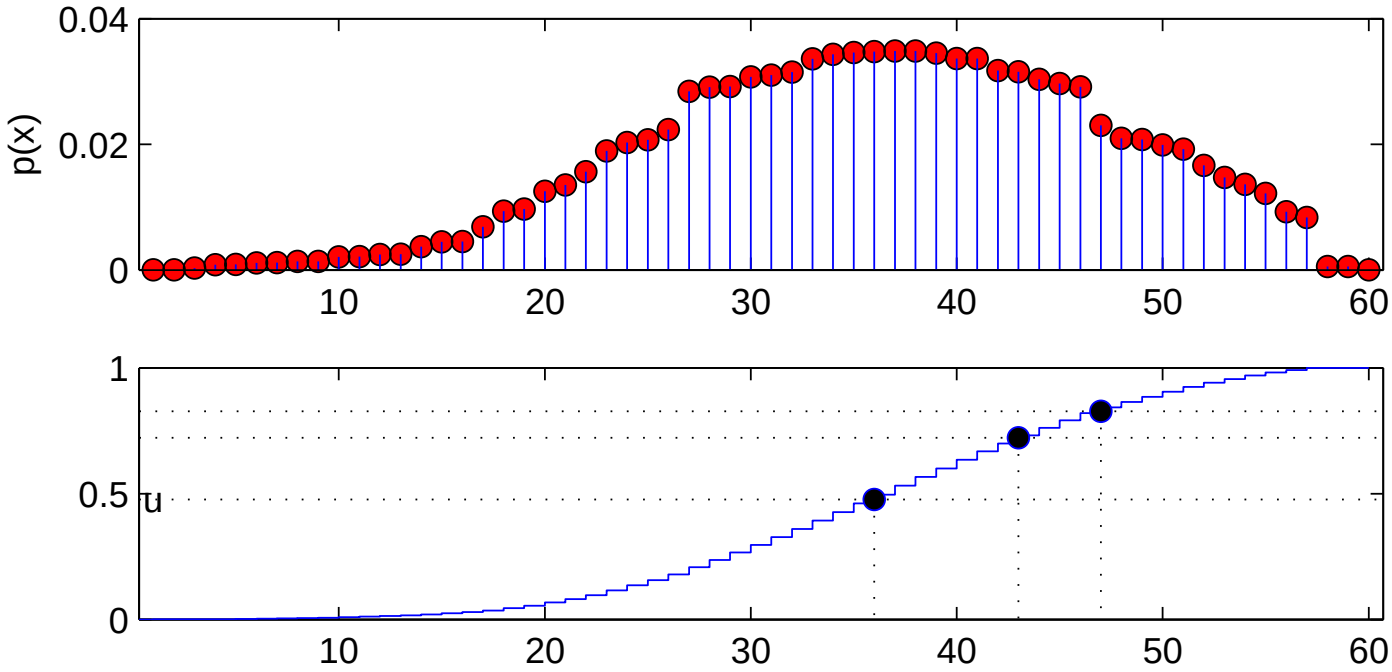
Generating random numbers from given distributions

- Inversion
- Transformation
- Rejection
- Reweighting – Importance sampling

Inversion



Inversion (cont.)



Generalised Inverse

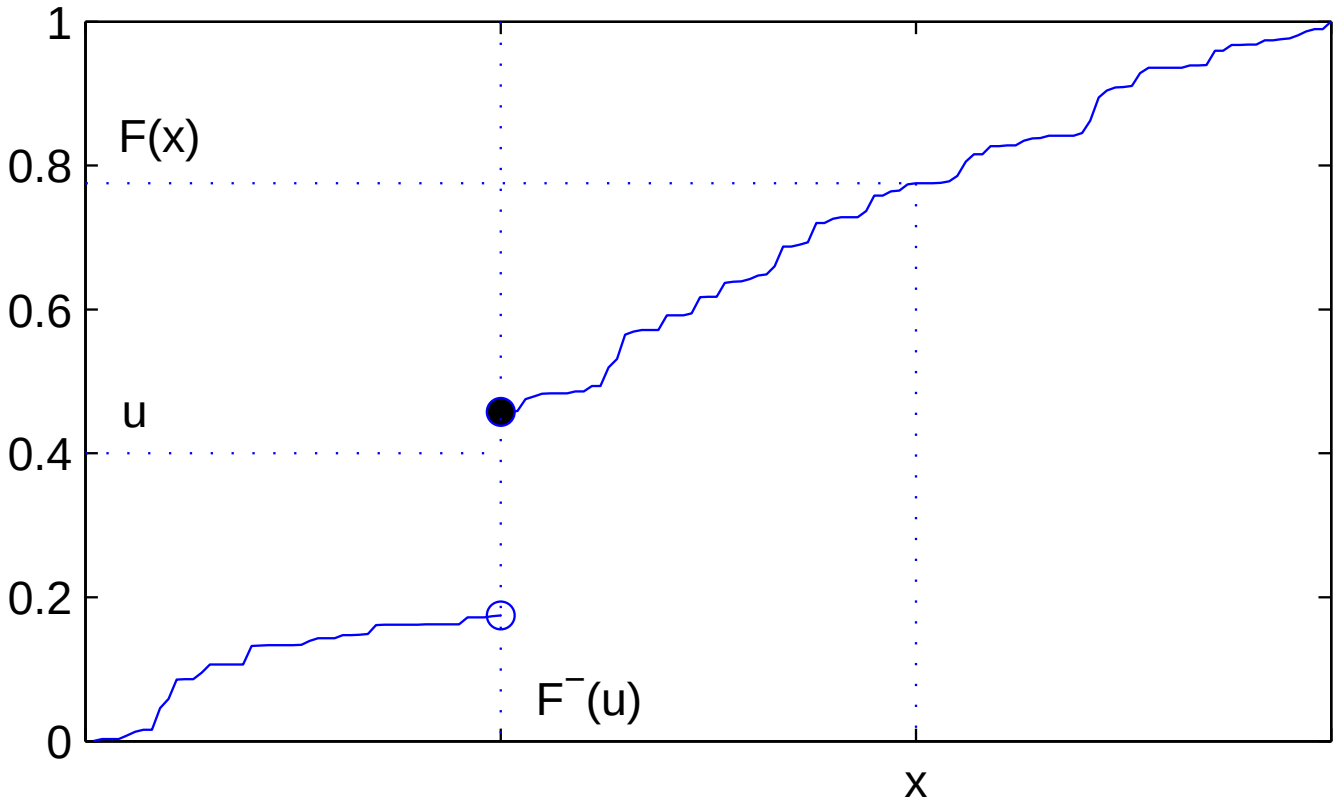
- Define the cumulative distribution function

$$\Pr\{X(\omega) \leq x\} \quad = \quad F(x)$$

- Generalised Inverse CDF

$$F^{-}(u) \quad \equiv \quad \inf\{x : F(x) \geq u\}$$

Generalised Inverse (cont.)



