ST509 Computational Statistics

Lecture 10: Integration

Seung Jun Shin

Department of Statistics
Korea University

E-mail: `sjshin@korea.ac.kr`

Numerical Integration I

- Integration is crucial in statistics.
- Consider
$$I = \int_a^b f(x)dx$$
- We may not have a closed form of I.
- Requires a numerical integration.

Numerical Integration II

- Mid-point rule computes

$$I \approx \sum_{i=1}^{n} f\left(\frac{x_{i-1} + x_i}{2}\right) \times (x_i - x_{i-1})$$

  where $a = x_0 < x_1 < \cdots < x_n = b$.

- Let $x_i - x_{i-1} = d$ then

$$I \approx \sum_{i=1}^{n} f\left(x_i - \frac{d}{2}\right) \times d$$

Numerical Integration III

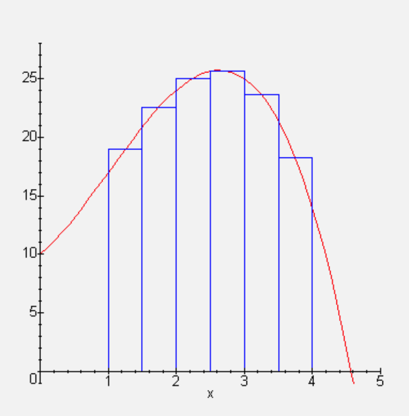


Figure: Illustration of numerical integration (mid-point rule)

Numerical Integration IV

- Trapezoidal rule computes

$$I \approx \sum_{i=1}^{n} \frac{x_{i-1} - x_i}{2} \{f(x_i) + f(x_{i-1})\}$$

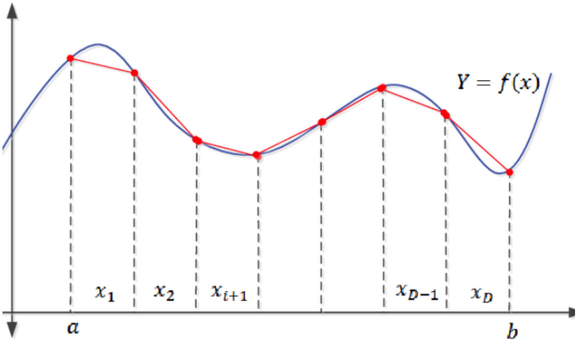
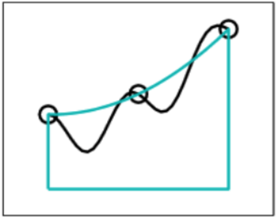


Figure: Illustration of numerical integration (Trapezoidal rule)

Numerical Integration V

- Simpson rule suggests to use higher order polynomials:



- Given

$$(x_0, f(x_0)), (x_1, f(x_1)), (x_2, f(x_2))$$

we have the following expression of quadratic polynomial that possess through the three points:

$$\frac{x-x_1}{x_0-x_1}\frac{x-x_2}{x_0-x_2}f(x_0)+\frac{x-x_0}{x_1-x_0}\frac{x-x_2}{x_1-x_2}f(x_1)+\frac{x-x_0}{x_2-x_0}\frac{x-x_1}{x_2-x_1}f(x_2)$$

Numerical Integration VI

- Let $x_0 = x_{i-1}, x_1 = (x_{i-1} + x_i)/2$ and $x_2 = x_i$, then

$$I \approx \sum_{i=1}^{n} (x_i - x_{i-1}) \frac{f(x_{i-1}) + 4f((x_{i-1} + x_i)/2) + f(x_i)}{6}$$

Numerical Integration VII

- The idea can be generalized for k points $(x_j, f(x_j)), j = 1, \cdots, k$:

$$\sum_{j=1}^{k} \underbrace{f(x_j) \prod_{l=1}^{k} \frac{x - x_l}{x_j - x_l}}_{P_j(x)}$$

- Let $x_0 = x_{i-1}, x_1 = \frac{1}{3}(x_{i-1} + x_i), x_2 = \frac{2}{3}(x_{i-1} + x_i)$ and $x_3 = x_i$, then

$$I \approx \sum_{i=1}^{n} (x_i - x_{i-1}) \frac{f(x_{i-1}) + 3f(1(x_{i-1} + x_i)/3) + 3f(2(x_{i-1} + x_i)/3) + f(x_i)}{8}$$

Numerical Integration VIII

- Gaussian quadrature on $[-1, 1]$ solves

$$\int_{-1}^{1} f(x)dx = \sum_{i=1}^{n} c_i f(x_i)$$

for $f(x) = x^0, x^1, \cdots, x^{2n-1}$ w.r.t $(c_1, \cdots, c_n)$ and $(x_1, \cdots, x_n)$.

