.

Chapter 4 Simulation and Monte Carlo Integration

4.1 Introduction to the Monte Carlo Method

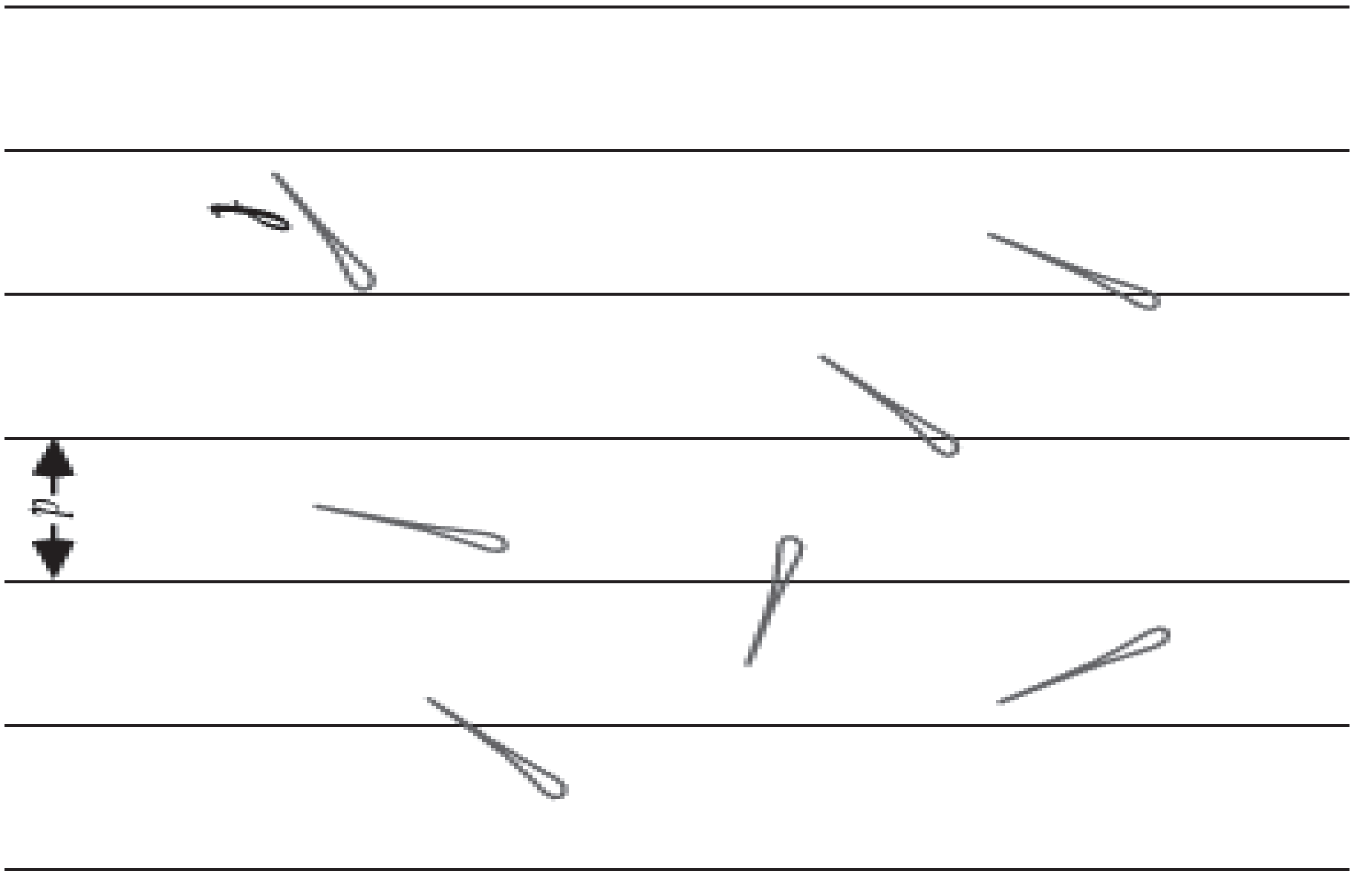
- **Monte Carlo Methods:**
  - Wikipedia: Monte Carlo methods are a broad class of computational algorithms that rely on repeated **random sampling** to obtain numerical results.
  - It was named, by Stanislaw Ulam and Nicholas Metropolis, after the Monte Carlo Casino.

4.1 Introduction to the Monte Carlo Method

- **Buffon's Needle (1777):** Consider the problem of calculating π.
  - Suppose there is a floor made of parallel strips of wood with the same width d. Drop a needle with length $l < d$ onto the floor. We consider the probability that the needle will lie across a line between two strips.
  - Let X be the distance from the center of the needle to the closest parallel line, and let θ be the acute angle between the needle and one of the parallel lines.
  - It is reasonable to assume that (X, θ) follows a uniform distribution with the density

$$f(x,\theta) = \begin{cases} \frac{2}{d} \cdot \frac{2}{\pi}, & 0 \le X \le \frac{d}{2}, 0 \le \theta \le \frac{\pi}{2}; \\ 0, & \text{elsewhere.} \end{cases}$$

4.1 Introduction to the Monte Carlo Method



Buffon’s Needle

4.1 Introduction to the Monte Carlo Method

- – The probability that the needle will lie across a line is

$$\begin{aligned} P\big(X < \frac{l}{2}\cdot\sin(\theta)\big) &= \int\int_{x<\frac{l}{2}\cdot\sin(\theta)} f(x,\theta)\,dxd\theta \\ &= \frac{2}{d}\cdot\frac{2}{\pi}\int_0^{\pi/2}\int_0^{\frac{l}{2}\cdot\sin(\theta)} 1\,dxd\theta = \frac{2l}{d\pi}. \end{aligned}$$

- We can estimate the value of π with the following random experiment.
  - Drop a needle onto the floor m times, let I be the number of times that the the needle lies across a line. Then $I/m \to 2l/d\pi$.
  - Estimate π by

$$\widehat{\pi} = \frac{2l/d}{I/m}.$$

4.1 Introduction to the Monte Carlo Method

- **Monte Carlo Integration/Summation:** Suppose we want to calculate $\int g(x)\,dx$ or $\sum_{x\in A} g(x)$.
  - Generate random samples $x^{(1)}, \cdots, x^{(m)}$ from a *trial distribution* (or *sampling distribution*) with PDF/PMF $q(x)$.
  - Calculate

$$\begin{aligned}\widehat{\Pi} &= \frac{1}{m}\sum_{j=1}^{m}\frac{g(x^{(j)})}{q(x^{(j)})} \\ &\xrightarrow{a.s.} E\left(\frac{g(x^{(j)})}{q(x^{(j)})}\right) = \int \frac{g(x)}{q(x)}\,q(x)dx = \int g(x)\,dx.\end{aligned}$$

- **Remarks:**
  - In the previous sample, $g(x,\theta) = I\big(x < \frac{l}{2} \cdot \sin(\theta)\big) f(x,\theta)$ and $q(x,\theta) = f(x,\theta)$.
  - The support of $q(x)$ should cover the support of $g(x)$, that is,

$$\mathcal{X}_g := \{x : g(x) \neq 0\} \subset \mathcal{X}_q := \{x : q(x) > 0\}. \quad \textbf{(Why?)}$$

  - The root mean squared error (RMSE) of $\widehat{\Pi}$ is

$$\begin{aligned} \left\{E\left[\widehat{\Pi} - \int g(x)\,dx\right]^2\right\}^{1/2} &= \left\{\mathrm{Var}\big(\widehat{\Pi}\big)\right\}^{1/2} \\ &= \left\{\frac{1}{m}\mathrm{Var}_q\left(\frac{g\big(x^{(j)}\big)}{q\big(x^{(j)}\big)}\right)\right\}^{1/2} \\ &= \frac{1}{\sqrt{m}}\mathrm{Var}_q^{1/2}\left(\frac{g\big(x^{(j)}\big)}{q\big(x^{(j)}\big)}\right). \end{aligned}$$

4.1 Introduction to the Monte Carlo Method

- – We have

$$
\begin{aligned}
&\mathrm{Var}_q\left(\frac{g(x^{(j)})}{q(x^{(j)})}\right)\\
=\ & E_q\left(\frac{g(x^{(j)})}{q(x^{(j)})}\right)^2 - E_q^2\left(\frac{g(x^{(j)})}{q(x^{(j)})}\right)\\
=\ & E_q\left(\frac{g(x^{(j)})}{q(x^{(j)})}\right)^2 - \left(\int g(x)\,dx\right)^2\\
=\ & E_q\left(\frac{|g(x^{(j)})|}{q(x^{(j)})}\right)^2 - \left(\int |g(x)|\,dx\right)^2 + \left(\int |g(x)|\,dx\right)^2 - \left(\int g(x)\,dx\right)^2\\
=\ & \mathrm{Var}_q\left(\frac{|g(x^{(j)}|)}{q(x^{(j)})}\right) + \left(\int |g(x)|\,dx\right)^2 - \left(\int g(x)\,dx\right)^2\\
\geq\ & \left(\int |g(x)|\,dx\right)^2 - \left(\int g(x)\,dx\right)^2.
\end{aligned}
$$

4.1 Introduction to the Monte Carlo Method

- – Suppose that $\int |g(x)|\,dx < \infty$. Then $|g(x)|/\int |g(x)|\,dx$ is a density function.

  – **"Optimal" choice of q:** If we let $q(x) = |g(x)|/\int |g(x)|\,dx$, then

  $$\mathrm{Var}_q\left(\frac{|g(x^{(j)}|)}{q(x^{(j)})}\right) = 0$$

  and

  $$\mathrm{Var}_q\left(\frac{g(x^{(j)})}{q(x^{(j)})}\right) = \left(\int |g(x)|\,dx\right)^2 - \left(\int g(x)\,dx\right)^2.$$

  – However, we **can not** use $q(x) = |g(x)|/\int |g(x)|\,dx$ in practice, since we need to compute the normalizing constant $\int |g(x)|\,dx$.

  – We should choose the trial distribution so that (1) it is easy to draw samples from $q(x)$; (2) it is easy to calculate the value of $q(x)$; (3) $q(x)$ is close to $|g(x)|/\int |g(x)|\,dx$.

4.2 Random Variable Generation

- **Pseudo-random Number:** One of the most common approaches to generating pseudo-random numbers starts with an initial positive integer $x_0 \neq 0$, then recursively computes x_n by letting

  $$x_n = a x_{n-1} \mod m$$

  where a and m are given positive integers.

  - x_0 is call the *random seed.*
  - x_n takes value in $\{1, 2, \cdots, m-1\}$. (Note that we need choose x_0 and a so that $x_n \neq 0$ for any n.)
  - Since x_n, $n = 0, 1, \cdots$, can only take finite number of values, the sequence must repeat itself after a certain number of iterations.
  - We want the sequence $\frac{x_0}{m}, \frac{x_1}{m}, \cdots$ performs as a sequence of i.i.d. random variables following the Uniform$(0, 1)$ distribution.

4.2 Random Variable Generation

- - In general, the constants a and m should be chosen to satisfy the following criteria.
    - For any initial seed, $\frac{x_0}{m}, \frac{x_1}{m}, \cdots$ has the "appearance" of being a sequence of independent Uniform$(0, 1)$ distributed random variables.
    - For any initial seed, the number of variables that can be generated before repetition begins is large.
    - The values can be computed efficiently on a digital computer.
  - For example, one choice for a 32-bit computer is $m = 2^{32} - 1$ and $a = 7^5 = 16,807$.
  - Sometimes, we also consider the recursions of the type

$$x_n = ax_{n-1} + c \mod m.$$

4.2 Random Variable Generation

- In the following, suppose that we can generate i.i.d. Uniform$(0, 1)$ distributed sequence $U_1, U_2, \cdots$.

- **Generating Discrete Random Variables:** Suppose $X \in \{a_0, a_1, \cdots\}$ is a discrete random variable with $P(X = a_i) = p_i$, where $p_0 + p_1 + \cdots = 1$. We can simulate X as follows.
  - Step 1: Generate a random number $U \sim$ Uniform$(0, 1)$.
  - Step 2: Set $i = 0$ and $F = p_0$.
  - Step 3: If $U < F$, set $X = a_i$ and stop.
  - Step 4: Let $i \leftarrow i + 1$ and $F \leftarrow F + p_i$.
  - Step 5: Go to Step 3.

- **Remarks:**
  - At iteration i, $F = p_0 + \cdots + p_i$.
  - We let $X = a_i$ if $p_0 + \cdots + p_{i-1} \leq U < p_0 + \cdots + p_{i-1} + p_i$. Obviously, $P(X = a_i) = p_i$.
  - If we want to generate i.i.d. random variables $X_1, \cdots, X_m$ with $P(X_j = a_i) = p_i$, we should save $F_i = p_0 + \cdots + p_i$.

4.2 Random Variable Generation

- **Example: Bernoulli(p) Variable Generation.**
  - Step 1: Generate a random number $U \sim \text{Uniform}(0,1)$.
  - Step 2: If $U < p$, set $X = 1$; otherwise, set $X = 0$.

- **Example: Binomial($n; p$) Variable Generation.** Note that if $X \sim \text{Binomial}(n;p)$,

$$P(X = i) = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} = \frac{n-i+1}{i} \cdot \frac{p}{1-p} \cdot P(X = i-1).$$

  - Step 1: Generate a random number $U \sim \text{Uniform}(0,1)$.
  - Step 2: Set $c = p/(1-p)$, $q = (1-p)^n$, $i = 0$ and $F = q$.
  - Step 3: If $U < F$, set $X = i$ and stop.
  - Step 4: Let $i \leftarrow i+1$, $q \leftarrow \frac{n-i+1}{i} \cdot c \cdot q$ and $F \leftarrow F + q$.
  - Step 5: Go to Step 3.

- **Remarks:**
  - Note that if $X \sim \text{Binomial}(n; p)$, $E(X) = np$. On average, we need compute $np + 1$ probabilities and make $np + 1$ comparisons. When $p > 0.5$, we can generate $Y \sim \text{Binomial}(n; 1 - p)$ and let $X = n - Y$.
  - We can also generate $Y_1, \cdots, Y_n$ i.i.d. following the Bernoulli(p) distribution and let $X = Y_1 + \cdots + Y_n$.
  - If we already have $F_0, F_1, \cdots$, generating X from Bernoulli distributed random variables requires n random numbers and makes n comparisons, whereas the previous method only requires 1 random number and make $np + 1$ comparisons (on average).

- **Example: Uniform**$\{1, \cdots, n\}$ **Variable Generation.**
  - Step 1: Generate a random number $U \sim \text{Uniform}(0, 1)$.
  - Step 2: Let $X = 1 + \lfloor nU \rfloor$, where $\lfloor nU \rfloor$ denotes the integer part of nU.

- **Example: Geometric**(p) **Variable Generation.** A Geometric(p) random variable X has the distribution $P(X = i) = p(1-p)^{i-1}$ for $i = 1, 2, \cdots$, where $0 < p < 1$. Note that

$$P(X \le i) = P(X = 1) + \cdots + P(X = i) = 1 - (1-p)^i.$$

To simulate X, we generate $U \sim \text{Uniform}(0,1)$, the let $X = i$ if

$$\begin{aligned} & P(X \le i-1) \le U < P(X \le i) \\ \Leftrightarrow\ & 1 - (1-p)^{i-1} \le U < 1 - (1-p)^i \\ \Leftrightarrow\ & i - 1 \le \log(1-U)/\log(1-p) < i. \end{aligned}$$

Hence, we can simulate $X \sim \text{Geometric}(p)$ as follows.

  - Step 1: Generate a random number $U \sim \text{Uniform}(0,1)$.
  - Step 2: Let $X = 1 + \lfloor \frac{\log(1-U)}{\log(1-p)} \rfloor$ (or let $X = 1 + \lfloor \frac{\log(U)}{\log(1-p)} \rfloor$).