Example: Exponential Distribution

$$
\begin{aligned}
\mathcal{E}(x;\lambda) &= \lambda \exp(-\lambda x) \\
F(x) &= \int_0^x \lambda \exp(-\lambda \tau) = -\exp(-\lambda x) + 1 \\
u &= F(x) = 1 - \exp(-\lambda x) \\
x &= -\log(1-u)/\lambda = F^{-}(u)
\end{aligned}
$$

Or

$$
\begin{aligned}
u &\sim \mathcal{U}[0,1] \\
1-u &= u' \sim \mathcal{U}[0,1] \\
x &= -\log(u')/\lambda
\end{aligned}
$$

Transform Method

The generalised inverse of the CDF is just one possible transformation method

- Generate $\xi_i \sim p(\xi_i)$
- Set $x_j = h_j(\xi_1, \ldots, \xi_N)$
- Example: Box-Müller Method for generating Gaussian RV's
  - Generate a polar coordinate $(\sqrt{a}, \theta)$

$$\begin{aligned} \theta &\sim \mathcal{U}(\theta; 0, 2\pi) \\ a &\sim \mathcal{E}(a; 1/2) \end{aligned}$$

  - Transform into cartesian coordinates

$$\begin{aligned} x_1 &= \sqrt{a}\cos(\theta) && (\sim \mathcal{N}(0,1)) \\ x_2 &= \sqrt{a}\sin(\theta) && (\sim \mathcal{N}(0,1)) \end{aligned}$$

Change of Variables

Example:

- We are given X with density $f_X(x)$

$$Y = aX + b$$

- We find the CDF of Y analytically as follows

$$\Pr\{Y \leq y\} = \Pr\{aX + b \leq y\} = \begin{cases} \Pr\{X \leq (y-b)/a\}, & a > 0 \\ \Pr\{X \geq (y-b)/a\}, & a < 0 \end{cases}$$

- The density $f_Y(y)$ is the derivative w.r.t. y

$$f_Y(y) = \frac{1}{|a|} f_X((y-b)/a)$$

Change of variables

More general case (see Gri. & Sti. pp108):

- Define a one-to-one mapping

$$g : (x_1, x_2) \quad \rightarrow (y_1, y_2)$$

- Invert the mapping, i.e. find

$$\begin{aligned} x_1 &= x_1(y_1, y_2) \\ x_2 &= x_2(y_1, y_2) \end{aligned}$$

- Define the *Jacobian* determinant

$$J = \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_2/\partial y_1 \\ \partial x_1/\partial y_2 & \partial x_2/\partial y_2 \end{vmatrix} = J(y_1, y_2)$$

Change of variables (cont.)

- The density of the transformed variable is

$$f_Y(y_1, y_2) = f_X(x_1(y_1, y_2), x_2(y_1, y_2))|J(y_1, y_2)|$$

- Keep in mind

$$f_Y(y)dy = f_X(x(y))\left|\frac{dx}{dy}\right|dy$$

Example: Box-Muller

- Generate

$$\begin{aligned} x_1 &\sim \mathcal{E}(x_1; 1/2) = \frac{1}{2}\exp(-x_1/2) \\ x_2 &\sim \mathcal{U}(x_2; 0, 2\pi) = \frac{1}{2\pi}\mathbb{I}\{0 \leq x_2 \leq 2\pi\} \end{aligned}$$

- Transform

$$\begin{aligned} y_1 &= \sqrt{x_1}\cos(x_2) \\ y_2 &= \sqrt{x_1}\sin(x_2) \end{aligned}$$

- Find the inverse mapping (the difficult bit)

$$\begin{aligned} y_1^2 + y_2^2 &= x_1 \\ \arctan(y_2/y_1) &= x_2 \end{aligned}$$

Example: Box-Muller (cont.)

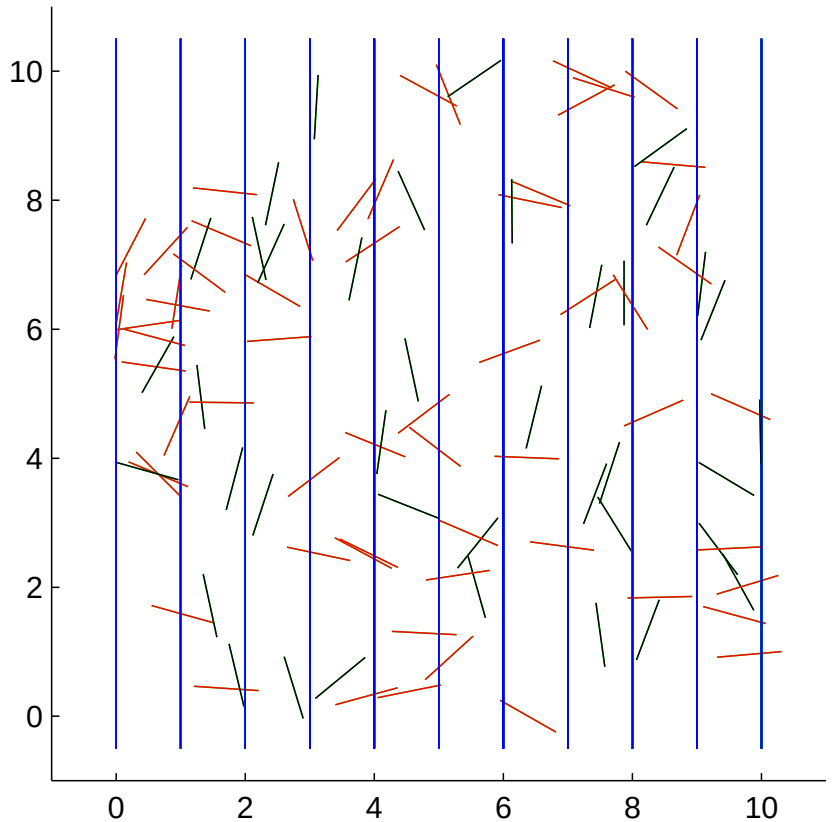
- Find the Jacobian determinant

$$J = \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_2/\partial y_1 \\ \partial x_1/\partial y_2 & \partial x_2/\partial y_2 \end{vmatrix} = \begin{vmatrix} 2y_1 & \frac{1}{1+(y_2/y_1)^2}\frac{-y_2}{y_1^2} \\ 2y_2 & \frac{1}{1+(y_2/y_1)^2}\frac{1}{y_1} \end{vmatrix} = 2$$

- The density is

$$\begin{aligned} f_Y(y_1,y_2) &= \mathcal{E}(x_1(y_1,y_2);1/2)\mathcal{U}(x_2(y_1,y_2);0,2\pi)|J(y_1,y_2)| \\ &= \frac{1}{2}\exp(-(y_1^2+y_2^2)/2)\frac{1}{2\pi}2 \\ &= \frac{1}{\sqrt{2\pi}}\exp(-y_1^2/2)\frac{1}{\sqrt{2\pi}}\exp(-y_2^2/2) \\ &= \mathcal{N}(y_1;0,1)\mathcal{N}(y_2;0,1) \end{aligned}$$