4.2 Random Variable Generation

- **Example: Random Permutation Generation.** Suppose we are interested in generating a permutation of the numbers $1, \cdots, n$ which is such that all $n!$ possible orderings are equally likely.
  - Step 1: Let $P_1, \cdots, P_n$ be any permutation of $1, \cdots, n$, for example, let $(P_1, \cdots, P_n) = (1, \cdots, n)$.
  - Step 2: Set $k = n$.
  - Step 3: Generate a random number $U \sim \text{Uniform}(0, 1)$ and let $I = 1 + \lfloor kU \rfloor$.
  - Step 4: Interchange the values of P_I and P_k.
  - Step 5: Let $k \leftarrow k - 1$; if $k > 1$ go to Step 3.
  - Step 6: $P_1, \cdots, P_n$ is the desired random permutation.

4.2 Random Variable Generation

- Next, we consider generating continuous random variables. Suppose we want to simulate X following a distribution with the CDF $F_X(x)$ and the PDF $f_X(x)$.

- **Definition:** Let F be a cumulative distribution fucntion, the generalized inverse of F is defined by

$$F^{-}(u) = \inf\{x : F(x) \geq u\} \quad \text{for } 0 < u < 1.$$

  - $F^{-}(u)$ is called the *uth quantile* of F.
  - Note that $F\big(F^{-}(u)\big)$ may not be equal to u.
  - We have $F\big(F^{-}(u)\big) \geq u$ and $F\big(F^{-}(u) - \varepsilon\big) < u$ for any $\varepsilon > 0$. (**Homework**)
  - If F is continuous at $F^{-}(u)$, then $F\big(F^{-}(u)\big) = u$.
  - For any CDF F and any $0 < u < 1$, $F(x) \geq u$ if and only if $x \geq F^{-}(u)$.

4.2 Random Variable Generation

- **Theorem:** Let F_X be the CDF of a random variable X. If $U \sim$ Uniform$(0,1)$, then $F_X^-(U)$ has the distribution F_X, which can be written as $F_X^-(U) \overset{d}{=} X$.

- **Proof.** Note that $F_X^-(u) \le x$ if and only if $u \le F_X(x)$. We have

$$P\big(F_X^-(U) \le x\big) = P\big(U \le F_X(x)\big) = F_X(x).$$

  This completes the proof.

- **Remarks:**

  - Let $X \sim F_X$. Under certain conditions (*e.g.*, F_X is continuous), we have

$$X \overset{d}{=} F_X^-(U) \;\Rightarrow\; F_X(X) \overset{d}{=} F_X\big(F_X^-(U)\big) = U.$$

    Therefore, $F_X(X) = \int_{-\infty}^{X} f_X(x)\,dx$ follows a Uniform$(0,1)$ distribution. The theorem is also called the *probability integral transform*.

4.2 Random Variable Generation

- – If we want to simulate $X \sim F_X$, we can first generate $U \sim \text{Uniform}(0, 1)$, then let $X = F_X^-(U)$.

  – Suppose $X \in \{a_0 < a_1 < \cdots\}$ is a discrete random variable with $P(X = a_i) = p_i$, where $p_0 + p_1 + \cdots = 1$. We can simulate X by letting

$$X = F_X^-(U) = \begin{cases} a_0, & \text{if } 0 < U \le p_0; \\ a_1, & \text{if } p_0 < U \le p_0 + p_1; \\ \cdots, & \cdots; \\ a_i, & \text{if } p_0 + \cdots + p_{i-1} < U \le p_0 + \cdots + p_{i-1} + p_i; \\ \cdots, & \cdots, \end{cases}$$

  where $U \sim \text{Uniform}(0, 1)$.

4.2 Random Variable Generation

- **Example: Exponential(β) Variable Generation.** Let X be an Exponential(β) distributed random variable with $f_X(x) = \frac{1}{\beta}e^{-x/\beta}$ and $F_X(x) = 1 - e^{-x/\beta}$ for $x > 0$.
  - Here $\beta > 0$. We have $E(X) = \beta$.
  - Define $\lambda = 1/\beta$, which is called the *rate parameter* of the exponential distribution.
  - If $U \sim$ Uniform$(0, 1)$, then

$$X \overset{d}{=} F_X^{-}(U) = -\beta \log(1 - U) \overset{d}{=} -\beta \log U.$$

  - If $X_1, X_2, \cdots$ are i.i.d. Exponential(1) random variables,

$$2\sum_{i=1}^{m} X_j \sim \chi^2_{2m}, \quad \beta\sum_{i=1}^{m} X_i \sim \text{Gamma}(m, \beta), \quad \text{and} \quad \frac{\sum_{i=1}^{m} X_i}{\sum_{i=1}^{m+n} X_i} \sim \text{Beta}(m, n).$$

- **Gamma Distribution:** A continuous random variable X follows a Gamma(α,β) ($\alpha, \beta > 0$) distribution if

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$.

(1) $E(X) = \alpha\beta$.

(2) $\mathrm{Var}(X) = E(X^2) - E^2(X) = \alpha\beta^2$.

(3) $M_X(u) = (1 - \beta u)^{-\alpha}$ for $u < 1/\beta$.

- **Remarks:**
  - Exponential(β) = Gamma(1, β).
  - If $X_1, \cdots, X_n$ are independent and $X_i \sim$ Gamma(α_i, β), then $X_1 + \cdots + X_n \sim$ Gamma($\alpha_1 + \cdots + \alpha_n$, β).

- **Chi-Squared Distribution:** A random variable follows *Chi-squared* distribution with *degrees of freedom* p, denoted by χ_p^2, if

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2} & x > 0, \\ 0 & \text{otherwise}, \end{cases}$$

- **Remarks:**

  – If X is a $N(0,1)$ random variable, then $X^2 \sim \chi_1^2$.

  – $\chi_p^2 = \text{Gamma}(p/2,\, 2)$.

  – $\chi_2^2 = \text{Gamma}(1,\, 2) = \text{Exponential}(2)$.

  – If $X_1, \cdots, X_n$ are independent and $X_i \sim \chi_{p_i}^2$, then $X_1 + \cdots + X_n \sim \chi_{p_1+\cdots+p_n}^2$.

4.2 Random Variable Generation

- **Example: Normal Variable Generation.** Suppose X_1 and X_2 are two i.i.d. $N(0,1)$ random variables. Let (r, θ) be the polar coordinates of (X_1, X_2), that is,

$$r = \sqrt{X_1^2 + X_2^2} \quad \text{and} \quad \theta = \text{angle}(X_1, X_2),$$

or equivalently,

$$X_1 = r\cos\theta \quad \text{and} \quad X_2 = r\sin\theta.$$

  – We can prove that $r > 0$ and $\theta \in [0, 2\pi)$ are independent,

$$r^2 \sim \chi_2^2 \overset{d}{=} -2\log U_1 \quad \text{and} \quad \theta \sim \text{Uniform}[0, 2\pi) \overset{d}{=} 2\pi U_2.$$

  – We can simulate two independent $N(0,1)$ random variables as follows.

    ∗ Generate U_1 and U_2 i.i.d. following the Uniform$(0,1)$ distribution.

    ∗ Let $X_1 = \sqrt{-2\log U_1}\cos\big(2\pi U_2\big)$ and $X_2 = \sqrt{-2\log U_1}\sin\big(2\pi U_2\big)$.

- – If $X \sim N(0,1)$, then $Y = \mu + \sigma X \sim N(\mu, \sigma^2)$.

  – A random vector $Y = (Y_1, \cdots, Y_p)^T$ follows a *multivariate normal distribution* $N(\boldsymbol{\mu}, \Sigma)$ if

$$\begin{aligned} f_{Y_1 \cdots Y_p}(\boldsymbol{y}) &= \frac{1}{\sqrt{det[2\pi\Sigma]}} exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right\} \\ &= \frac{1}{(2\pi)^{p/2}\sqrt{det[\Sigma]}} exp\left\{ -\frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} r_{i,j}(x_i - \mu_i)(x_j - \mu_j) \right\}, \end{aligned}$$

  where $\boldsymbol{y} = (y_1, \cdots, y_p)^T$, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^T$, $\Sigma = \{\Sigma_{i,j}\}_{p\times p}$ is a $p \times p$ symmetric and positive definite matrix, and $R = \{r_{i,j}\}_{p\times p} = \Sigma^{-1}$. We have

$$E(Y_i) = \mu_i \quad \text{and} \quad \mathrm{Cov}(Y_i, Y_j) = \Sigma_{i,j}.$$

- – **Bivariate Normal Distribution:** Suppose (X, Y) follow a bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, its joint density function is

$$\begin{aligned} & f_{XY}(x, y) \\ & = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ & \quad exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}, \end{aligned}$$

then

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

- – Suppose that $Y = (Y_1, \cdots, Y_p)^T \sim N(\boldsymbol{\mu}, \Sigma)$.

  ∗ We have $Y_i \sim N(\mu_i, \Sigma_{i,i})$.

  ∗ The moment generating function of Y is $M_Y(\boldsymbol{u}) = E\big[\exp\{\boldsymbol{u}^T Y\}\big] = \exp\{\boldsymbol{u}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{u}^T \Sigma \boldsymbol{u}\}$, where $\boldsymbol{u} = (u_1, \cdots, u_p)^T$.

  ∗ Let A be a $k \times p$ matrix and let b be a $k \times 1$ vector, then

$$AY + b \sim N\big(A\boldsymbol{\mu} + b, A\Sigma A^T\big).$$

  – For a $p \times p$ non-negative definite symmetric matrix Σ, we can find a $p \times p$ symmetric matrix, denoted by $\Sigma^{1/2}$, satisfying $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$.

  – We can simulate $Y = (Y_1, \cdots, Y_p)^T \sim N(\boldsymbol{\mu}, \Sigma)$ by letting

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = \boldsymbol{\mu} + \Sigma^{1/2} \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix},$$

  where $X_1, \cdots, X_p$ are i.i.d. $N(0, 1)$ random variables.

4.2 Random Variable Generation

- **Rejection Method:** Suppose we want to simulate a random variable with density function $f(x)$, and we have a method for generating a random variable having density $g(x)$ with $\mathcal{X}_g \supset \mathcal{X}_f$. Assume that there exists a positive constant c such that

$$\frac{f(x)}{g(x)} \leq c \quad \text{for all } x.$$

Then we can generate a random variable X having density $f(x)$ as follows.

  - Step 1: Generate Y having density g.
  - Step 2: Generate a random number $U \sim \text{Uniform}(0, 1)$.
  - Step 3: If $U \leq \frac{f(Y)}{cg(Y)}$, set $X = Y$ and stop. Otherwise, go to Step 1.

- **Remark:** The method is also called the *acceptance-rejection method*, and g is called the *instrumental distribution*.

- **Theorem:** (1) The number of iterations that the algorithm need is a geometric random variable with mean c. (2) The random variable X generated by the rejection method has density f.

- **Proof.**

(1) Let Y_i and U_i be the random number Y and U generated at iteration i, respectively. Let D be the number of iterations needed before stopping. We have

$$P(D=i) = P\left(U_1 > \frac{f(Y_1)}{cg(Y_1)}\right)\cdots P\left(U_{i-1} > \frac{f(Y_{i-1})}{cg(Y_{i-1})}\right) P\left(U_i \le \frac{f(Y_i)}{cg(Y_i)}\right).$$

Note that Y_k and U_k are independent. It is easy to find

$$\begin{aligned} P\left(U_k \le \frac{f(Y_k)}{cg(Y_k)}\right) &= \int\int_{u\le\frac{f(y)}{cg(y)}} 1\cdot g(y)du\,dy \\ &= \int\int_0^{\frac{f(y)}{cg(y)}} 1\cdot g(y)du\,dy = \int \frac{f(y)}{c}dy = \frac{1}{c}. \end{aligned}$$

4.2 Random Variable Generation

- – Therefore, $P(D=i) = \big(1-1/c\big)^{i-1} \cdot (1/c)$ for $i = 1, 2, \cdots$, and D is a geometric random variable with mean c.

(2) We have

$$\begin{aligned}
P(X \le x) &= \sum_{i=1}^{\infty} P(Y_i \le x, D = i) \\
&= \sum_{i=1}^{\infty} P\left(Y_i \le x, U_i \le \frac{f(Y_i)}{cg(Y_i)} \,\Big|\, U_1 > \frac{f(Y_1)}{cg(Y_1)}, \cdots U_{i-1} > \frac{f(Y_{i-1})}{cg(Y_{i-1})}\right) \\
&\qquad\qquad \times P\left(U_1 > \frac{f(Y_1)}{cg(Y_1)}, \cdots U_{i-1} > \frac{f(Y_{i-1})}{cg(Y_{i-1})}\right) \\
&= \sum_{i=1}^{\infty} \int_{-\infty}^{x} \int_{0}^{\frac{f(y)}{cg(y)}} 1 \cdot g(y) du\, dy \cdot \big(1-1/c\big)^{i-1} \\
&= \sum_{i=1}^{\infty} \int_{-\infty}^{x} f(y)\, dy \cdot \frac{1}{c} \cdot \big(1-1/c\big)^{i-1} = \int_{-\infty}^{x} f(y)\, dy.
\end{aligned}$$

Hence, $f_X(x) = f(x)$.

- **Remarks:**
  - Suppose that $X \sim f(x) = K \cdot \bar{f}(x)$, where $\bar{f}(x)$ is given, but the normalizing constant $K = 1/\int \bar{f}(x)\,dx$ is unknown.
    - For example, we have $f_{\boldsymbol{\theta}|X}(\theta|x) \propto \pi(\theta) f_{X|\boldsymbol{\theta}}(x|\theta)$ for Bayesian inference.
    - Assume that we know $\frac{\bar{f}(x)}{g(x)} \le M$ for all x. Then $\frac{f(x)}{g(x)} \le c := KM$. The condition in Step 3 of the rejection method becomes
$$U \le \frac{f(Y)}{KM \cdot g(Y)} \quad \Leftrightarrow \quad U \le \frac{\bar{f}(Y)}{M \cdot g(Y)}.$$
      The normalizing constant K need not to be known to apply the rejection method.
  - The average number of iterations need for the rejection method is $E(D) = c \ge 1$. We want c to be as small as possible for efficiency. In general, we want g is close to f so that c would be close to 1.

- **Example: Generating Gamma from Exponential.** We want to generate a Gamma(α,β) distributed random variable with density $f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}e^{-x/\beta}$ for $x > 0$. Consider the rejection method with $g(x;b) = \frac{1}{b}e^{-x/b}$ for $x > 0$.
  - We have
    $$\frac{f(x)}{g(x;b)} = \frac{b}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1}e^{-x(b-\beta)/(\beta b)}.$$
  - When $0 < \alpha < 1$, $\sup_{x>0} \frac{f(x)}{g(x;b)} = \infty$. When $\alpha = 1$, Gamma(1,β) is an exponential distribution.
  - We only consider the case when $\alpha > 1$. When $b > \beta$, the maximum value of $f(x)/g(x;b)$ is obtained at $x = \frac{(\alpha-1)\beta b}{b-\beta}$ and we have
    $$\frac{f(x)}{g(x;b)} \leq \frac{b}{\Gamma(\alpha)\beta^\alpha} \left(\frac{(\alpha-1)\beta b}{b-\beta}\right)^{\alpha-1} e^{1-\alpha} = \frac{(\alpha-1)^{\alpha-1}b^\alpha}{\Gamma(\alpha)\beta(b-\beta)^{\alpha-1}} e^{1-\alpha} := c(b).$$

4.2 Random Variable Generation

-
  - Obviously, $c(b)$ is minimized when $b = \alpha\beta$, and the minimum value is

    $$c_* = \alpha^\alpha \, e^{1-\alpha}/\Gamma(\alpha).$$

  - When we use the rejection method to simulate a Gamma(α,β) variable with $g(x; b) = \frac{1}{b}e^{-x/b}$ for $x > 0$, we should choose $b = \alpha\beta$, that is, we use an exponential distribution having the same mean as the Gamma distribution.
  - When $\alpha = 1.5$, $c_* = 1.2573$; when $\alpha = 10$, $c_* = 3.4008$.
  - Note that we do not need to know $\Gamma(\alpha)$ when using the rejection method.

- **Example: Generating Normal from Double Exponential.** Consider generating a $N(0,1)$ variable by the rejection method using a double-exponential distribution with density $g(x;\lambda) = (\lambda/2)e^{-\lambda|x|}$ for $-\infty < x < \infty$, where $\lambda > 0$.
  - **How to generate a double-exponential variable?**
  - We have

$$\begin{aligned}\frac{f(x)}{g(x;\lambda)} &= \frac{1}{\sqrt{2\pi}}e^{-x^2/2}(2/\lambda)e^{\lambda|x|} \\ &= \sqrt{\frac{2}{\pi}}\,\lambda^{-1}\exp\left\{-\frac{1}{2}(|x|-\lambda)^2+\lambda^2/2\right\} \leq \sqrt{\frac{2}{\pi}}\,\lambda^{-1}e^{\lambda^2/2} := c(\lambda).\end{aligned}$$

  - $c(\lambda)$ is minimized when $\lambda = 1$, and the minimum value is

$$c_* = \sqrt{\frac{2e}{\pi}} = 1.3155.$$

- **Example: Truncated Normal Variable Generation.** We want to generate a truncated $N(0,1)$ random variable with the density

$$f(x) = \frac{\phi(x)}{1-\Phi(b)} \cdot I(x \geq b),$$

where Φ and ϕ are the CDF and PDF of the $N(0,1)$ distribution, respectively.

  - Consider the instrumental distribution $g_1(x) = \phi(x)$. Obviously,

$$\frac{f(x)}{g_1(x)} \leq \frac{1}{1-\Phi(b)} := c_1.$$

Note that $\frac{f(x)}{c_1 g_1(x)} = I(x \geq b)$. The rejection method is equivalent to

$*$ Step 1: Generate $Y \sim N(0,1)$.

$*$ Step 2: If $Y \geq b$, set $X = Y$ and stop. Otherwise, go to Step 1.

Where $\Phi(b)$ is close to 1, this method can be very inefficient.

- – Consider the *translated exponential distribution*

$$g_2(x;\lambda) = \lambda e^{-\lambda(x-b)} \cdot I(x \geq b),$$

where $\lambda > 0$. Then for $x \geq b$,

$$\begin{aligned}
\frac{f(x)}{g_2(x;\lambda)} &= \frac{1}{\sqrt{2\pi}\big(1-\Phi(b)\big)\lambda} \cdot \exp\big\{-x^2/2 + \lambda(x-b)\big\} \\
&= \frac{1}{\sqrt{2\pi}\big(1-\Phi(b)\big)\lambda} \cdot \exp\big\{-(x-\lambda)^2/2 - \lambda b + \lambda^2/2\big\} \\
&\leq \begin{cases} \frac{1}{\sqrt{2\pi}\big(1-\Phi(b)\big)\lambda} \cdot \exp\big\{-\lambda b + \lambda^2/2\big\}, & \text{if } \lambda \geq b; \\ \frac{1}{\sqrt{2\pi}\big(1-\Phi(b)\big)\lambda} e^{-b^2/2}, & \text{if } 0 < \lambda < b, \end{cases} \\
&:= c_2(\lambda).
\end{aligned}$$

where $c_2(\lambda)$ is minimized at $\lambda_* = b/2 + \sqrt{1 + b^2/4}$.

- – The algorithm for rejection method with the instrumental distribution $g_2(x;\lambda_*) = \lambda_* e^{-\lambda_*(x-b)} \cdot I(x \geq b)$, where $\lambda_* = b/2 + \sqrt{1+b^2/4}$, is as follows.
  - $*$ Step 1: Generate Y having density $g_2(x;\lambda_*)$.
  - $*$ Step 2: Generate a random number $U \sim \text{Uniform}(0,1)$.
  - $*$ Step 3: If $U \leq \frac{f(Y)}{g_2(Y;\lambda_*)c(\lambda_*)} = \exp\big\{-(Y-\lambda_*)^2/2\big\}$, set $X = Y$ and stop. Otherwise, go to Step 1.

  – The average number of iterations need for the rejection method using the instrumental distribution $g_2(x;\lambda_*)$ is

$$c_2(\lambda_*) = \frac{1}{\sqrt{2\pi}\big(1-\Phi(b)\big)\lambda_*} \cdot \exp\big\{-\lambda_* b + \lambda_*^2/2\big\}.$$

  When $b = -1$, $c_1 = 1.1886$ and $c_2(\lambda_*) = 1.7230$; when $b = 0$, $c_1 = 2$ and $c_2(\lambda_*) = 1.3155$; when $b = 1$, $c_1 = 6.3030$ and $c_2(\lambda_*) = 1.1409$; when $b = 2$, $c_1 = 43.9558$ and $c_2(\lambda_*) = 1.0711$.

4.2 Random Variable Generation

- – When it requires substantial computing time at each evaluation of the density f, we may use the *envelope acceptance-rejection method* to generate random variables from f.

  – **Envelope Acceptance-Rejection Method:** Suppose there exist a density (instrumental distribution) $g(x)$, a function $g_l(x)$ and a constant $c > 0$ such that

  $$g_l(x) \leq f(x) \leq c\, g(x),$$

  where $g_l(x)$ is easy to evaluate and $g(x)$ is easy to draw samples from. We can generate a random variable X having density $f(x)$ as follows.

  ∗ Step 1: Generate $Y \sim g(x)$ and $U \sim \text{Uniform}(0,1)$.

  ∗ Step 2: If $U \leq \frac{g_l(Y)}{cg(Y)}$, set $X = Y$ and stop.

  ∗ Step 3: If $U \leq \frac{f(Y)}{cg(Y)}$, set $X = Y$ and stop. Otherwise, go to Step 1.

4.3 Monte Carlo Variance Reduction Methods

- We consider calculating $\int h(x)f(x)\,dx$, where $f(x)$ is a density function.
  - When $h(x) = x^k$ and $f(x) = f_X(x)$, $\int h(x)f(x)\,dx = E(X^k)$.
  - When $h(x) = I(x \in A)$ and $f(x) = f_X(x)$,

$$\int h(x)f(x)\,dx = \int_A f(x)\,dx = P(X \in A).$$

  - When $h(x) = x$ and $f(x) = f_{X|Z}(x|z)$,

$$\int h(x)f(x)\,dx = \int x f_{X|Z}(x|z)\,dx = E(X|Z = z).$$

  - If we can generate $x^{(1)}, x^{(2)}, \cdots, x^{(m)}$ i.i.d. from $f(x)$, for example, using the rejection method, then

$$\widehat{\Pi}_0 := \frac{1}{m}\sum_{j=1}^{m} h\big(x^{(i)}\big) \xrightarrow{a.s.} \int h(x)f(x)\,dx := E_f\big[h(X)\big].$$

- **Stratified Sampling:** Let $\mathcal{X}_f = \{x : f(x) > 0\}$. Suppose we divide $\mathcal{X}_f$ into K disjoint sets $D_1, \cdots, D_K$. Then

$$\int h(x) f(x)\, dx = \int_{D_1} h(x) f(x)\, dx + \cdots + \int_{D_K} h(x) f(x)\, dx.$$

  - Define $a_k = \int_{D_k} f(x)\, dx$ and $f_k(x) = \frac{1}{a_k} f(x) I(x \in D_k)$ for $k = 1, \cdots, K$. Here $f_k(x)$ is the density of the distribution $f(x)$ conditional on D_k.
  - Suppose that a_k is known and we can generate samples $x^{(k,1)}, \cdots, x^{(k,m_k)}$ from $f_k(x)$, then

$$\Pi_{S,k} = \int_{D_k} h(x) f(x)\, dx = a_k \cdot \int h(x) f_k(x)\, dx$$

can be estimated by

$$\widehat{\Pi}_{S,k} = a_k \cdot \frac{1}{m_k} \sum_{j=1}^{m_k} h\big(x^{(k,j)}\big).$$

- – Note that

  $$\begin{aligned}\widehat{\Pi}_S &= \widehat{\Pi}_{S,1} + \cdots + \widehat{\Pi}_{S,K} \\ &= \frac{a_1}{m_1}\sum_{j=1}^{m_1} h\big(x^{(1,j)}\big) + \cdots + \frac{a_K}{m_K}\sum_{j=1}^{m_K} h\big(x^{(K,j)}\big)\end{aligned}$$

  is an unbiased estimator for $\int h(x)f(x)\,dx$. We have

  $$\begin{aligned}\mathrm{MSE}(\widehat{\Pi}_S) &= \mathrm{Var}(\widehat{\Pi}_S) \\ &= \sum_{k=1}^{K} \frac{a_k^2}{m_k}\mathrm{Var}_{f_k}\big[h\big(x^{(k,j)}\big)\big].\end{aligned}$$

  – Let $m = m_1 + \cdots + m_K$, the total number of samples. We compare variances of $\widehat{\Pi}_0$ and $\widehat{\Pi}_S$ for a fixed m.

4.3 Monte Carlo Variance Reduction Methods

- – When m is fixed, $\mathrm{Var}(\widehat{\Pi}_S)$ is minimized when

$$m_k = m \cdot \frac{a_k \mathrm{Var}_{f_k}^{1/2}\left[h\left(x^{(k,j)}\right)\right]}{\sum_{s=1}^{K} a_s \mathrm{Var}_{f_s}^{1/2}\left[h\left(x^{(s,j)}\right)\right]}, \quad \textbf{(Homework)}$$

and the minimum value is

$$\min\left\{\mathrm{Var}(\widehat{\Pi}_S)\right\} = \frac{1}{m}\left\{\sum_{k=1}^{K} a_k \mathrm{Var}_{f_k}^{1/2}\left[h\left(x^{(k,j)}\right)\right]\right\}^2.$$

In many cases, we have $\mathrm{Var}_{f_k}\left[h\left(x^{(k,j)}\right)\right] < \mathrm{Var}_f\left[h\left(x^{(j)}\right)\right]$. Also note that $a_1 + \cdots + a_K = 1$. Then

$$\begin{aligned}\min\left\{\mathrm{Var}(\widehat{\Pi}_S)\right\} &\le \frac{1}{m}\left\{\mathrm{Var}_f^{1/2}\left[h\left(x^{(j)}\right)\right]\right\}^2 \\ &= \frac{1}{m}\mathrm{Var}_f\left[h\left(x^{(j)}\right)\right] = \mathrm{Var}(\widehat{\Pi}_0).\end{aligned}$$

- – If $\mathrm{Var}_{f_k}\big[h\big(x^{(k,j)}\big)\big]$, $k = 1, \cdots, K$, are unknown, but $\mathrm{Var}_{f_k}\big[h\big(x^{(1,j)}\big)\big] \approx \cdots \approx \mathrm{Var}_{f_K}\big[h\big(x^{(K,j)}\big)\big]$, we can let $m_k = a_k m$. Define $d(x) = k$ if $x \in D_k$. Note that

$$\mathrm{Var}_{f_k}\big[h\big(x^{(k,j)}\big)\big] = \mathrm{Var}_f\big[h\big(x^{(j)}\big)|d\big(x^{(j)}\big) = k\big].$$

Then

$$\begin{aligned} \mathrm{Var}(\widehat{\Pi}_S) &= \frac{1}{m}\sum_{k=1}^{K} a_k \mathrm{Var}_{f_k}\big[h\big(x^{(k,j)}\big)\big] \\ &= \frac{1}{m}\sum_{k=1}^{K} a_k \mathrm{Var}_f\big[h\big(x^{(j)}\big)|d\big(x^{(j)}\big) = k\big] \\ &= \frac{1}{m} E_f\Big\{\mathrm{Var}_f\big[h\big(x^{(j)}\big)|d\big(x^{(j)}\big)\big]\Big\} \\ &\leq \frac{1}{m}\mathrm{Var}_f\big[h\big(x^{(j)}\big)\big] = \mathrm{Var}(\widehat{\Pi}_0). \end{aligned}$$

4.3 Monte Carlo Variance Reduction Methods

- **Control Variate Method:** We want to calculate $E_f\big[h(X)\big] = \int h(x)f(x)\,dx$. Suppose $\mu_g = E_f\big[g(X)\big]$ is known. We can generate $x^{(1)}, \cdots, x^{(m)}$ i.i.d. from $f(x)$ and estimate $E_f\big[h(X)\big]$ by

$$\widehat{\Pi}_C = \frac{1}{m}\sum_{j=1}^{m}\left[h(x^{(j)}) + b\big(g(x^{(j)}) - \mu_g\big)\right].$$

  - $\widehat{\Pi}_C$ is an unbiased estimator of $E_f\big[h(X)\big]$. The MSE of $\widehat{\Pi}_C$ is

$$\begin{aligned}\text{MSE}(\widehat{\Pi}_C) &= \text{Var}(\widehat{\Pi}_C)\\ &= \frac{1}{m}\Big[\text{Var}_f\big(h(x^{(j)})\big) + 2b\text{Cov}_f\big(h(x^{(j)}), g(x^{(j)})\big) + b^2\text{Var}_f\big(g(x^{(j)})\big)\Big].\end{aligned}$$

  - $\text{Var}(\widehat{\Pi}_C)$ is minimized when $b = -\text{Cov}_f\big(h(x^{(j)}), g(x^{(j)})\big)/\text{Var}_f\big(g(x^{(j)})\big)$, and the minimum value is

$$\frac{1}{m}\text{Var}_f\big(h(x^{(j)})\big)\,\big[1-\rho_f^2(h,g)\big] = \text{Var}(\widehat{\Pi}_0)\,\big[1-\rho_f^2(h,g)\big],$$

where $\rho_f(h,g) = \frac{\text{Cov}_f\big(h(x^{(j)}),g(x^{(j)})\big)}{\text{Var}_f^{1/2}\big(h(x^{(j)})\big)\text{Var}_f^{1/2}\big(g(x^{(j)})\big)}$.

4.3 Monte Carlo Variance Reduction Methods

- – For example, we want to calculate $P(X > a) = E\big[I(X > a)\big]$ for a given a, where X has density f.
  - $*$ Assume we know that f is symmetric around μ.
  - $*$ Assume that $a > \mu$. We can generate $x^{(1)}, \cdots, x^{(m)}$ i.i.d. from $f(x)$ and estimate $P(X > a)$ by

$$\frac{1}{m}\sum_{j=1}^{m}\big[I(x^{(j)} > a) + b\big(I(x^{(j)} > \mu) - 0.5\big)\big],$$

where

$$\begin{aligned} b &\approx -\frac{\mathrm{Cov}_f\big(I(x^{(j)} > a), I(x^{(j)} > \mu)\big)}{\mathrm{Var}_f\big(I(x^{(j)} > \mu)\big)} \\ &= -\frac{P(X > a) - P(X > a)/2}{1/4} \\ &= -2P(X > a). \end{aligned}$$

4.3 Monte Carlo Variance Reduction Methods

- **Antithetic Variate Method:** We want to calculate $E_f\big[h(X)\big] = \int h(x)f(x)\,dx$. Let $F(x)$ be the CDF with density $f(x)$. Consider the estimator

$$\widehat{\Pi}_A = \frac{1}{m}\sum_{j=1}^{m/2}\Big[h\big(F^{-}(U^{(j)})\big) + h\big(F^{-}(1-U^{(j)})\big)\Big],$$

  where both $F^{-}(U^{(j)})$ and $F^{-}(1-U^{(j)})$ follow the distribution with density $f(x)$.