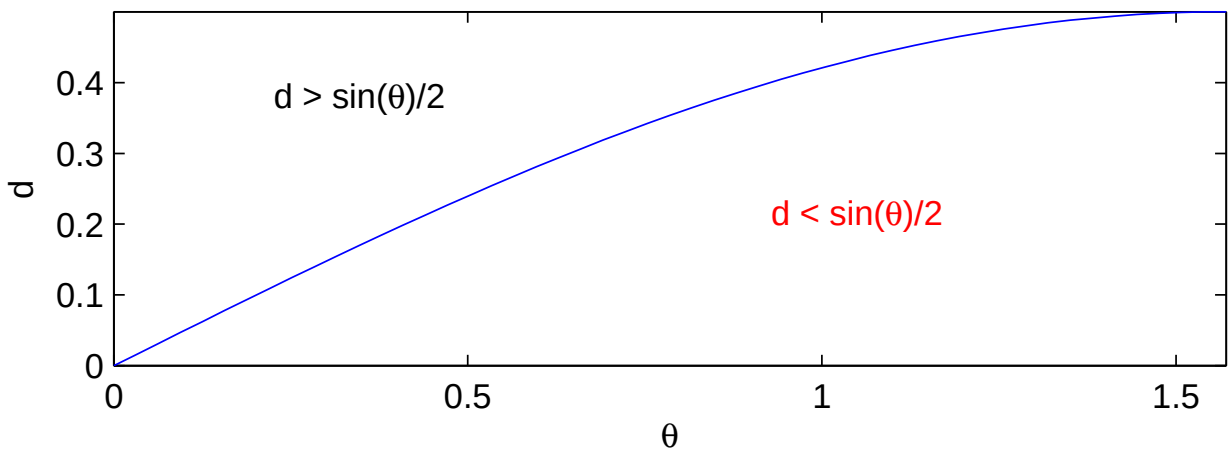
Rejection Sampling : Recall Buffon's needle



Buffon's needle

- d : Distance from the middle of the needle to the nearest line
- θ : Acute angle between the parallel lines and the needle
- A needle touches a line iff

$$\frac{d}{\sin\theta} < \frac{1}{2}$$



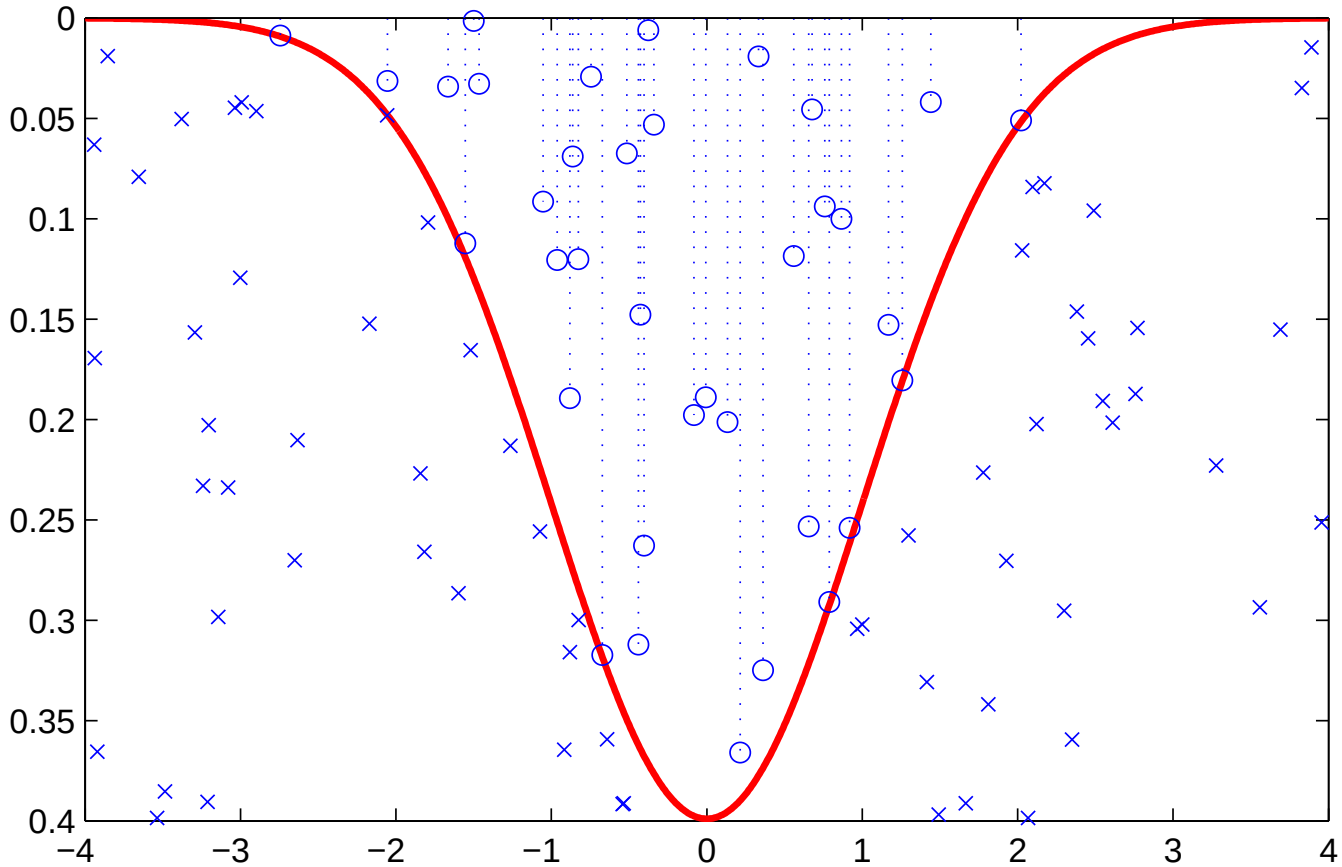
Rejection Sampling

- The solution to the Buffon needle problem is an instance of **rejection sampling**
- Basic idea

$$p(x) \quad = \quad \int_0^{p(x)} 1 d\tau = \int \mathbb{I}\{0 \leq \tau \leq p(x)\} d\tau \equiv \int p(x,\tau) d\tau$$

$\rightsquigarrow p(x)$ is a *marginal density* of the uniform distribution on the area under the curve $p(x)$. (Think of fishes in a lake)

Fishes in a lake



Rejection Sampling. Version 1.

- Construct an easy to sample density $q(x)$ such that $p(x) < Mq(x)$
- Draw $x^{(i)} \sim q(x)$, $i = 1 \ldots N$
- Generate the τ component

$$\begin{aligned} u &\sim \mathcal{U}[0,1] \\ \tau^{(i)} &= Mq(x^{(i)})u \end{aligned}$$

  The pair (x, τ) are uniformly distributed under the curve $Mq(x)$
- If $\tau^{(i)} < p(x^{(i)})$, accept: discard $\tau^{(i)}$, keep $x^{(i)}$, else reject: discard $(x^{(i)}, \tau^{(i)})$

Note that we don't need to explicitly generate $\tau^{(i)}$

Acceptance Probability

$$\begin{aligned} \tau^{(i)} &< p(x^{(i)}) \\ Mq(x^{(i)})u &< p(x^{(i)}) \\ u &< \frac{p(x^{(i)})}{Mq(x^{(i)})} \equiv \alpha(x^{(i)}) \end{aligned}$$

α is the acceptance probability.