  - Define $g(u) := h\big(F^{-}(u)\big)$. We have

$$\begin{aligned}\mathrm{MSE}(\widehat{\Pi}_A) &= \mathrm{Var}(\widehat{\Pi}_A)\\ &= \frac{1}{m}\cdot\frac{1}{2}\cdot\Big[\mathrm{Var}_f\big(h(x^{(j)})\big) + 2\mathrm{Cov}\big(g(U^{(j)}), g(1-U^{(j)})\big) + \mathrm{Var}_f\big(h(x^{(j)})\big)\Big].\end{aligned}$$

  - Suppose that $h(x)$ is a **monotonic** (increasing/decreasing) function of x, *e.g.*, $h(x) = x$, then $g(u)$ is also **monotonic**.

4.3 Monte Carlo Variance Reduction Methods

- – Let $U^{(1)}$ and $U^{(2)}$ be independent. Then

$$\begin{aligned} 0 &\geq E\Big[\big(g(U^{(1)}) - g(U^{(2)})\big)\big(g(1-U^{(1)}) - g(1-U^{(2)})\big)\Big] \\ &= E\big[g(U^{(1)})g(1-U^{(1)})\big] + E\big[g(U^{(2)})g(1-U^{(2)})\big] \\ &\qquad\qquad -E\big[g(U^{(1)})g(1-U^{(2)})\big] - E\big[g(U^{(2)})g(1-U^{(1)})\big] \\ &= 2E\big[g(U^{(1)})g(1-U^{(1)})\big] - 2E\big[g(U^{(1)})\big]\cdot E\big[g(1-U^{(1)})\big] \\ &= 2\mathrm{Cov}\big(g(U^{(1)}), g(1-U^{(1)})\big) \end{aligned}$$

  – Hence, $\mathrm{Var}(\widehat{\Pi}_A) \leq \frac{1}{m}\cdot\frac{1}{2}\cdot\Big[\mathrm{Var}_f\big(h(x^{(j)})\big) + \mathrm{Var}_f\big(h(x^{(j)})\big)\Big] = \mathrm{Var}(\widehat{\Pi}_0)$.

  – Similarly, if $f(x)$ is symmetric around μ, we can generate $x^{(1)}, \cdots, x^{(m/2)}$ i.i.d. from $f(x)$ and use the estimator

$$\widehat{\Pi}_A = \frac{1}{m}\sum_{j=1}^{m/2}\Big[h\big(x^{(j)}\big) + h\big(2\mu - x^{(j)}\big)\Big].$$

  – **If $h(x)$ is not a monotonic function, the conclusion may not hold.**

- **Rao-Blackwellization:** Suppose $X = (X_1, X_2)$. Assume that we are able to compute $E_f\big(h(X)|X_2 = x_2\big)$. We can generate $x^{(j)} = (x_1^{(j)}, x_2^{(j)})$, $j = 1, \cdots, m$, i.i.d. from $f(x)$ and estimate $E_f\big[h(X)\big]$ by

$$\widehat{\Pi}_0 = \frac{1}{m}\sum_{j=1}^{m} h(x_1^{(j)}, x_2^{(j)}) \quad \text{or} \quad \widehat{\Pi}_R = \frac{1}{m}\sum_{j=1}^{m} E_f\big(h(X)|X_2 = x_2^{(j)}\big).$$

We have

$$\begin{aligned} \text{MSE}(\widehat{\Pi}_R) &= \text{Var}(\widehat{\Pi}_R) \\ &= \frac{1}{m}\text{Var}_f\big[E_f\big(h(X_1, X_2)|X_2\big)\big] \\ &\le \frac{1}{m}\text{Var}_f\big(h(X_1, X_2)\big) \\ &= \text{Var}(\widehat{\Pi}_0). \end{aligned}$$

- **Remark:** One basic principle in Monte Carlo computation: **One should carry out analytical computation as much as possible.**

4.4 Importance Sampling

- **Importance Sampling:** We want to calculate $E_f\left[h(X)\right] = \int h(x) f(x)\, dx$. Suppose that we can generate samples $x^{(1)}, x^{(2)}, \cdots, x^{(m)}$ i.i.d. from a *trial distribution (proposal distribution)* $q(x)$ with $\mathcal{X}_q \supset \mathcal{X}_f$, then

$$\begin{aligned}
\widehat{\Pi}_1 &:= \frac{1}{m} \sum_{j=1}^{m} h(x^{(j)})\, \frac{f(x^{(j)})}{q(x^{(j)})} \\
&\xrightarrow{a.s.} E_q\left(h(x^{(j)})\, \frac{f(x^{(j)})}{q(x^{(j)})} \right) \\
&= \int_{\mathcal{X}_q} h(x)\, \frac{f(x)}{q(x)}\, q(x) dx = \int_{\mathcal{X}_q} h(x) f(x)\, dx = \int_{\mathcal{X}_f} h(x) f(x)\, dx.
\end{aligned}$$

- **Remarks:**
  - To minimize the MSE of $\widehat{\Pi}_1$, we want to choose $q(x) = \frac{|h(x)| f(x)}{\int |h(x)| f(x)\, dx}$.
  - In practice, we may need to calculate $\int h(x) f(x)\, dx$ for several different h's. We often choose $q(x)$ close to the *target distribution* $f(x)$.

4.4 Importance Sampling

- – Compared with uniform sampling (or Newton-Côtes quadrature), the importance sampling idea suggests that one should focus on the "important" region to improve efficiency.

  – Define $w^{(j)} := w(x^{(j)}) = f(x^{(j)})/q(x^{(j)})$, then

  $$\widehat{\Pi}_1 = \frac{1}{m}\sum_{j=1}^{m} w^{(j)} h(x^{(j)}),$$

  where $w^{(j)}$ is called the *importance weight*.

  – For any function h with $\int |h(x)| f(x)\,dx < \infty$, we also have

  $$\widehat{\Pi}_2 := \frac{\sum_{j=1}^{m} w^{(j)} h(x^{(j)})}{\sum_{j=1}^{m} w^{(j)}} \xrightarrow{a.s.} \int h(x) f(x)\,dx. \quad (\textbf{Why?}).$$

4.4 Importance Sampling

- – We say that the set $\{(x^{(j)}, w^{(j)})\}_{j=1}^m$ is *properly weighted* with respect to (*w.r.t.*) the target distribution $f(x)$. **It can be interpreted as that we use a discrete distribution**

$$f(x) \simeq \sum_{j=1}^{m} \frac{w^{(j)}}{\sum_{k=1}^m w^{(k)}} \delta(x - x^{(j)}) \quad \textbf{or} \quad P(X = x^{(j)}) = \frac{w^{(j)}}{\sum_{k=1}^m w^{(k)}}$$

  to approximate the target distribution $f(x)$, where $\delta(\cdot)$ is the Dirac delta function. Here $\delta(u) = 0$ for $u \neq 0$, $\delta(0) = \infty$ and $\int_{-\infty}^{\infty} \delta(u)\, du = 1$.

  – Comparing $\widehat{\Pi}_1$ with $\widehat{\Pi}_2$.

  - $\widehat{\Pi}_1$ is an unbiased estimator for $\int h(x) f(x)\, dx$, but $\widehat{\Pi}_2$ is biased.
  - When calculating $\widehat{\Pi}_2$, the multiplicative constants in f or q can be ignored.
  - $\widehat{\Pi}_2$ is in the form of a weighted average, it is often more robust than $\widehat{\Pi}_1$.

4.4 Importance Sampling

- – Comparing importance sampling with the rejection method.
    - ∗ It is difficult to compare efficiency of the importance sampling and the rejection method in general cases (it is case-specified).
    - ∗ The importance sampling method is often easier to implement, since we do not need to find an upper bound for $f(x)/q(x)$.
    - ∗ The rejection method generates "useless" samples when rejecting, but the importance sampling method uses all generated samples.

  – In practice, the efficiency of $\widehat{\Pi}_2$ is often measured using the *effective sample size (ESS)*, which is defined as

$$\mathrm{ESS}_m(w) = \frac{m}{1 + \mathrm{Var}_q\big(w(x^{(j)})\big)}.$$

  It can be interpreted as that the m weighted samples perform as $\mathrm{ESS}_m(w)$ i.i.d. samples drawn from the target distribution $f(x)$.

4.4 Importance Sampling

- – We have

$$\begin{aligned}\mathrm{Var}_q\big(w(x^{(j)})\big) &= E_q\big(w(x^{(j)})\big)^2 - \big[E_q\big(w(x^{(j)})\big)\big]^2 \\ &= \int \frac{f^2(x)}{q^2(x)}\, q(x)\, dx - 1 = \int \frac{f(x)}{q(x)}\, f(x)\, dx - 1.\end{aligned}$$

  It is called the χ^2-*divergence* between f and q, which measures the difference between the target distribution and the proposal distribution.

  – Usually, we want the proposal distribution $q(x)$ has a fatter tail than the target distribution $f(x)$, otherwise, $\mathrm{Var}_q\big(w(x^{(j)})\big)$ could be infinite.

  – Since the importance weights $w^{(1)}, \cdots, w^{(m)}$ **may not be normalized** in practice (*we may ignore some multiplicative constants so that* $E\big(w(x^{(j)})\big) \neq 1$), we can estimate $\mathrm{Var}_q\big(w(x^{(j)})\big)$ by

$$\frac{1}{m}\sum_{j=1}^{m}\left(\frac{w^{(j)}}{\bar{w}_m}\right)^2 - 1,$$

  where $\bar{w}_m = \frac{1}{m}\sum_{j=1}^{m} w^{(j)}$.

4.4 Importance Sampling

- **Marginalization:** Consider the target distribution $f(x)$. Suppose we generate $x^{(1)}, \cdots, x^{(m)}$ from the trail distribution $q(x)$ and let $w^{(j)} = f(x^{(j)})/q(x^{(j)})$, then $\{(x^{(j)}, w^{(j)})\}_{j=1}^m$ is *properly weighted* with respect to $f(x)$, that is,

$$\frac{\sum_{j=1}^m w^{(j)} h(x^{(j)})}{\sum_{j=1}^m w^{(j)}} \xrightarrow{a.s.} \int h(x) f(x)\, dx$$

for any h with finite $\int |h(x)| f(x)\, dx$.

  – The proper weights $\{w^{(j)}\}_{j=1}^m$ are **not unique**. For example, let $f(y|x)$ and $q(y|x)$ be conditional densities satisfying

$$\{(x, y) : f(x)f(y|x) > 0\} \subset \{(x, y) : q(x)q(y|x) > 0\}.$$

Then for each $x^{(j)}$, we generate $y^{(j)}$ from $q(y|x^{(j)})$ and let

$$\widetilde{w}^{(j)} = \frac{f(x^{(j)})f(y^{(j)}|x^{(j)})}{q(x^{(j)})q(y^{(j)}|x^{(j)})} = w^{(j)} \frac{f(y^{(j)}|x^{(j)})}{q(y^{(j)}|x^{(j)})}.$$

4.4 Importance Sampling

- – It is easy to show that **(Why?)**

$$\frac{\sum_{j=1}^{m} \widetilde{w}^{(j)} h(x^{(j)})}{\sum_{j=1}^{m} \widetilde{w}^{(j)}} \xrightarrow{a.s.} \int h(x) f(x)\, dx$$

  for any h with finite $\int |h(x)| f(x)\, dx$. Then $\{(x^{(j)}, \widetilde{w}^{(j)})\}_{j=1}^{m}$ is also properly weighted with respect to $f(x)$.

  – Note that $E\big(\widetilde{w}^{(j)} | x^{(j)}\big) = w^{(j)}$, $\{w^{(j)}\}_{j=1}^{m}$ are "better" weights because

$$\begin{aligned} \mathrm{Var}\big(w^{(j)}\big) &= \mathrm{Var}\big[E\big(\widetilde{w}^{(j)} | x^{(j)}\big)\big] \\ &\leq \mathrm{Var}\big(\widetilde{w}^{(j)}\big). \end{aligned}$$

  (*However,* $f(x^{(j)}, y^{(j)})$ *or* $\widetilde{w}^{(j)}$ *may be easier to compute than* $f(x^{(j)})$ *or* $w^{(j)}$ *in some cases.*)

4.5 Sequential Importance Sampling

- We want to generate **high dimensional** random samples $x_{0:T} = (x_0, x_1, \cdots, x_T)$ from the target distribution $p(x_{0:T})$ (p is a PDF/PMF).
  - Some sampling methods can not be applied to generate high dimensional samples, for example, the inverse CDF transform method.
  - We can write the target distribution as

    $$p(x_{0:T}) = p(x_0)p(x_1|x_0)\cdots p(x_T|x_{0:T-1}).$$

    If we can **sequentially** generate

    $$x_0^{(j)} \sim p(x_0), \quad x_1^{(j)} \sim p(x_1|x_0^{(j)}), \quad \cdots, \quad x_T^{(j)} \sim p(x_T|x_{0:T-1}^{(j)}),$$

    then $(x_0^{(j)}, \cdots, x_T^{(j)}) \sim p(x_{0:T})$.
  - In many cases, it is not easy to draw samples from $p(x_t|x_{0:t-1})$.

- **Example: Optimal Trading Path.** Let x_t be the holding position of a financial asset in shares at time t, then $x_{0:T} = (x_0, x_1, \cdots, x_T)$ forms a trading path.
  - We want to find an “optimal” trading path to maximize utility function

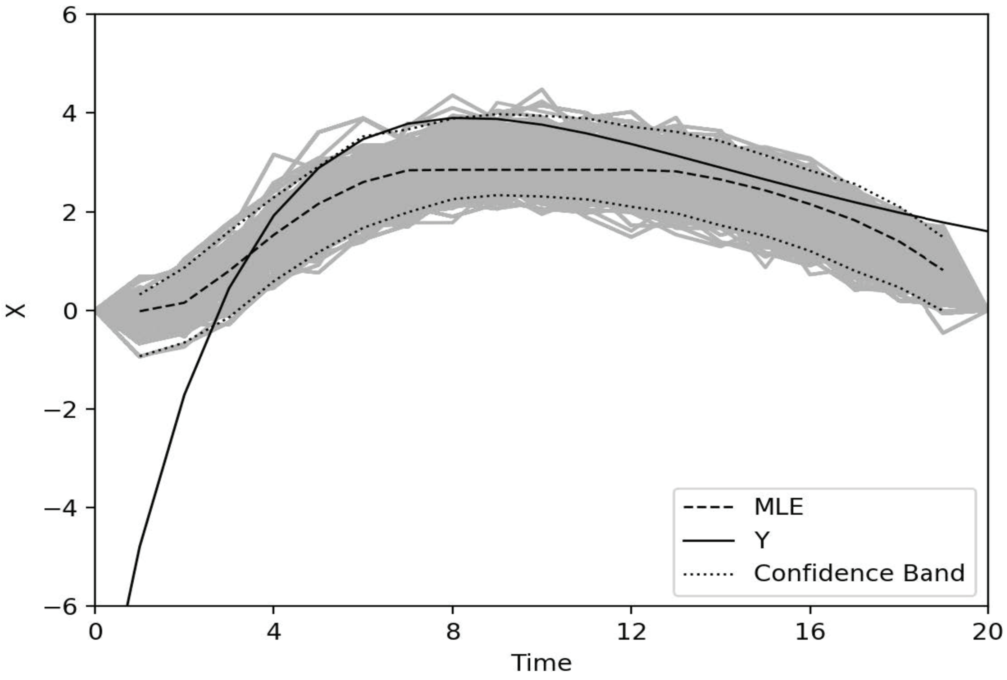
    $$u(x_{0:T}) = -\sum_{t=1}^{T-1} l(y_t - x_t) - \sum_{t=1}^{T} c(x_t - x_{t-1}).$$

    subject to $x_0 = 0$ and $x_T = 0$.
  - $(y_1, \cdots, y_{T-1})$ is the ‘optimal” trading path in an ideal world without trading costs. It can be obtained through maximizing the risk-adjusted expected return using historical data. ($y_{1:T-1}$ **is known.**)
  - $l(\cdot)$ is the utility loss due to the departure of the trading path $x_{1:T-1}$ from the ideal path $y_{1:T-1}$.
  - $c(\cdot)$ denotes the trading costs.