Rejection Sampling. Final Version.

1. Construct an easy to sample density $q(x)$ such that $p(x) < Mq(x)$
2. Draw $x^{(n)} \sim q(x)$
3. Accept $x^{(n)}$ with probability

$$\alpha(x^{(n)}) = \frac{p(x^{(n)})}{Mq(x^{(n)})}$$

   - $p(x)$ can not have heavier tails than $q(x)$

Use in Bayesian Computation

- Suppose we don't know the normalisation constant Z

$$p(x|y) = \frac{1}{p(y)} p(x,y) \equiv \frac{1}{Z}\phi(x)$$

$$\begin{aligned} p(x|y) &< Mq(x) \\ \phi(x) &< ZMq(x) \equiv \bar{M}q(x) \end{aligned}$$

- We can still use the acceptance probability

$$\alpha(x^{(n)}) = \frac{\phi(x^{(n)})}{\bar{M}q(x^{(n)})}$$

- It may be very difficult to find $\bar{M}$ in higher dimensions.

Adaptive Rejection Sampling

- Idea: Adapt the rejection envelope adaptively to reduce the rejection ratio
- Suitable for log-concave ($\frown$) densities, which are analytically complex
- Gilks and Wild, 1992, Appl. Statist. Vol 41, No 2, pp. 337-348

Importance Sampling (IS)

- Consider a probability distribution $p(x)$, that is hard to sample.
- IS idea: Estimate expectations (or features) of $p(x)$ by a properly weighted sample, using a "simpler" proposal distribution q.

Importance Sampling (cont.)

- Change of measure with **weight function** $W(x) \equiv p(x)/q(x)$

$$\langle f(x) \rangle_{p(x)} = \int dx f(x) \frac{p(x)}{q(x)} q(x) = \left\langle f(x) \frac{p(x)}{q(x)} \right\rangle_{q(x)}$$

- The Expectation

$$\langle f(x) \rangle_{p(x)} \equiv \langle f(x) W(x) \rangle_{q(x)}$$

- Monte Carlo estimate

$$\tilde{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}) W(x^{(i)})$$

Unknown normalising constant

Consider a probability distribution with (possibly unknown) normalisation constant

$$p(x) = \frac{1}{Z}\phi(x) \qquad\qquad Z = \int dx \phi(x).$$

- Change of measure with **weight function** $W(x) \equiv \phi(x)/q(x)$

$$\langle f(x) \rangle_{p(x)} = \frac{1}{Z} \int dx f(x) \frac{\phi(x)}{q(x)} q(x) = \frac{1}{Z} \left\langle f(x) \frac{\phi(x)}{q(x)} \right\rangle_{q(x)}$$

Unknown normalising constant (cont.)

- Rewrite Z :

$$Z = \int dx \phi(x) = \int dx \frac{\phi(x)}{q(x)} q(x) = \langle W(x) \rangle_{q(x)}$$

- Ratio of two expectations

$$\langle f(x) \rangle_{p(x)} = \frac{\langle f(x) W(x) \rangle_{q(x)}}{\langle W(x) \rangle_{q(x)}}$$

- Monte Carlo estimate

$$\hat{\mu}_N = \frac{\sum_{i=1}^{N} W^{(i)} f(x^{(i)})/N}{\sum_{i'=1}^{N} W^{(i')}/N}$$

Normalised weights w

$$
\begin{aligned}
\hat{\mu}_N &= \frac{\sum_{i=1}^{N} W^{(i)} f(x^{(i)})}{\sum_{i'=1}^{N} W^{(i')}} \\
&= \sum_{i=1}^{N} \left(W^{(i)} / \sum_{i'=1}^{N} W^{(i')} \right) f(x^{(i)}) \\
&\equiv \sum_{i=1}^{N} \tilde{w}^{(i)} f(x^{(i)})
\end{aligned}
$$

$$
\tilde{w}^{(i)} \equiv W^{(i)} / \sum_{i'=1}^{N} W^{(i')}
$$

Importance Sampling

- Draw $i = 1, \ldots N$ independent samples from q

$$x^{(i)} \sim q(x)$$

- Calculate the **importance weights**

$$W^{(i)} \quad = \quad W(x^{(i)}) = \phi(x^{(i)})/q(x^{(i)})$$

- If $Z = 1$, i.e., $p(x) = \phi(x)$, construct the estimate $\tilde{\mu}$

$$\langle f(x) \rangle_{p(x)} \quad \approx \quad \tilde{\mu} = \frac{1}{N} \sum_{i=1}^{N} W(x^{(i)}) f(x^{(i)})$$

Importance Sampling (cont.)

- If Z is unknown, approximate the normalizing constant

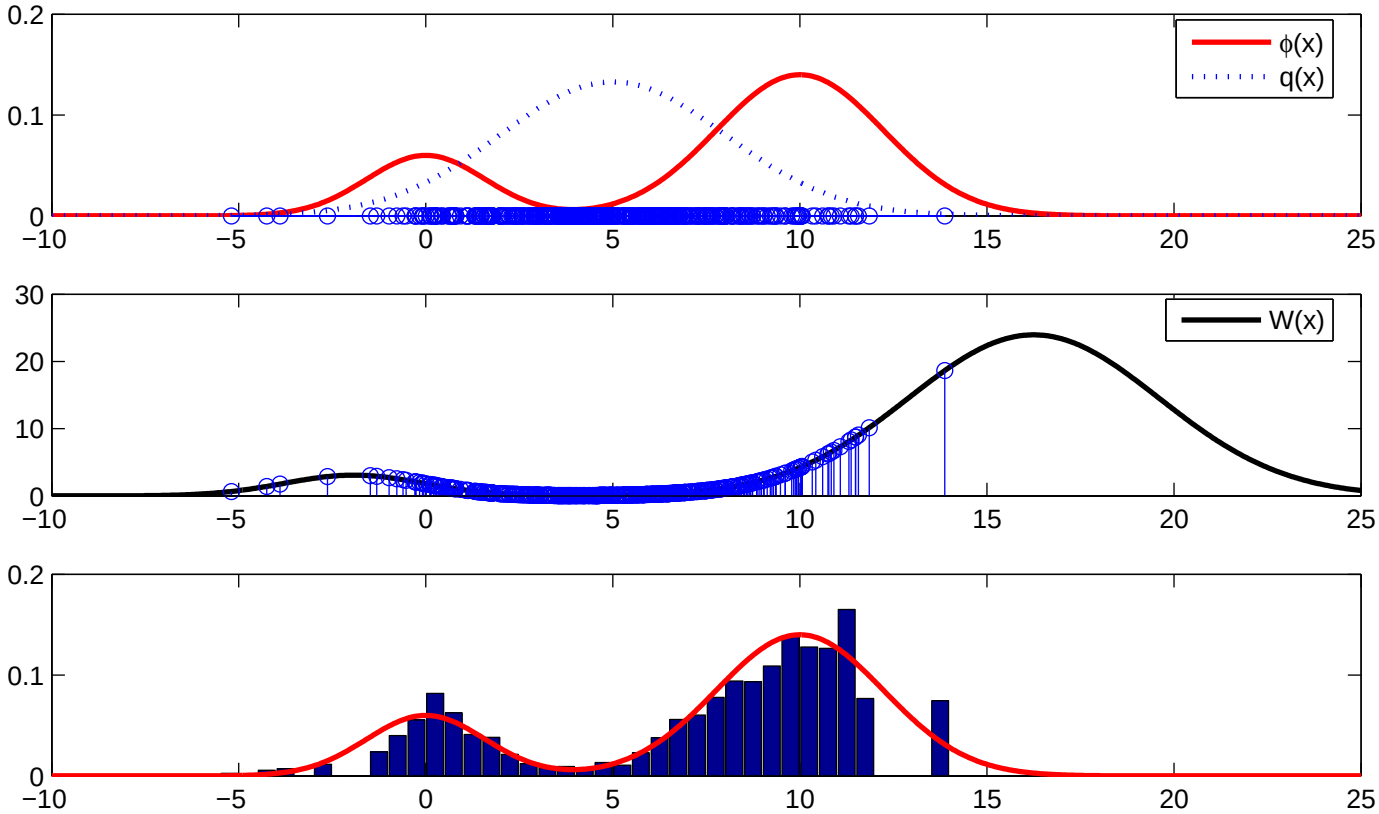
$$Z = \langle W(x) \rangle_{q(x)} \approx \frac{1}{N} \sum_{i=1}^{N} W^{(i)}$$

- Construct the estimate $\hat{\mu}$

$$\begin{aligned} \langle f(x) \rangle_{p(x)} &= \frac{\langle f(x) W(x) \rangle_{q(x)}}{\langle W(x) \rangle_{q(x)}} \approx \hat{\mu} \\ &= \frac{\sum_{i=1}^{N} W^{(i)} f(x^{(i)})/N}{\sum_{i=1}^{N} W^{(i)}/N} \equiv \sum_{i=1}^{N} \tilde{w}^{(i)} f(x^{(i)}) \equiv \hat{E}_N \end{aligned}$$

Here $\tilde{w}^{(i)} = W^{(i)} / \sum_{j=1}^{N} W^{(j)}$ are *normalized importance weights*.

Importance Sampling (cont.)



The Quality of the estimator

- Characterised by the (asymptotic) variance

$$
\begin{aligned}
Var\{f(x)W(x)\} &= \left\langle \left(f(x)W(x) - \langle f(x)W(x)\rangle\right)^2 \right\rangle_{q(x)} \\
&= \left\langle f^2(x)W^2(x)\right\rangle_{q(x)} - \langle f(x)W(x)\rangle^2_{q(x)} \\
&= \left\langle f^2(x)W^2(x)\right\rangle_{q(x)} - \langle f(x)\rangle^2_{p(x)} \\
&= \left\langle f^2(x)W^2(x)\right\rangle_{q(x)} - \mu_f^2
\end{aligned}
$$

- Alternatively, via Effective Sample Size (ESS)

Effective Sample Size (ESS)

For a weight function $W(x) = p(x)/q(x)$ we compute

$$ESS(N) = \frac{N}{1 + Var\{W\}_q}$$

- When $p(x) = q(x)$, the variance is zero, so $ESS(N) = N$.
- With increasing variance, the ESS decreases.

In practice, we estimate ESS via the coefficient of variation given samples $x^{(i)}$, $i = 1 \ldots N$ with weights $W^{(i)} = W(x^{(i)})$.

$$CV^2(W^{(1:N)}) = \frac{\sum_{i=1}^{N}(W^{(i)} - \bar{W})^2}{(N-1)\bar{W}^2}$$

$$\bar{W} = \frac{1}{N}\sum_{i=1}^{N} W^{(i)}$$

Example

- We will compute the expected value of the Beta distribution via importance sampling

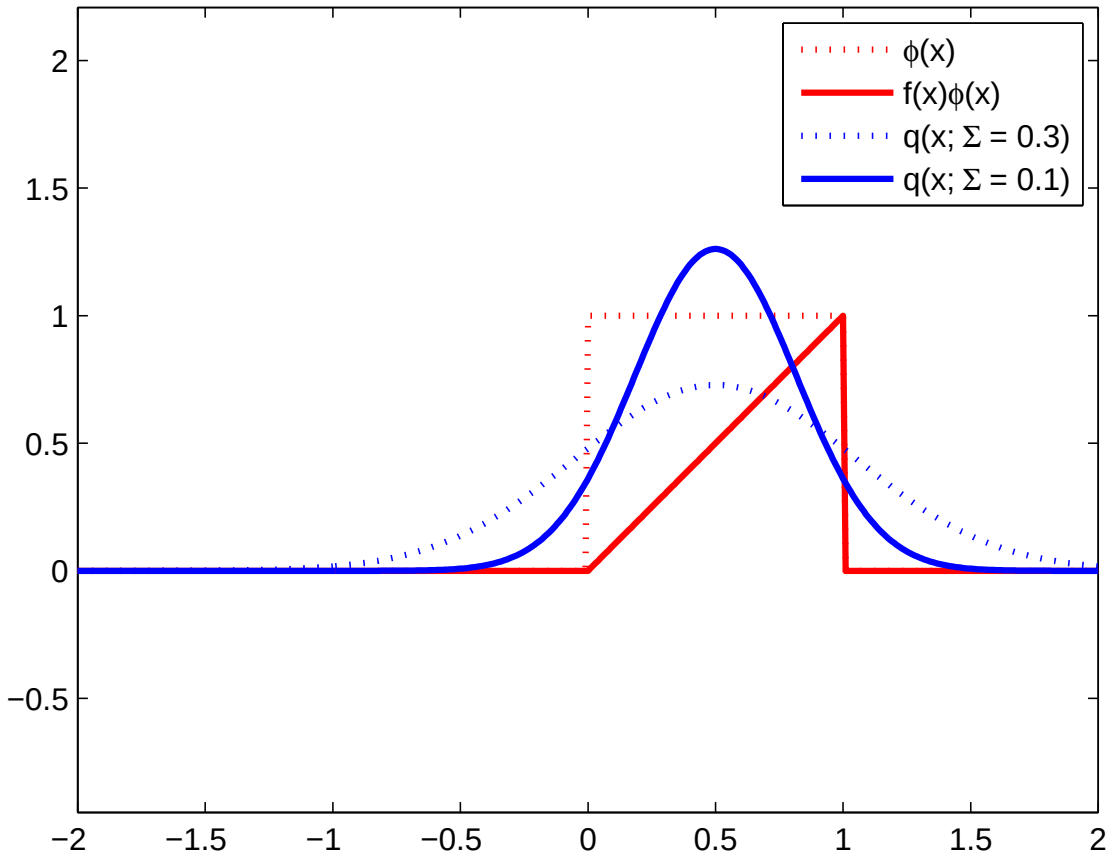
$$
\begin{aligned}
E &= \langle f(x) \rangle_{p(x)} \\
f(x) &\equiv x \\
p(x) &\equiv \mathcal{B}(x; a, b) = \phi(x)/Z \\
\phi(x) &\equiv \exp\left((a-1)\log x + (b-1)\log(1-x)\right)[0 \leq x \leq 1] \\
Z &\equiv \exp\left(-\log\Gamma(a+b) + \log\Gamma(a) + \log\Gamma(b)\right)
\end{aligned}
$$