4.5 Sequential Importance Sampling

- – Generally, we can not find analytical solution of the optimal trading path. We consider generating samples from $p(x_{0:T}) \propto \exp\{u(x_{0:T})/\tau\}$ subject to $x_0 = 0$ and $x_T = 0$, where $\tau > 0$ is called the *temperature*.

  – It is not easy to obtain $p(x_t | x_0 = 0, x_{1:t-1}, x_T = 0)$ in this example.



Trading Path Samples

4.5 Sequential Importance Sampling

- **Chain Structured Model:** A model of $x_{0:T}$ is called a *chain structured model* or *Markovian structured model* if its distribution can be written as

$$p(x_{0:T}) \propto \exp\Big\{-H(x_{0:T})\Big\} = \exp\Big\{-\sum_{t=1}^{T} h_t(x_{t-1}, x_t)\Big\}.$$

– Such a model has the following **Markovian property**:

$$\begin{aligned} p\big(x_i | x_{0:(i-1)}, x_{(i+1):T}\big) &= \frac{\exp\Big\{-\sum_{t=1}^{T} h_t(x_{t-1}, x_t)\Big\}}{\int \exp\Big\{-\sum_{t=1}^{T} h_t(x_{t-1}, x_t)\Big\}\, dx_i} \\ &= \frac{\exp\Big\{-h_i(x_{i-1}, x_i) - h_{i+1}(x_i, x_{i+1})\Big\}}{\int \exp\Big\{-h_i(x_{i-1}, x_i) - h_{i+1}(x_i, x_{i+1})\Big\}\, dx_i} \\ &= p\big(x_i | x_{i-1}, x_{i+1}\big). \end{aligned}$$

– The trading path example is a chain structured model with

$$h_t(x_{t-1}, x_t) = \big[l(y_t - x_t) + c(x_t - x_{t-1})\big]/\tau.$$

4.5 Sequential Importance Sampling

- Assume that in a chain structured mode, x_t takes value in **a finite set** $\mathcal{S}_t = \{s_{t,1}, \cdots, s_{t,K_t}\}$.
- For simplicity, we assume $\mathcal{S}_t = \mathcal{S} = \{s_1, \cdots, s_K\}$ for all t.
- We can find the "optimal" path

$$\begin{aligned} x_{0:T}^* &= \arg \min_{x_0,\cdots,x_T \in \mathcal{S}} H(x_{0:T}) \\ &= \arg \min_{x_0,\cdots,x_T \in \mathcal{S}} \sum_{t=1}^{T} h_t(x_{t-1}, x_t) \\ &= \arg \max_{x_0,\cdots,x_T \in \mathcal{S}} p(x_{0:T}) \end{aligned}$$

using the Viterbi algorithm (Viterbi, 1967).

4.5 Sequential Importance Sampling

- **Viterbi Algorithm:**
  - For $x_1 = s_1, \cdots, s_K$, calculate $m_1(x_1) = \min_{x_0 \in \mathcal{S}} h_1(x_0, x_1)$.
  - For $t = 2, 3, \cdots, T$, recursively compute

$$m_t(x_t) \overset{\triangle}{=} \min_{x_0,\cdots,x_{t-1}\in\mathcal{S}} \sum_{s=1}^{t} h_s(x_{s-1}, x_s)$$

$$= \min_{x_{t-1}\in\mathcal{S}} \left[\min_{x_0,\cdots,x_{t-2}\in\mathcal{S}} \sum_{s=1}^{t-1} h_s(x_{s-1}, x_s) + h_t(x_{t-1}, x_t) \right] = \min_{x_{t-1}\in\mathcal{S}} \left[m_{t-1}(x_{t-1}) + h_t(x_{t-1}, x_t) \right]$$

for $x_t = s_1, \cdots, s_K$.

  - At time T, output

$$\min_{x_T\in\mathcal{S}} m_T(x_T) = \min_{x_0,\cdots,x_{t-1}\in\mathcal{S}} \sum_{s=1}^{T} h_s(x_{s-1}, x_s) = \min_{x_0,\cdots,x_T\in\mathcal{S}} H(x_{0:T}).$$

- **Q: How to output the "optimal" path**

$$x_{0:T}^{*} = \arg \min_{x_0,\cdots,x_T\in\mathcal{S}} \sum_{s=1}^{T} h_s(x_{s-1}, x_s)?$$

4.5 Sequential Importance Sampling

- – **Expectation Calculation:** Suppose that we want to calculate

$$E_p\big[g(x_i)\big] = \frac{\sum_{x_0,\cdots,x_T} \exp\Big\{-\sum_{t=1}^T h_t(x_{t-1},x_t)\Big\} g(x_i)}{\sum_{x_0,\cdots,x_T} \exp\Big\{-\sum_{t=1}^T h_t(x_{t-1},x_t)\Big\}}$$

$$\triangleq \frac{A}{B}.$$

We can calculate B (or A) using the following algorithm (we can use the same method to calculate A).

4.5 Sequential Importance Sampling

- – **Algorithmic steps to compute B:**

  $\ast$ For $x_1 = s_1, \cdots, s_K$, compute $V_1(x_1) = \sum_{x_0 \in \mathcal{S}} \exp\big\{-h_1(x_0, x_1)\big\}$.

  $\ast$ For $t = 2, 3, \cdots, T$, recursively compute

$$\begin{aligned} V_t(x_t) &\triangleq \sum_{x_0, \cdots, x_{t-1} \in \mathcal{S}} \exp\Big\{-\sum_{s=1}^{t} h_s(x_{s-1}, x_s)\Big\} \\ &= \sum_{x_{t-1} \in \mathcal{S}} \Bigg[\exp\big\{-h_t(x_{t-1}, x_t)\big\} \sum_{x_0, \cdots, x_{t-2} \in \mathcal{S}} \exp\Big\{-\sum_{s=1}^{t-1} h_s(x_{s-1}, x_s)\Big\}\Bigg] \\ &= \sum_{x_{t-1} \in \mathcal{S}} \Big[V_{t-1}(x_{t-1}) \exp\big\{-h_t(x_{t-1}, x_t)\big\}\Big] \end{aligned}$$

  for $x_t = s_1, \cdots, s_K$.

  $\ast$ At time T, output

$$B = \sum_{x_T \in \mathcal{S}} V_T(x_T).$$

4.5 Sequential Importance Sampling

- – **Exact Simulation:** Note that

$$p(x_{0:T}) = p(x_T)p(x_{T-1}|x_T)\cdots p(x_0|x_{1:T}).$$

We can generate $x_{0:T}^{(j)} = (x_0^{(j)}, \cdots, x_T^{(j)})$ from $p(x_{0:T})$ as follows.

$*$ Generate $x_T^{(j)}$ from $p(x_T) = V_T(x_T)/B$.

$*$ For $t = T-1, \cdots, 0$, recursively generate $x_t^{(j)}$ from

$$p(x_t|x_{(t+1):T}^{(j)}) = \frac{p(x_t, x_{(t+1):T}^{(j)})}{p(x_{(t+1):T}^{(j)})}$$

$$= \frac{\sum_{x_0,\cdots,x_{t-1}\in\mathcal{S}} \exp\Big\{-\sum_{s=1}^{t} h_s(x_{s-1},x_s) - h_{t+1}(x_t, x_{t+1}^{(j)}) - \sum_{s=t+2}^{T} h_s(x_{s-1}^{(j)}, x_s^{(j)})\Big\}}{\sum_{x_0,\cdots,x_t\in\mathcal{S}} \exp\Big\{-\sum_{s=1}^{t} h_s(x_{s-1},x_s) - h_{t+1}(x_t, x_{t+1}^{(j)}) - \sum_{s=t+2}^{T} h_s(x_{s-1}^{(j)}, x_s^{(j)})\Big\}}$$

$$= \frac{V_t(x_t)\exp\big\{-h_{t+1}(x_t, x_{t+1}^{(j)})\big\}}{\sum_{x_t\in\mathcal{S}} V_t(x_t)\exp\big\{-h_{t+1}(x_t, x_{t+1}^{(j)})\big\}} = \frac{V_t(x_t)\exp\big\{-h_{t+1}(x_t, x_{t+1}^{(j)})\big\}}{V_{t+1}(x_{t+1}^{(j)})}.$$

4.5 Sequential Importance Sampling

- **State Space Model:** A state space model consists of a **latent** state variable sequence $\{x_t, t = 0, 1, \cdots\}$ and observations $\{y_t, t = 1, 2, \cdots\}$. The model is defined by

$$\begin{aligned} \text{state equation} &: \quad x_t \sim p(x_t \,|\, x_{0:t-1}, y_{1:t-1}) = p(x_t \,|\, x_{t-1}) := g_t(x_t \,|\, x_{t-1}), \\ \text{observation equation} &: \quad y_t \sim p(y_t \,|\, x_{0:t}, y_{1:t-1}) = p(y_t \,|\, x_t) := \zeta_t(y_t \,|\, x_t). \end{aligned}$$

The joint density of this model can be calculated by

$$\begin{aligned} p(x_{0:t}, y_{1:t}) &= p(x_0) \prod_{s=1}^{t} p(x_s, y_s \,|\, x_{0:s-1}, y_{1:s-1}) \\ &= p(x_0) \prod_{s=1}^{t} p(x_s \,|\, x_{0:s-1}, y_{1:s-1}) p(y_s \,|\, x_{0:s}, y_{1:s-1}) = g_0(x_0) \prod_{s=1}^{t} g_s(x_s \,|\, x_{s-1}) \zeta_s(y_s \,|\, x_s). \end{aligned}$$

$$\begin{tikzcd}
\cdots & y_{t-1} & y_t & y_{t+1} & \cdots \\
\cdots \arrow[r] & x_{t-1} \arrow[u] \arrow[r] & x_t \arrow[u] \arrow[r] & x_{t+1} \arrow[u] \arrow[r] & \cdots
\end{tikzcd}$$

4.5 Sequential Importance Sampling

- We want to make inference of the unobservable states x_t given the observations $y_1, y_2, \cdots$.
  - *Filtering*: estimate $p(x_t \mid y_{1:t})$ or $E(x_t \mid y_{1:t})$.
  - *Prediction*: estimate $p(x_{t+\Delta} \mid y_{1:t})$ or $E(x_{t+\Delta} \mid y_{1:t})$ for $\Delta > 0$.
  - *Smoothing*: estimate $p(x_{t-\delta} \mid y_{1:t})$ or $E(x_{t-\delta} \mid y_{1:t})$ for $\delta > 0$.
  - $E(x_t \mid y_{1:t})$ is the "best" function of $y_{1:t}$ to estimate x_t in terms of MSE.
  - $E(x_t \mid y_{1:t+\delta})$ is a better estimator for x_t than $E(x_t \mid y_{1:t})$.
  - We have

$$\begin{aligned} E(x_t \mid y_{1:t}) &= \int x_t \, p(x_{0:t} \mid y_{1:t}) \, dx_{0:t} \\ &\propto \int x_t \, p(x_{0:t}, y_{1:t}) \, dx_{0:t} = \int x_t \, g_0(x_0) \prod_{s=1}^{t} g_s(x_s \mid x_{s-1}) \zeta_s(y_s \mid x_s) \, dx_{0:t}, \end{aligned}$$

which does not have a closed-form solution in most cases.

- **Example: Target Tracking.** Consider a target moving with random acceleration on a plane. The state equation can be written as

$$\begin{pmatrix} x_{t,1} \\ x_{t,2} \\ v_{t,1} \\ v_{t,2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & T_0 & 0 \\ 0 & 1 & 0 & T_0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \\ v_{t-1,1} \\ v_{t-1,2} \end{pmatrix} + \begin{pmatrix} T_0^2/2 & 0 \\ 0 & T_0^2/2 \\ T_0 & 0 \\ 0 & T_0 \end{pmatrix} \begin{pmatrix} u_{t,1} \\ u_{t,2} \end{pmatrix},$$

where $(x_{t,1}, x_{t,2})$ and $(v_{t,1}, v_{t,2})$ are the position and velocity of the target, respectively, T_0 is the time duration between two observations, and $u_t = (u_{t,1}, u_{t,2})$ is the random acceleration. The observation is

$$\begin{pmatrix} y_{t,1} \\ y_{t,2} \end{pmatrix} = \begin{pmatrix} x_{t,1} \\ x_{t,2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{t,1} \\ \varepsilon_{t,2} \end{pmatrix},$$

which is a noised measurement of the target location.

4.5 Sequential Importance Sampling

- **Example: Wireless Communication.** In a digital wireless communication problem, the received signal sequence $\{y_t\}$ is modelled as

$$y_t = \xi_t s_t + v_t,$$

where $\{\xi_t\}$ is the transmitted channel, $s_t \in \{-1, 1\}$ is the transmitted digital signal, $\{v_t\}$ are i.i.d. noises following the $N(0, \sigma^2)$ distribution.

  – The latent state is $x_t = (\xi_t, s_t)$.

  – Assume that $\{s_t\}$ and $\{\xi_t\}$ are independent,

$$\xi_t = \rho \xi_{t-1} + u_t,$$

where $u_t \sim N(0, \delta^2)$, and

$$p(s_t = -1 \mid s_{0:t-1}) = p(s_t = 1 \mid s_{0:t-1}) = 0.5.$$

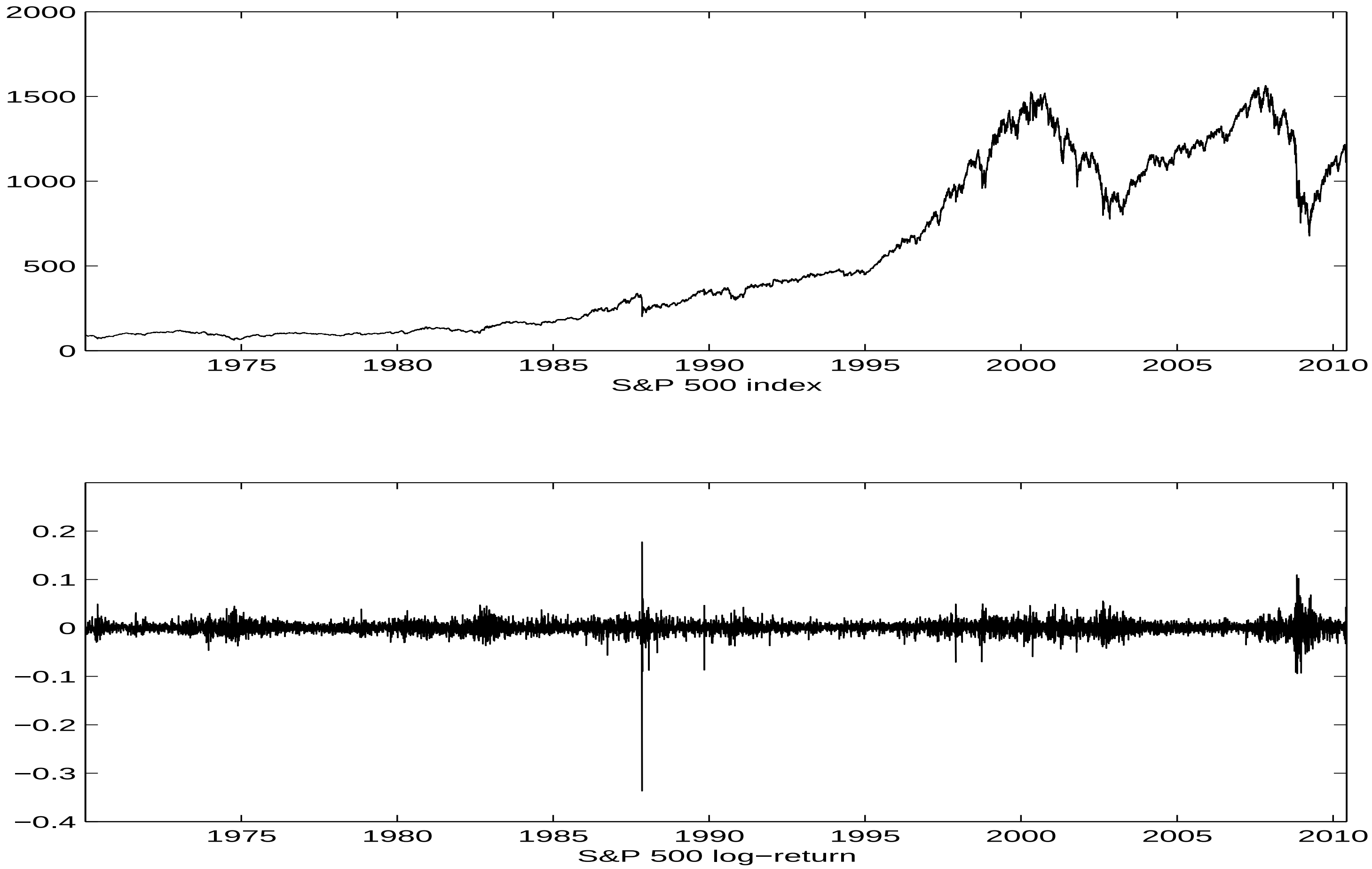
  – We want to estimate $p(s_t = 1 \mid y_1, \cdots, y_t, \cdots, y_{t+h}) = \frac{1}{2} + \frac{1}{2} E(s_t \mid y_{1:t+h})$.