- We will suppose that we don't know Z
- We will use a Gaussian as a proposal

$$
q(x) = \mathcal{N}(x; \mu, \Sigma)
$$

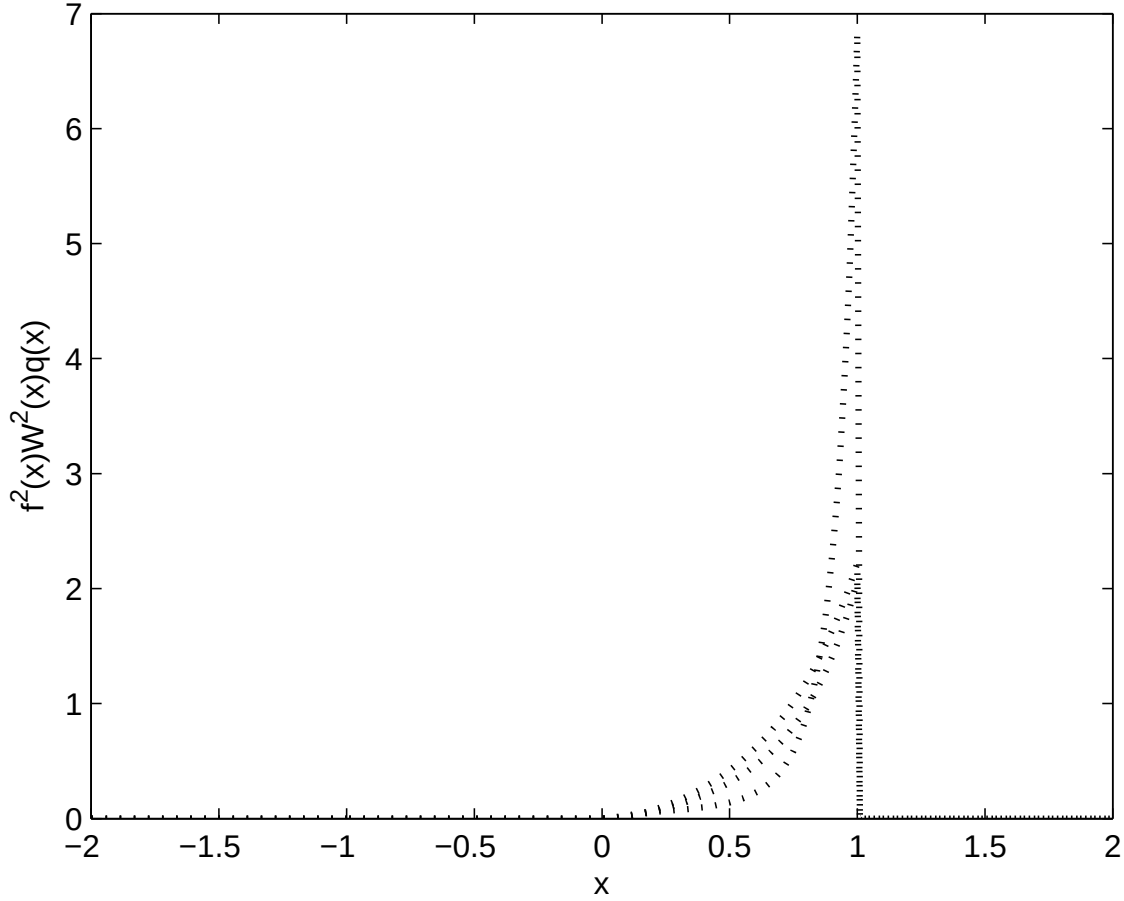
Example

- $\phi(x) \propto \mathcal{B}(x; 1, 1)$



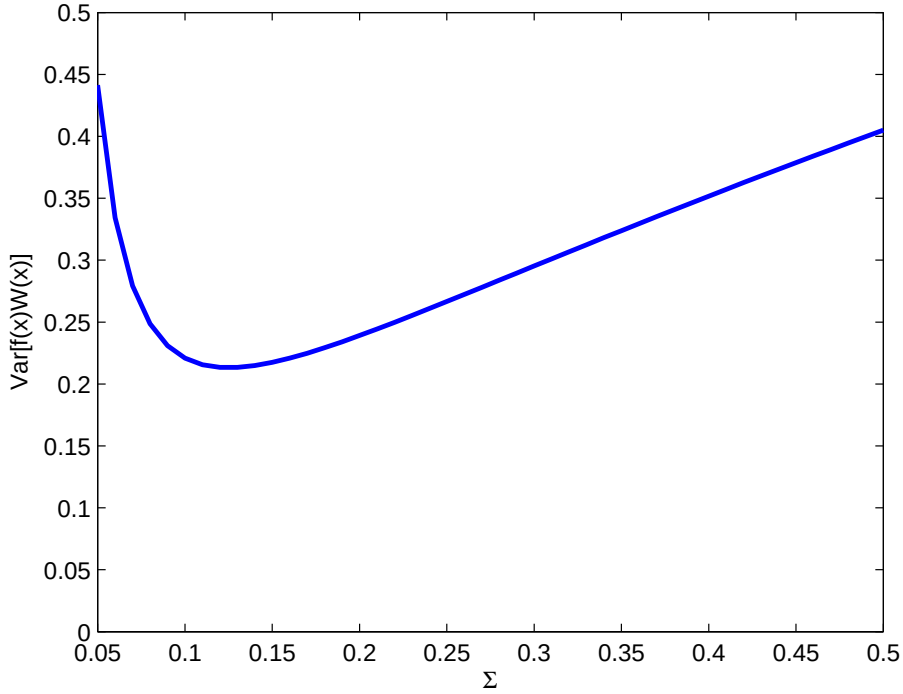
Example

- The variance is equal to the area under the curve $f(x)^2 W(x)^2 q(x) = f(x)^2 \phi(x)^2 / q(x)$