4.5 Sequential Importance Sampling

- **Example: Stochastic Volatility Model.** Let $y_t := \log(P_t/P_{t-1})$ be the observed log-return of a financial asset at time t, where P_t is the price at time t.
  - Assume that y_t follows a normal distribution $N(0, \sigma_t^2)$.
  - The variance of y_t is an unobservable state variable. Assume that $\{\log \sigma_t^2\}$ follows an autoregressive (AR(1)) process.
  - We have the following state space model

$$\begin{aligned} \text{state equation} \ :& \quad \log \sigma_t^2 = \alpha + \beta \log \sigma_{t-1}^2 + u_t, \\ \text{observation equation} \ :& \quad y_t | \sigma_t^2 \sim N(0, \sigma_t^2), \end{aligned}$$

where $\beta > 0$ and $u_t \sim N(0, \delta^2)$.



Price Series and Log-return of S&P 500 Index

- **Linear-Gaussian State Space Model:** Consider the state space model:

$$\begin{aligned}\text{state equation} &: \quad x_t = c + Ax_{t-1} + u_t,\\ \text{observation equation} &: \quad y_t = d + Bx_t + v_t.\end{aligned}$$

  - Here x_t and y_t are random variables/vectors, where x_t is the latent state and y_t is the observation at time t. $u_t \sim N(0, \Sigma_{uu})$ and $v_t \sim N(0, \Sigma_{vv})$ are independent noises. Also assume that $x_0 \sim N(\mu_0, \Sigma_0)$.
  - μ_0, Σ_0, A, B, c, d, Σ_{uu}, and Σ_{vv} are known.
  - It is easy to show that $(x_{0:t}, y_{1:t})$ follows a high-dimensional Gaussian distribution. Then the conditional distribution $p(x_t|y_{1:t})$ is also a Gaussian distribution, denoted by $N(\mu_t, \Sigma_t)$.
  - If we want to find $p(x_t|y_{1:t})$ or $E(x_t|y_{1:t})$, we only need to determine μ_t and Σ_t.

4.5 Sequential Importance Sampling

- **Multivariate Gaussian Distribution:** Suppose X and Y are $n \times 1$ and $m \times 1$ random vectors, respectively. $Z = (X', Y')'$ follows multivariate normal distribution $N(\mu, \Sigma)$ with joint pdf

$$f_{XY}(x, y) = \frac{1}{\sqrt{det[2\pi\Sigma]}} \exp\left\{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right\},$$

where $z = (x', y')'$, $\mu = (\mu_X', \mu_Y')'$,

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}.$$

  - Given y, $f_{X|Y}(x \mid Y = y)$ follows the normal distribution

$$N\left(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y), \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\right).$$

  - **We can verify that** $f_{XY}(x, y) = f_{X|Y}(x \mid y) f_Y(y)$**.** (*You may need to use the Schur complement to prove it.*)

- **Schur Complement.** Let I_n be the $n \times n$-identity matrix. vector. Because

$$\begin{pmatrix} I_n & -\Sigma_{XY}\Sigma_{YY}^{-1} \\ 0 & I_m \end{pmatrix} \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} \begin{pmatrix} I_n & 0 \\ -\Sigma_{YY}^{-1}\Sigma_{YX} & I_m \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} & 0 \\ 0 & \Sigma_{YY} \end{pmatrix},$$

let $F = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$, then

$$\begin{aligned} &\begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} I_n & 0 \\ -\Sigma_{YY}^{-1}\Sigma_{YX} & I_m \end{pmatrix} \begin{pmatrix} F^{-1} & 0 \\ 0 & \Sigma_{YY}^{-1} \end{pmatrix} \begin{pmatrix} I_n & -\Sigma_{XY}\Sigma_{YY}^{-1} \\ 0 & I_m \end{pmatrix} \\ &= \begin{pmatrix} F^{-1} & -F^{-1}\Sigma_{XY}\Sigma_{YY}^{-1} \\ -\Sigma_{YY}^{-1}\Sigma_{YX}F^{-1} & \Sigma_{YY}^{-1} + \Sigma_{YY}^{-1}\Sigma_{YX}F^{-1}\Sigma_{XY}\Sigma_{YY}^{-1} \end{pmatrix}. \end{aligned}$$

- **Kalman Filter:** Suppose at time $t-1$, we already obtain $p(x_{t-1}|y_{1:t-1}) \sim N(\mu_{t-1}, \Sigma_{t-1})$.
  - At time t, $p(x_t, y_t \mid y_{1:t-1}) \sim N(\mu, \Sigma)$, where

$$\mu := (\mu_X', \mu_Y')' = \Big((c + A\mu_{t-1})', \big[d + B(c + A\mu_{t-1})\big]'\Big)'$$

and

$$\Sigma \;=\; \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} = \begin{pmatrix} A\Sigma_{t-1}A' + \Sigma_{uu} & \big(A\Sigma_{t-1}A' + \Sigma_{uu}\big)B' \\ B\big(A\Sigma_{t-1}A' + \Sigma_{uu}\big) & B\big(A\Sigma_{t-1}A' + \Sigma_{uu}\big)B' + \Sigma_{vv} \end{pmatrix}.$$

  - Then $p(x_t|y_{1:t-1}, y_t) \sim N(\mu_t, \Sigma_t)$ with

$$\mu_t = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y_t - \mu_Y) \quad \text{and} \quad \Sigma_t = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}.$$

- **Remark:** Obviously, we have

$$E(x_t \,|\, y_{1:t}) = \mu_t \quad \text{and} \quad E(x_{t+1} \,|\, y_{1:t}) = c + A\mu_t.$$

4.5 Sequential Importance Sampling

- For the state space models, we can only find analytic solutions of $p(x_t \mid y_{1:t})$ or $E(x_t \mid y_{1:t})$ in some special cases.
  - Case 1: For any t, x_t takes value in a finite set.
  - Case 2: the linear-Gaussian state space model.
- In most cases,

$$\begin{aligned} p(x_t \mid y_{1:t}) &\propto \int p(x_{0:t}, y_{1:t})\, dx_{0:t-1} \\ &= \int g_0(x_0) \prod_{s=1}^{t} g_s(x_s \mid x_{s-1}) \zeta_s(y_s \mid x_s)\, dx_{0:t-1} \end{aligned}$$

does not have a closed form.

4.5 Sequential Importance Sampling*

- **Importance Sampling for State Space Model:** We consider generating samples $x_{0:t}^{(1)}, \cdots, x_{0:t}^{(m)}$ from a trial distribution $q(x_{0:t})$ and let

$$\begin{aligned} w_t^{(j)} &= \frac{p(x_{0:t}^{(j)} \mid y_{1:t})}{q(x_{0:t}^{(j)})} \\ &\propto \frac{p(x_{0:t}^{(j)}, y_{1:t})}{q(x_{0:t}^{(j)})} = \frac{g_0(x_0^{(j)}) \prod_{s=1}^{t} g_s(x_s^{(j)} \mid x_{s-1}^{(j)}) \zeta_s(y_s \mid x_s^{(j)})}{q(x_{0:t}^{(j)})}. \end{aligned}$$

  Then

$$p(x_{0:t} \mid y_{1:t}) \approx \sum_{j=1}^{m} \frac{w_t^{(j)}}{\sum_{k=1}^{m} w_t^{(k)}} \delta(x_{0:t} - x_{0:t}^{(j)})$$

  and

$$E(x_t \mid y_{1:t}) \approx \sum_{j=1}^{m} \frac{w_t^{(j)} x_t^{(j)}}{\sum_{k=1}^{m} w_t^{(k)}} = \frac{\sum_{j=1}^{m} w_t^{(j)} x_t^{(j)}}{\sum_{k=1}^{m} w_t^{(k)}}$$

- **Remark:** Since $p(x_t \mid y_{1:t})$ may not have a closed-form, we can not use the "marginalized" weight $p(x_t^{(j)} \mid y_{1:t}) / q(x_t^{(j)})$.

4.5 Sequential Importance Sampling*

- **Sequential Importance Sampling:** Generate samples $x_{0:t}^{(j)}, j = 1, \cdots, m$, as follows.
  - At $t = 0$, generate $x_0^{(j)}$ from $q(x_0)$ and let $w_0^{(j)} = g_0(x_0^{(j)})/q(x_0^{(j)})$.
  - For $t = 1, 2, \cdots$,
    - (Sampling.) Generate $x_t^{(j)}$ from distribution $q(x_t \mid x_{0:t-1}^{(j)})$.
    - (Updating Weights.) Let

$$w_t^{(j)} = w_{t-1}^{(j)} \eta_t^{(j)},$$

where

$$\eta_t^{(j)} := \frac{g_t(x_t^{(j)} \mid x_{t-1}^{(j)}) \zeta_t(y_t \mid x_t^{(j)})}{q(x_t^{(j)} | x_{0:t-1}^{(j)})}$$

is called the *incremental weight*.

- **Remarks:**
  - The weighted sample set $\{(x_{0:t}^{(j)}, w_t^{(j)})\}_{j=1}^{m}$ obtained at time t can be used at time $t+1$.
  - The sequential importance sampling (SIS) method is often used for "on-line" estimation, that is, estimate $E(x_t|y_{1:t})$ for $t = 1, 2, \cdots$ recursively, without restarting from $t = 0$.
  - The sample $x_{0:t}^{(j)}$ is built up **sequentially** according to a series of **low dimensional** conditional distributions

$$q(x_{0:t}) = q(x_0)q(x_1|x_0)\cdots q(x_t|x_{0:t-1}).$$

  - At each time t, the "correct" weight should be

$$w_t^{(j)} = \frac{g_0(x_0^{(j)})\prod_{s=1}^{t} g_s(x_s^{(j)} \mid x_{s-1}^{(j)})\zeta_s(y_s \mid x_s^{(j)})}{q(x_0^{(j)})\prod_{s=1}^{t} q(x_s^{(j)} \mid x_{0:s-1}^{(j)})}.$$

4.5 Sequential Importance Sampling*

- – From the algorithm, we have

$$\begin{aligned}
w_t^{(j)} = w_{t-1}^{(j)}\eta_t^{(j)} &= w_{t-1}^{(j)} \frac{g_t(x_t^{(j)} \mid x_{t-1}^{(j)})\zeta_t(y_t \mid x_t^{(j)})}{q(x_t^{(j)}|x_{0:t-1}^{(j)})} \\
&= w_{t-2}^{(j)} \frac{\prod_{s=t-1}^{t} g_s(x_s^{(j)} \mid x_{s-1}^{(j)})\zeta_s(y_s \mid x_s^{(j)})}{\prod_{s=t-1}^{t} q(x_s^{(j)}|x_{0:s-1}^{(j)})} \\
&= \cdots \\
&= \frac{g_0(x_0^{(j)}) \prod_{s=1}^{t} g_s(x_s^{(j)} \mid x_{s-1}^{(j)})\zeta_s(y_s \mid x_s^{(j)})}{q(x_0^{(j)}) \prod_{s=1}^{t} q(x_s^{(j)}|x_{0:s-1}^{(j)})} \\
&= \frac{p(x_{0:t}^{(j)}, y_{0:t})}{q(x_{0:t}^{(j)})} \\
&\propto \frac{p(x_{0:t}^{(j)} \mid y_{0:t})}{q(x_{0:t}^{(j)})}.
\end{aligned}$$

4.5 Sequential Importance Sampling*

- – Therefore, for any function $h(x_{0:t})$ with finite expectation, we have

$$\sum_{j=1}^{m} \frac{w_t^{(j)} h(x_{0:t}^{(j)})}{\sum_{k=1}^{m} w_t^{(k)}} \xrightarrow{a.s.} E\big[h(x_{0:t}) \,|\, y_{1:t}\big].$$

  – Since we often ignore some normalizing constants in the importance weights (*e.g.*, $p(y_{1:t})$), $\sum_{j=1}^{m} w_t^{(j)}$ could be very large or very small when t is large.

  – In practice, we may record $\alpha_t^{(j)} = \log w_t^{(j)}$, $j = 1, \cdots, m$. Let

$$\alpha_{t,\max} = \max\{\alpha_t^{(1)}, \cdots, \alpha_t^{(m)}\}.$$

  Then $E(x_t \,|\, y_{1:t})$ can be estimated by

$$\frac{\sum_{j=1}^{m} \boxed{\exp\{\alpha_t^{(j)} - \alpha_{t,\max}\}} \cdot x_t^{(j)}}{\sum_{j=1}^{m} \exp\{\alpha_t^{(j)} - \alpha_{t,\max}\}} = \frac{\sum_{j=1}^{m} w_t^{(j)} x_t^{(j)}}{\sum_{j=1}^{m} w_t^{(j)}}.$$

4.5 Sequential Importance Sampling*

- – **Choices of the trial distribution** $q(x_t \,|\, x^{(j)}_{0:t-1})$**:**

  (1) Only use the state equation (Gordon et al., 1993).

  · Let $q(x_t \,|\, x^{(j)}_{0:t-1}) = g_t(x_t \,|\, x^{(j)}_{t-1})$.

  · The incremental weight is $\eta^{(j)}_t = \zeta_t(y_t \,|\, x^{(j)}_t)$.

  (2) Use the state equation and the observation equation (Kong et al., 1994; Liu and Chen, 1998).

  · Let $q(x_t \,|\, x^{(j)}_{0:t-1}) = \frac{g_t(x_t \,|\, x^{(j)}_{t-1})\zeta_t(y_t \,|\, x^{(j)}_{t-1}, x_t)}{\int g_t(x_t \,|\, x^{(j)}_{t-1})\zeta_t(y_t \,|\, x^{(j)}_{t-1}, x_t)\, dx_t}$.

  · The incremental weight is $\eta^{(j)}_t = \int g_t(x_t \,|\, x^{(j)}_{t-1})\zeta_t(y_t \,|\, x^{(j)}_{t-1}, x_t) dx_t$.

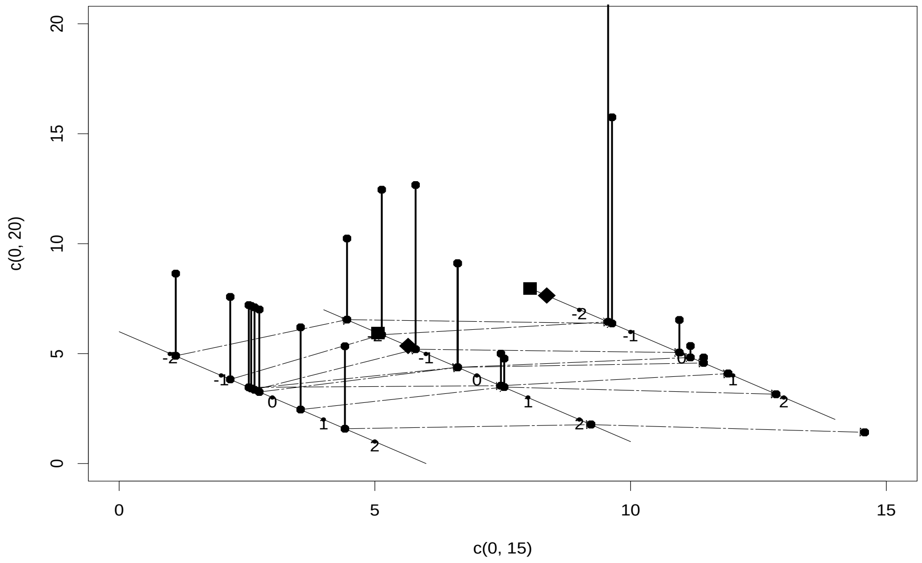
  (3) Only use the observations (Lin et al., 2005) .

  · Let $q(x_t \,|\, x^{(j)}_{0:t-1}) \propto \zeta_t(y_t \,|\, x_t)$.

  · The incremental weight is $\eta^{(j)}_t = g_t(x^{(j)}_t \,|\, x^{(j)}_{t-1})$.

4.5 Sequential Importance Sampling*

- – As t increases, $\mathrm{Var}(w_t^{(j)})$ increases and $w_t^{(j)}$ becomes increasingly skewed, resulting in many unrepresentative samples of $x_{0:t}^{(j)}$. This phenomena is called *sample degeneracy*.



4.5 Sequential Importance Sampling*

- A *resampling step* is often used to deal with the "sample degeneracy" problem.

- **Resampling:** At time t, suppose we have obtained $\{(x_{0:t}^{(j)}, w_t^{(j)}), j = 1, \cdots, m\}$ properly weighted with respect to $p(x_{0:t} \mid y_{1:t})$.

  – For each sample $x_{0:t}^{(j)}$, $j = 1, \cdots, m$, assign a *priority score* $\beta_t^{(j)} > 0$.

  – For $j = 1, \cdots, m$,

    $*$ Choose K_j from $\{1, \cdots, m\}$ with probability $P(K_j = i) = \beta_t^{(i)} / \sum_{l=1}^m \beta_t^{(l)}$.

    $*$ Set $x_{0:t}^{*(j)} = x_{0:t}^{(K_j)}$ and $w_t^{*(j)} = \dfrac{w_t^{(K_j)}}{\beta_t^{(K_j)}/(m^{-1}\sum_{l=1}^m \beta_t^{(l)})}$.

  – Return the new set $\left\{(x_{0:t}^{(j)}, w_t^{(j)})\right\}_{j=1}^m \leftarrow \left\{(x_{0:t}^{*(j)}, w_t^{*(j)})\right\}_{j=1}^m$.

- **Remark:** We often choose $\beta_t^{(j)} = w_t^{(j)}$.

4.5 Sequential Importance Sampling*

- **Multinomial Resampling:** At time t, suppose we have obtained $\{(x_{0:t}^{(j)}, w_t^{(j)}), j = 1, \cdots, m\}$ properly weighted with respect to $p(x_{0:t} \mid y_{1:t})$.
  - For each sample $x_{0:t}^{(j)}$, $j = 1, \cdots, m$, assign a *priority score* $\beta_t^{(j)} > 0$.
  - Generate $U_1, \cdots, U_m$ i.i.d. from the Uniform$(0, 1)$ distribution.
  - For $j = 1, \cdots, m$,
    - $*$ Let $K_j = k$ if
$$\frac{\sum_{j=1}^{k-1} \beta_t^{(j)}}{\sum_{l=l}^{m} \beta_t^{(l)}} < U_j \le \frac{\sum_{j=1}^{k} \beta_t^{(j)}}{\sum_{l=l}^{m} \beta_t^{(l)}}.$$
    - $*$ Set $x_{0:t}^{*(j)} = x_{0:t}^{(K_j)}$ and $w_t^{*(j)} = \frac{w_t^{(K_j)}}{\beta_t^{(K_j)}/(m^{-1}\sum_{l=1}^{m}\beta_t^{(l)})}$.
  - Return the new set $\left\{(x_{0:t}^{(j)}, w_t^{(j)})\right\}_{j=1}^{m} \leftarrow \left\{(x_{0:t}^{*(j)}, w_t^{*(j)})\right\}_{j=1}^{m}$.

- **Systematic Resampling:** At time t, suppose we have obtained $\{(x_{0:t}^{(j)}, w_t^{(j)}), j = 1, \cdots, m\}$ properly weighted with respect to $p(x_{0:t} \mid y_{0:t})$.
  - For each sample $x_{0:t}^{(j)}$, $j = 1, \cdots, m$, assign a *priority score* $\beta_t^{(j)} > 0$.
  - **Generate random number U_1 from the Uniform$(0, 1/m)$ distribution. Let $U_j = U_1 + (j-1)/m$, $j = 2, \cdots, m$.**
  - For $j = 1, \cdots, m$,
    - Let $K_j = k$ if
$$\frac{\sum_{j=1}^{k-1} \beta_t^{(j)}}{\sum_{l=l}^{m} \beta_t^{(l)}} < U_j \leq \frac{\sum_{j=1}^{k} \beta_t^{(j)}}{\sum_{l=l}^{m} \beta_t^{(l)}}.$$
    - Set $x_{0:t}^{*(j)} = x_{0:t}^{(K_j)}$ and $w_t^{*(j)} = \frac{w_t^{(K_j)}}{\beta_t^{(K_j)}/(m^{-1}\sum_{l=1}^{m} \beta_t^{(l)})}$.
  - Return the new set $\left\{(x_{0:t}^{(j)}, w_t^{(j)})\right\}_{j=1}^{m} \leftarrow \left\{(x_{0:t}^{*(j)}, w_t^{*(j)})\right\}_{j=1}^{m}$.

4.5 Sequential Importance Sampling*

- **Remarks:**
  - Resampling tries to remove "bad" samples (with small $\beta_t^{(j)}$) and duplicate "good" samples (with large $\beta_t^{(j)}$) **at intermediate steps.**
  - We often **choose** $\beta_t^{(j)} = w_t^{(j)}$, then the new weight after resampling becomes

    $$w_t^{*(j)} = \frac{w_t^{(K_j)}}{\beta_t^{(K_j)}/(m^{-1}\sum_{l=1}^m \beta_t^{(l)})} = \frac{w_t^{(K_j)}}{w_t^{(K_j)}/(m^{-1}\sum_{l=1}^m w_t^{(l)})} = \frac{1}{m}\sum_{l=1}^m w_t^{(l)},$$

    which is a constant not depending on j. The sample variance of $\left\{w_t^{*(j)}\right\}_{j=1}^m$ is 0.
  - For the multinomial resampling, it is easy to find that

    $$P\big(K_j = k \mid \beta_t^{(l)}, x_{0:t}^{(l)}, w_t^{(l)}, l = 1, \cdots, m\big) = \frac{\beta_t^{(k)}}{\sum_{l=l}^m \beta_t^{(l)}}.$$

4.5 Sequential Importance Sampling*

- – For the **multinomial resampling**,

$$
\begin{aligned}
& E\Big(\frac{1}{m}\sum_{j=1}^{m} w_t^{*(j)} h(x_{0:t}^{*(j)}) \,|\, \beta_t^{(l)}, x_{0:t}^{(l)}, w_t^{(l)}, l = 1, \cdots, m\Big) \\
&= \frac{1}{m}\sum_{j=1}^{m} E\Big(w_t^{*(j)} h(x_{0:t}^{(K_j)}) \,|\, \beta_t^{(l)}, x_{0:t}^{(l)}, w_t^{(l)}, l = 1, \cdots, m\Big) \\
&= \frac{1}{m}\sum_{j=1}^{m} \Big\{ \sum_{k=1}^{m} P\big(K_j = k \,|\, \beta_t^{(l)}, x_{0:t}^{(l)}, w_t^{(l)}, l = 1, \cdots, m\big) \\
&\qquad \cdot \frac{w_t^{(k)}}{\beta_t^{(k)}/(m^{-1}\sum_{l=1}^{m}\beta_t^{(l)})} \cdot h(x_{0:t}^{(k)}) \Big\} \\
&= \frac{1}{m}\sum_{j=1}^{m} \Big\{ \sum_{k=1}^{m} \frac{\beta_t^{(k)}}{\sum_{l=l}^{m}\beta_t^{(l)}} \cdot \frac{w_t^{(k)}(m^{-1}\sum_{l=1}^{m}\beta_t^{(l)})}{\beta_t^{(k)}} \cdot h(x_{0:t}^{(k)}) \Big\} \\
&= \frac{1}{m}\sum_{j=1}^{m} \Big\{ m^{-1}\sum_{k=1}^{m} w_t^{(k)} h(x_{0:t}^{(k)}) \Big\} = \frac{1}{m}\sum_{k=1}^{m} w_t^{(k)} h(x_{0:t}^{(k)}) \xrightarrow{a.s.} E\big[h(x_{0:t}) \,|\, y_{1:t}\big].
\end{aligned}
$$

- – For the **multinomial resampling**, we can show that

$$\frac{1}{m}\sum_{j=1}^{m} w_t^{*(j)} h(x_{0:t}^{*(j)}) \xrightarrow{a.s.} E\big[h(x_{0:t}) \,|\, y_{1:t}\big]$$

and

$$\frac{1}{m}\sum_{j=1}^{m} w_t^{*(j)} \cdot 1 \xrightarrow{a.s.} 1.$$

Hence,

$$\frac{\sum_{j=1}^{m} w_t^{*(j)} h(x_{0:t}^{*(j)})}{\sum_{j=1}^{m} w_t^{*(j)}} \xrightarrow{a.s.} E\big[h(x_{0:t}) \,|\, y_{1:t}\big].$$

The sample set $\big\{(x_{0:t}^{*(j)}, w_t^{*(j)})\big\}_{j=1}^{m}$ obtained after resampling is also properly weighted with respect to $p(x_{0:t} \,|\, y_{1:t})$.

4.5 Sequential Importance Sampling*

- – For the **systematic resampling**, we can also prove that

$$E\Big(\frac{1}{m}\sum_{j=1}^{m} w_t^{*(j)} h(x_{0:t}^{*(j)}) \,\big|\, \beta_t^{(l)}, x_{0:t}^{(l)}, w_t^{(l)}, l=1,\cdots,m\Big) = \frac{1}{m}\sum_{k=1}^{m} w_t^{(k)} h(x_{0:t}^{(k)}).$$

and

$$\frac{\sum_{j=1}^{m} w_t^{*(j)} h(x_{0:t}^{*(j)})}{\sum_{j=1}^{m} w_t^{*(j)}} \xrightarrow{a.s.} E\big[h(x_{0:t}) \,|\, y_{1:t}\big].$$

$*$ *To prove the conclusion, note that* $U_j \sim$ *Uniform*$\big(\frac{j-1}{m}, \frac{j}{m}\big)$ *and*

$$P\big(K_j = k \,|\, \beta_t^{(l)}, x_{0:t}^{(l)}, w_t^{(l)}, l=1,\cdots,m\big) = \frac{\Big|\Big(\frac{\sum_{j=1}^{k-1}\beta_t^{(j)}}{\sum_{l=l}^{m}\beta_t^{(l)}}, \frac{\sum_{j=1}^{k}\beta_t^{(j)}}{\sum_{l=l}^{m}\beta_t^{(l)}}\Big) \cap \big(\frac{j-1}{m}, \frac{j}{m}\big)\Big|}{1/m},$$

where $|\cdot|$ *denotes the length of the set. Then*

$$\sum_{j=1}^{m} P\big(K_j = k \,|\, \beta_t^{(l)}, x_{0:t}^{(l)}, w_t^{(l)}, l=1,\cdots,m\big) = \frac{m\beta_t^{(k)}}{\sum_{l=l}^{m}\beta_t^{(l)}}.$$

4.5 Sequential Importance Sampling*

- – In the resmapling step, to estimate $E\big[h(x_{0:t})\,|\,y_{1:t}\big]$, we need to resample the whole path $x_{0:t}^{(j)}$, not only $x_t^{(j)}$. (In practice, if we only want to estimate $E\big[h(x_t)\,|\,y_{1:t}\big]$, we may only record $x_t^{(j)}$.)

  – For the systematic resampling, we can find $K_1, \cdots, K_m$ in $O(m)$ comparisons. **(Why?)**

  – **Resampling introduces extra variation to the current step** (systematic resampling is better than multinomial resampling), **but it will benefit future sampling steps.**

  – At each time t, we should make inference of x_t before resampling.

- **Sequential Monte Carlo / Particle Filter:**
  - At $t = 0$, generate $x_0^{(j)}$ from $q(x_0)$ and let $w_0^{(j)} = g_0(x_0^{(j)})/q(x_0^{(j)})$.
  - For $t = 1, 2, \cdots$,
    - $*$ (Sampling.) Generate $x_t^{(j)}$ from distribution $q(x_t \mid x_{0:t-1}^{(j)})$.
    - $*$ (*Updating Weights*:) Set

$$w_t^{(j)} = w_{t-1}^{(j)} \eta_t^{(j)} = w_{t-1}^{(j)} \frac{g_t(x_t^{(j)} \mid x_{t-1}^{(j)}) \zeta_t(y_t \mid x_t^{(j)})}{q(x_t^{(j)} | x_{0:t-1}^{(j)})}.$$

    - $*$ (*Inference*:) Estimate $E(x_t \mid y_{1:t})$ by $\sum_{j=1}^m w_t^{(j)} x_t^{(j)} / \sum_{j=1}^m w_t^{(j)}$.
    - $*$ (*Resampling*:) Resampling with the priority scores $\left\{\beta_t^{(j)}\right\}_{j=1}^m$. (Usually we let $\beta_t^{(j)} = w_t^{(j)}$.)