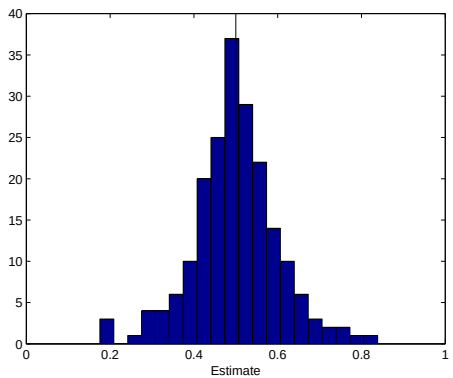


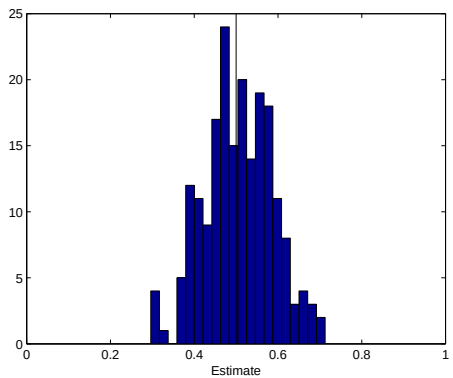
Example

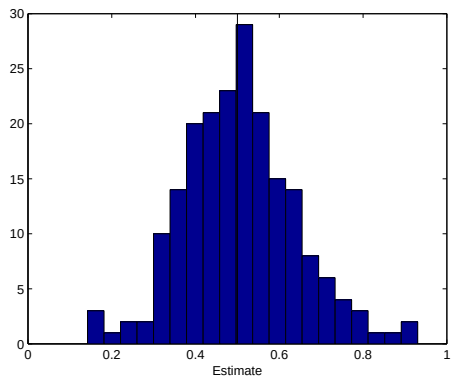
- The asymptotic variance as a function of the proposal density
- The variance is minimised when the proposal is 'close' to the target



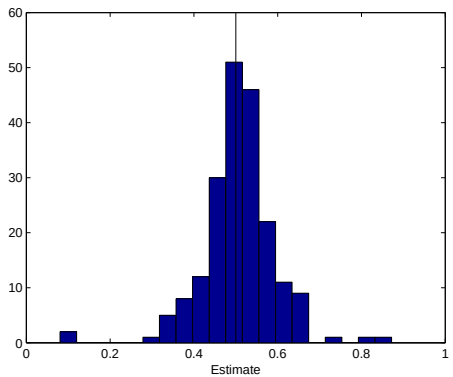
- Histogram of independent estimates with $N = 20$. $\Sigma = \{0.01, 0.1, 2\}$
- The experiment is repeated 200 times and the histogram is shown

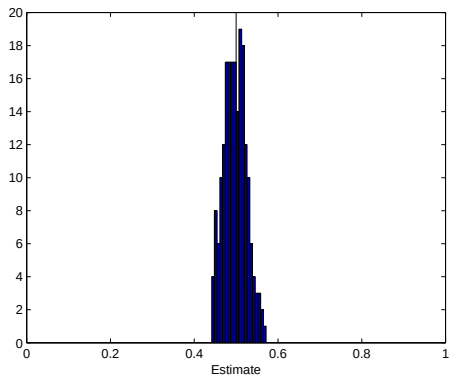


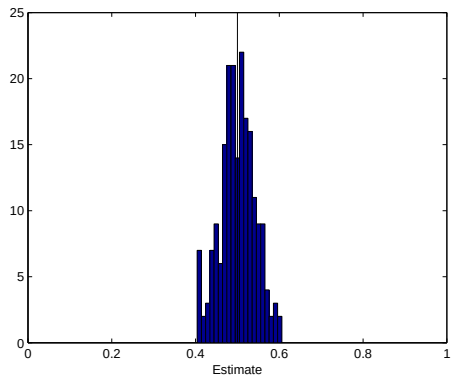




- Histogram of independent estimates with $N = 200$. $\Sigma = \{0.01, 0.1, 2\}$
- As expected, the variance scales with $N^{-1/2}$.







Example 2

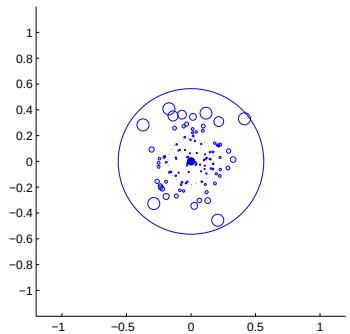
- We will compute the area of a circle centered at 0 with radius $\rho = 1/\sqrt{\pi}$ using importance sampling
- The proposal is an (isotropic) Gaussian

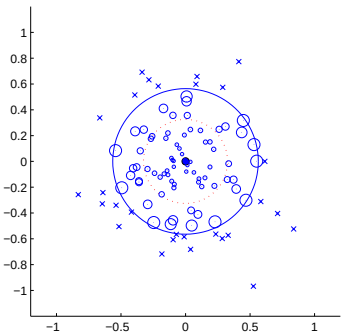
$$q(x) \quad = \quad \mathcal{N}(x; 0, \gamma\rho^2 I)$$

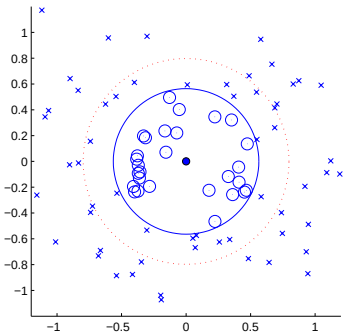
  here $\gamma > 0$ is a scalar that adjusts the variance
- We will investigate
  - The estimation quality
  - Variance of the weights - Effective sample size (ESS)
  - The effect of the proposal

Typical Draws

- Number of samples: $N = 100$
- $\gamma = \{1/10, 1/3, 2\}$, 1σ contour shown with red dotted ellipse

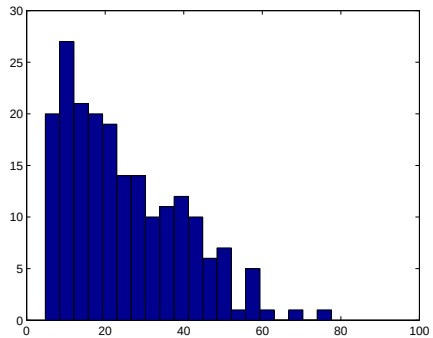


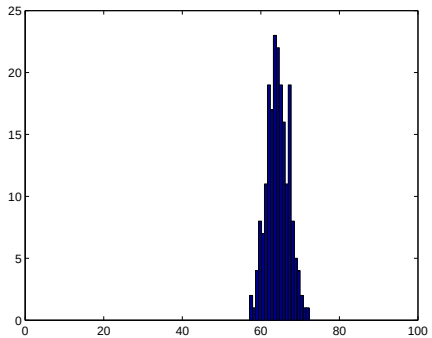


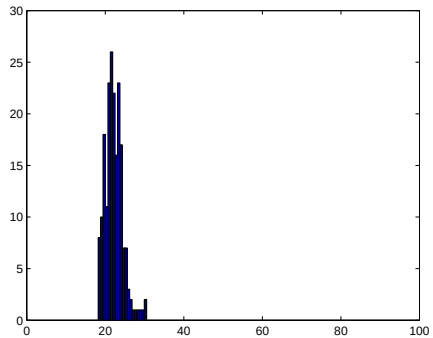


Histogram of ESS

- Number of samples: $N = 100$
- $\gamma = \{1/10, 1/3, 2\}$, 200 independent trials