4.5 Sequential Importance Sampling*

- **Remarks:**

  – Note that $\{(x_{0:t}^{(j)}, w_t^{(j)})\}_{j=1}^m$ obtained at time t is properly weighted with respect to $p(x_{0:t} \mid y_{1:t})$, that is, for any function h with finite expectation,

$$\frac{\sum_{j=1}^m w_t^{(j)} h(x_{0:t}^{(j)})}{\sum_{j=1}^m w_t^{(j)}} \xrightarrow{a.s.} E\big[h(x_{0:t}) \,|\, y_{1:t}\big].$$

  – Smoothing: At time t (before resampling), we can estimate

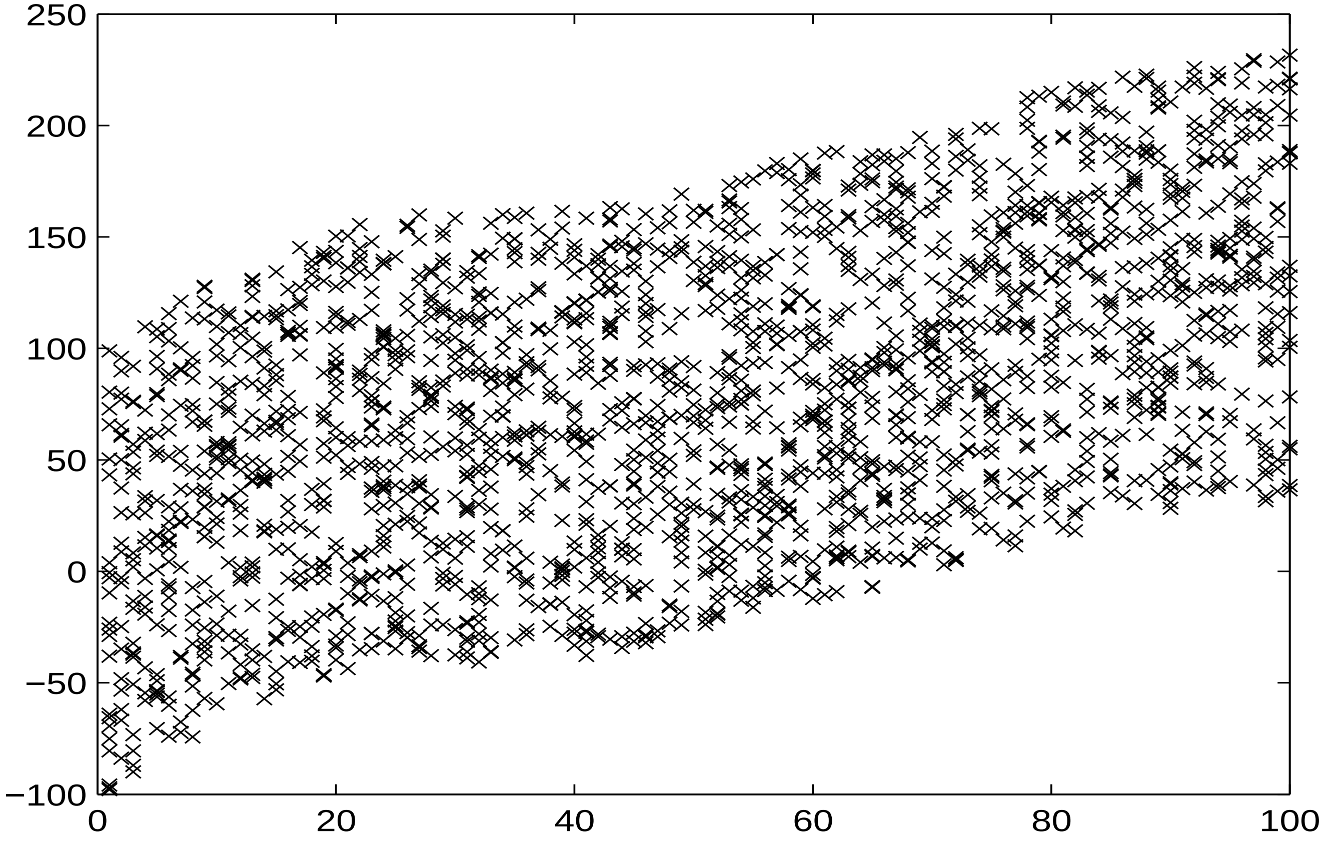
$$E(x_{t-\delta} \,|\, y_{1:t}) \approx \frac{\sum_{j=1}^m w_t^{(j)} x_{t-\delta}^{(j)}}{\sum_{j=1}^m w_t^{(j)}}.$$

  – Prediction: At time t (before resampling), we can estimate

$$\begin{aligned} E(x_{t+1} \,|\, y_{1:t}) \quad &\approx \quad \frac{\sum_{j=1}^m w_t^{(j)} E\big(x_{t+1} \,|\, x_{0:t} = x_{0:t}^{(j)}\big)}{\sum_{j=1}^m w_t^{(j)}} \\ &\xrightarrow{a.s.} \quad E\big[E\big(x_{t+1} \,|\, x_{0:t}\big) \,|\, y_{1:t}\big] \\ &\quad\;\; = E\big[E\big(x_{t+1} \,|\, x_{0:t}, y_{1:t}\big) \,|\, y_{1:t}\big] = E(x_{t+1} \,|\, y_{1:t}). \end{aligned}$$

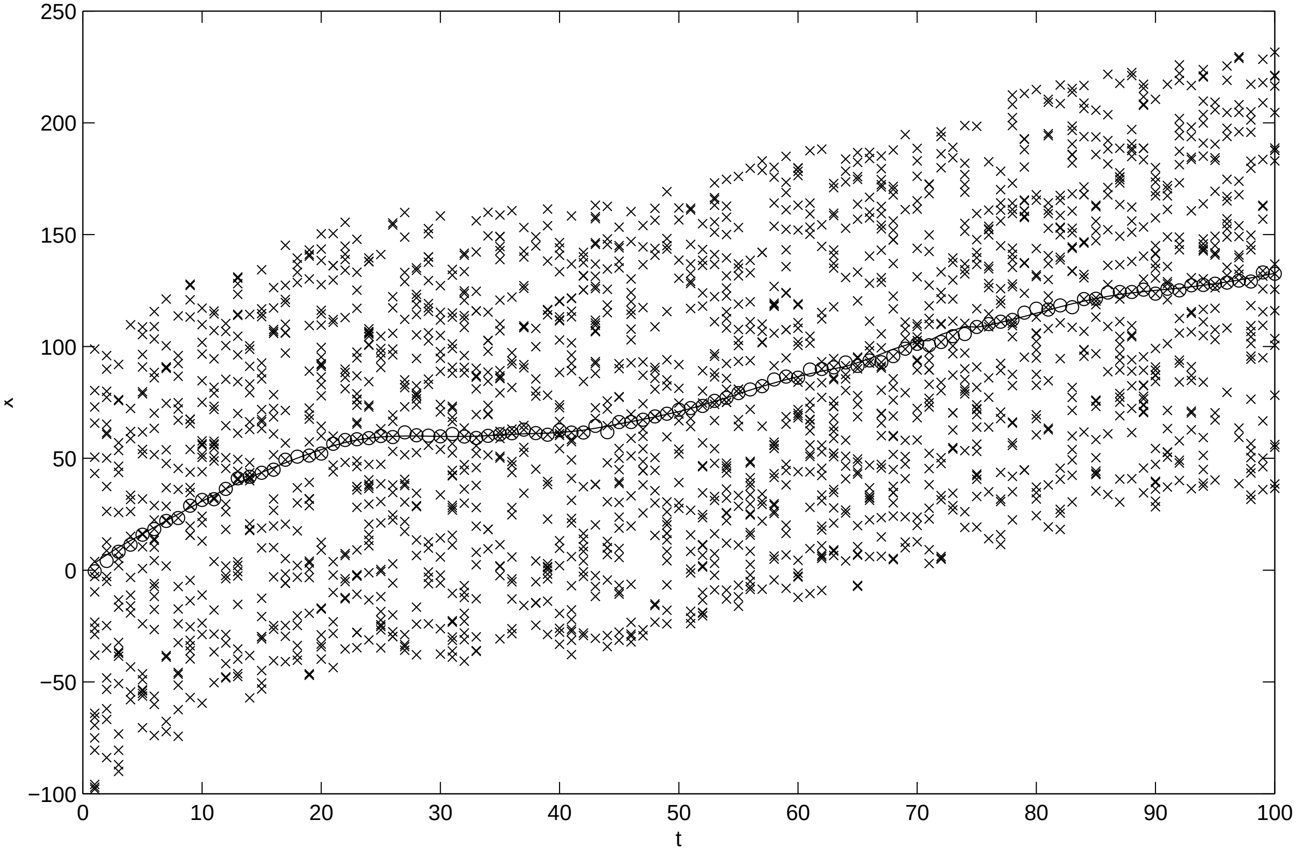
4.5 Sequential Importance Sampling*

- **Example: Target Tracking in Clutter.** Consider the problem of tracking a single target in one dimensional space (Avitzour, 1995).
  - The state equation can be written as
    $$\begin{pmatrix} x_t \\ v_t \end{pmatrix} = \begin{pmatrix} 1 & T_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ v_{t-1} \end{pmatrix} + \begin{pmatrix} T_0^2/2 \\ T_0 \end{pmatrix} u_t,$$
    where x_t and v_t denote the one dimensional location and velocity of the target, respectively; $u_t \sim N(0, \sigma^2)$ is the random acceleration.
  - At each time t, the target can be observed with probability p_d independently.
  - If the target is observed, the observation is $y_t = x_t + \varepsilon_t$, where $\varepsilon_t \sim N(0, \delta^2)$.
  - In additional to the true signal, there are false signals. False signals follow a spatially homogeneous Poisson process with rate λ.



The target is observed with probability $p_d = 0.8$.

4.5 Sequential Importance Sampling*



Estimated Trace (circle: true position, solid line: estimated position.)

Homework

1. A deck of 100 cards (numbered $1, 2, \cdots, 100$) is shuffled and then turned over one card at a time. Say that a "hit" occurs whenever card i is the ith card to be turned over for $i = 1, \cdots, 100$.

   (a) Find the expectation and variance of the total number of hits. (*Note: Let A_i be the event that the ith card to be turned over is card i. Then the total number of hits is $\sum_{i=1}^{100} I(A_i)$.*)

   (b) Write a simulation program to estimate the expectation and variance of the total number of hits. Compare your estimates with the exact answers.

2. Let F be any c.d.f. and let F^- be the generalized inverse of F. Prove that $F\big(F^-(u)\big) \geq u$ and $F\big(F^-(u) - \varepsilon\big) < u$ for any $\varepsilon > 0$, where $0 < u < 1$.

Homework

3. Suppose X_1 and X_2 are two i.i.d. $N(0,1)$ random variables. Let (r,θ) be the polar coordinates of (X_1,X_2). Find the joint PDF of (r,θ). Show that r and θ are independent.

4. Suppose that we want to generate a random variable X whose density function is

$$f(x) = \frac{1}{2}x^2e^{-x}, \quad x > 0$$

by using the rejection method with an exponential distribution having density $g(x) = \lambda e^{-\lambda x}$, $x > 0$. Find the value of λ that minimizes the expected number of iterations of the algorithm used to generate X.

5. Consider a distribution having density

$$f(x) = 30(x^2 - 2x^3 + x^4), \quad 0 \le x \le 1.$$

(1) Develop an algorithm to generate random variables from $f(x)$. Use the algorithm to draw $1,000,000$ random samples.

(2) Plot the histogram of the random samples you generated, and compare it with the density function $f(x)$.

6. Prove that $\text{Var}(\widehat{\Pi}_S)$ is minimized when

$$m_k = m \cdot \frac{a_k \text{Var}_{f_k}^{1/2}\big[h\big(x^{(k,j)}\big)\big]}{\sum_{s=1}^K a_s \text{Var}_{f_s}^{1/2}\big[h\big(x^{(s,j)}\big)\big]}.$$

7. For any two random variables X and Y, prove that

$$\text{Var}(Y) = E\big[\text{Var}(Y|X)\big] + \text{Var}\big[E(Y|X)\big].$$

Homework

8. Suppose $X \sim N(0,4)$. Use $\widehat{\Pi}_2$ with $m = 10,000$ samples to calculate
$$E\left(\frac{X^5}{1+(X-3)^2}\right).$$
Try trial distributions $q_1(x) \sim N(0,1)$, $q_2(x) \sim N(0,4)$ and $q_3(x) \sim N(0,9)$. Repeat the experiment 100 times. Report the mean and variance of the 100 estimates using different trial distributions.

9. Suppose that X follows the standard normal distribution $N(0,1)$.

   (1) Derive an algorithm to calculate $P(X > 4)$. (*Note: You should not use function* $\Phi(4)$ *in this question.*)

   (2) Use the rejection method (*i.e.*, $\widehat{\Pi}_0$) and the importance sampling method ($\widehat{\Pi}_1$ and $\widehat{\Pi}_2$) with $10,000$ samples to calculate $E\big(X|X > 4\big)$. Repeat the experiment 100 times. Report the boxplots of the 100 estimates using different methods. (*Note: For the rejection method, the rejected samples are included in the* $10,000$ *samples.*)

10. In the trading path example, let $T = 20$ and

$$\begin{aligned} c(x_t - x_{t-1}) &= 2\big[(x_t - x_{t-1})^2 + 2|x_t - x_{t-1}|\big], \\ l(y_t - x_t) &= \frac{1}{2}(y_t - x_t)^2, \\ y_t &= 25\exp\{-(t+1)/8\} - 40\exp\{-(t+1)/4\}. \end{aligned}$$

Assume that $x_0 = 0$, $x_T = 0$ and $x_t \in \{-2, -1.9, -1.8 \cdots, 5.8, 5.9, 6\}$ for $t = 1, \cdots, T-1$. Find the optimal trading path and the maximum value of the utility function

$$u(x_{0:T}) = -\sum_{t=1}^{T-1} l(y_t - x_t) - \sum_{t=1}^{T} c(x_t - x_{t-1}).$$

subject to $x_0 = 0$ and $x_T = 0$.

Homework

11. Consider a one-dimensional Ising model with

$$p(x_{0:T}) = \frac{1}{B}\exp\left\{0.1 \cdot (8x_0x_1 + x_1x_2 + \cdots + x_{T-1}x_T)\right\},$$

where $x_i \in \{-1, 1\}$. Let $T = 100$.

(1) Find the exact values of B and $E_p\big(x_0x_1 + x_1x_2 + \cdots + x_{T-1}x_T\big)$.

(2) Use the Monte Carlo method to compute $E_p\big(x_0x_1+x_1x_2+\cdots+x_{T-1}x_T\big)$.

12. Let X and Y be two random variables. Prove that $E(X \mid Y)$ is the best function of Y to estimate X in terms of mean squared error, that is,

$$E(X \mid Y) = \arg\min_g E\big[X - g(Y)\big]^2,$$

where the minimization is over all measurable and square-integrable functions of Y.

Homework

13. Consider the following state space model

$$\text{state equation} : \quad x_t = 0.5x_{t-1} + \frac{25x_{t-1}}{1+x_{t-1}^2} + 8\cos\big(1.2(t-1)\big) + u_t,$$

$$\text{observation equation} : \quad y_t = x_t^2/20 + v_t,$$

where the initial state is $x_0 = 0$, $u_t \sim N(0,1)$ and $v_t \sim N(0,1)$ are independent noises.

(1) Use the particle filter (with resampling) with $m = 1,000$ samples to estimate $E(x_t \,|\, y_{1:t})$ and $E(x_t \,|\, y_{1:t+1})$ for $t = 1, \cdots, 50$. Plot your estimates and the true state path $x_{1:50}$ in one figure.

(2) Repeat the experiment using 100 independent data sets. Report

$$\text{RMSE}(\delta) := \left\{ \frac{1}{100 \cdot 50} \sum_{l=1}^{100} \sum_{t=1}^{50} \Big[\widehat{E}(x_t \,|\, y_{1:t+\delta}) - x_t \Big]^2 \right\}^{1/2}$$

for $\delta = 0, 1, 2, 3, 4$. Try the particle filter with resampling and without resampling.

References

References

Avitzour, D. (1995), "Stochastic simulation Bayesian approach to multitarget tracking," *IEE Proceedings on Radar, Sonar and Navigation*, 142, 41–44.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993), "Novel approach to nonlinear / non-Gaussian Bayesian state estimation," *IEE Proceedings on Radar and Signal Processing*, 140, 107–113.

Kong, A., Liu, J., and Wong, W. (1994), "Sequential imputations and Bayesian missing data problems," *Journal of the American Statistical Association*, 89, 278–288.

Lin, M., Zhang, J., Cheng, Q., and Chen, R. (2005), "Independent particle filters," *Journal of the American Statistical Association*, 100, 1412–1421.

Liu, J. and Chen, R. (1998), "Sequential Monte Carlo methods for dynamic systems," *Journal of the American Statistical Association*, 93, 1032–1044.

Viterbi, A. (1967), "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, 13, 260–269.