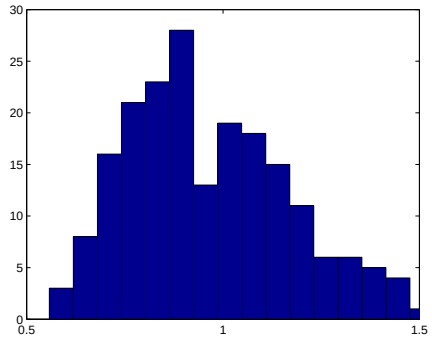


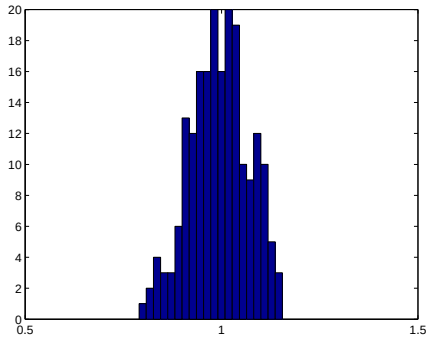


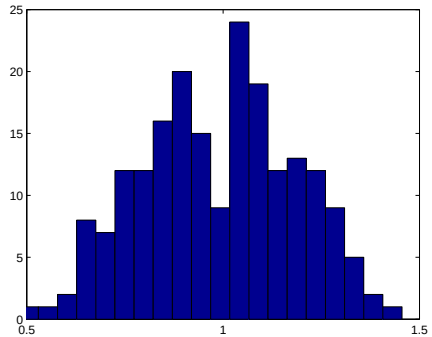


Histogram of Estimates

- Number of samples: $N = 100$
- $\gamma = \{1/10, 1/3, 2\}$, 200 independent trials

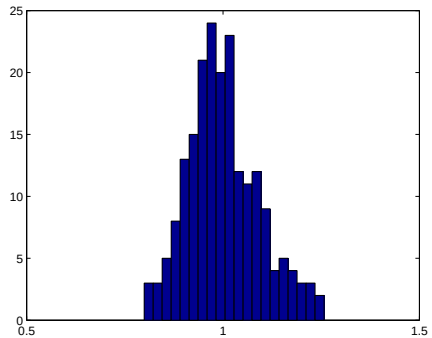


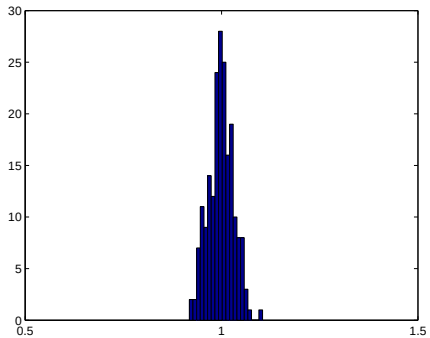


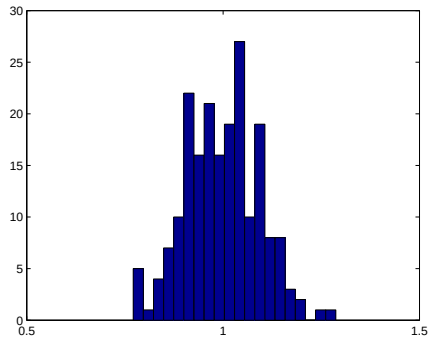


Histogram of Estimates

- Number of samples: $N = 500$
- $\gamma = \{1/10, 1/3, 2\}$, 200 independent trials







Variance reduction

There are many proposals,

- how do we choose a proposal?
- is there a “best” proposal distribution?

Optimal Proposal Distribution

Task: Estimate $\langle f(x) \rangle_{p(x)}$

- IS constructs the estimator $\mu_f = \langle f(x)W(x) \rangle_{q(x)}$
- Minimize the variance of the estimator

$$\begin{aligned} Var\{f(x)W(x)\} &= \left\langle \left(f(x)W(x) - \langle f(x)W(x) \rangle\right)^2 \right\rangle_{q(x)} \\ &= \left\langle f^2(x)W^2(x) \right\rangle_{q(x)} - \langle f(x)W(x) \rangle^2_{q(x)} \\ &= \left\langle f^2(x)W^2(x) \right\rangle_{q(x)} - \langle f(x) \rangle^2_{p(x)} \\ &= \left\langle f^2(x)W^2(x) \right\rangle_{q(x)} - \mu_f^2 \end{aligned}$$

- Minimize the first term: only it depends on q

Optimal Proposal Distribution

- (By Jensen's inequality) The first term is lower bounded:

$$\left\langle f^2(x)W^2(x)\right\rangle_{q(x)} \geq \left\langle |f(x)|W(x)\right\rangle^2_{q(x)} = \left(\int |f(x)|\, p(x)dx\right)^2$$

- Since this is a lower bound, we can't do better that this
- We will look for a distribution q^* that attains this lower bound. Take

$$q^*(x) = \frac{|f(x)|p(x)}{\int |f(x')|p(x')dx'}$$

Optimal Proposal Distribution (cont.)

- The weight function for this particular proposal q^* is

$$W_*(x) = p(x)/q^*(x) = \frac{\int |f(x')|p(x')dx'}{|f(x)|}$$

- We show that q^* attains its lower bound

$$\begin{aligned}
\left\langle f^2(x)W_*^2(x)\right\rangle_{q^*(x)} &= \left\langle f^2(x)\frac{\left(\int |f(x')|p(x')dx'\right)^2}{|f(x)|^2}\right\rangle_{q^*(x)} \\
&= \left(\int |f(x')|p(x')dx'\right)^2 = \langle |f(x)|\rangle^2_{p(x)} \\
&= \langle |f(x)|W_*(x)\rangle^2_{q^*(x)}
\end{aligned}$$

- $\Rightarrow$ There are distributions q^* that are even "better" than the distribution itself!

Reducing Variance of importance weights: A link to alpha divergences

The α-divergence between two distributions is defined as

$$D_\alpha(p||q) \equiv \frac{1}{\beta(1-\beta)}\left(1-\int dx p(x)^\beta q(x)^{1-\beta}\right)$$

where $\beta = (1+\alpha)/2$ and p and q are two probability distributions

- $\lim_{\beta\to 0} D_\alpha(p||q) = KL(q||p)$
- $\lim_{\beta\to 1} D_\alpha(p||q) = KL(p||q)$
- $\beta = 2$, $(\alpha = 3)$

$$D_3(p||q) \equiv \frac{1}{2}\int dx p(x)^2 q(x)^{-1} - \frac{1}{2} = \frac{1}{2}\left\langle W(x)^2\right\rangle_{q(x)} - \frac{1}{2}$$

Best q (in a constrained family) is typically a heavy-tailed approximation to p

Summary

- Random Number Generation
  - Pseudo-random numbers
- Inversion
- Transformation
- Rejection Sampling
- Importance Sampling