Numerical Integration IX

- For $n = 2$:

$$\begin{aligned}
f(x) = 1 &\quad\Rightarrow\quad \int_{-1}^{1} 1dx = 2 = c_1 + c_2 \\
f(x) = x &\quad\Rightarrow\quad \int_{-1}^{1} xdx = 0 = c_1 x_1 + c_2 x_2 \\
f(x) = x^2 &\quad\Rightarrow\quad \int_{-1}^{1} x^2 dx = \frac{2}{3} = c_1 x_1^2 + c_2 x_2^2 \\
f(x) = x^3 &\quad\Rightarrow\quad \int_{-1}^{1} x^3 dx = 0 = c_1 x_1^3 + c_2 x_2^3
\end{aligned}$$

which yields

$$\int_{-1}^{1} f(x)dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

Numerical Integration X

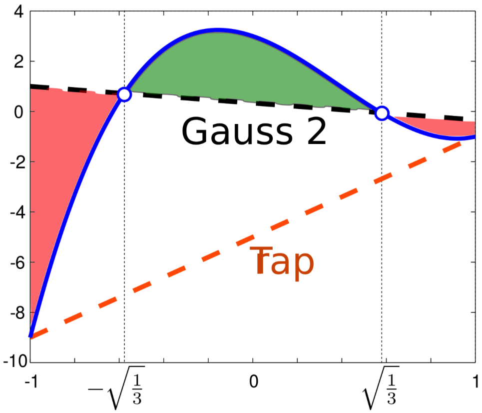


Figure: Comparison: Blue line is the polynomial $f(x) = 7x^3 - 8x^2 - 3x + 3$ whose integral is $2/3$. The trapezoidal rule returns $f(-1) + f(1) = -10$. The 2-point Gaussian quadrature rule returns $f(-\sqrt{1/3}) + f(\sqrt{1/3}) = 2/3$.

Numerical Integration XI

n	$x_i, i = 1, \cdots, n$	$c_i, i = 1, \cdots, n$
2	$-1/\sqrt{3}$	1
	$1/\sqrt{3}$	1
3	$-\sqrt{3/5}$	$5/9$
	0	$8/9$
	$\sqrt{3/5}$	$5/9$
4	$-\sqrt{(15+2\sqrt{30})/35}$	$(18-\sqrt{30})/36$
	$-\sqrt{(15-2\sqrt{30})/35}$	$(18+\sqrt{30})/36$
	$\sqrt{(15-2\sqrt{30})/35}$	$(18+\sqrt{30})/36$
	$\sqrt{(15+2\sqrt{30})/35}$	$(18-\sqrt{30})/36$

Numerical Integration XII

- Transformation from $[a, b]$ to $[-1, 1]$.

$$t = \frac{b-a}{2}x + \frac{b+a}{2}$$

- Then we have

$$\begin{aligned}\int_a^b f(t)dt &= \int_{-1}^{1} f\left(\frac{b-a}{2}x + \frac{a+b}{2}\right)\left(\frac{b-a}{2}\right)dx \\ &= \int_{-1}^{1} g_{a,b}(x)dx \\ &\approx g_{a,b}\left(-\frac{1}{\sqrt{3}}\right) + g_{a,b}\left(\frac{1}{\sqrt{3}}\right)\end{aligned}$$

Monte Carlo Integration I

- In statistics, integration is expectation. For a random variable $\mathbf{X} \sim f$,

$$E\{h(\mathbf{X})\} = \int h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

  where $g(\mathbf{x})$ is an arbitrary function.
- Previous approaches are not applicable unless p is very small (curse of dimensionality)
- However, we have a beautiful alternative, the Law of Large Numbers!
- For $X_1, \cdots, X_n \overset{iid}{\sim} f(\mathbf{x})$

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \quad \to \quad E\{h(\mathbf{X})\}$$

- Key is the generation of random sample!

Monte Carlo Integration II

- For example, we like to compute

$$I = \int_S g(\mathbf{x}) d\mathbf{x}$$

where $S \in \mathbb{R}^p$.

- Suppose we can generate a uniform random variable $\mathbf{X}_1, \cdots, \mathbf{X}_n$ on S. (i.e., $f(\mathbf{x}) = |S|^{-1}$ with $|S|$ being the area of S.)

$$\frac{1}{n} \sum_{i=1}^{n} g(\mathbf{X}_i) \quad \rightarrow \quad \int g(\mathbf{x}) \frac{1}{|S|} d\mathbf{x}.$$

which yields

$$I \approx \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{X}_i) \times |S|$$

Monte Carlo Integration III

- Suppose $X \sim F$ where $F(x) = P(X \leq x)$. Let $Y = F(X)$, then

$$P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y.$$

- Probability Integral Transformation:
  1. Generate $u_1, \cdots, u_n \overset{iid}{\sim} U(0,1)$.
  2. $x_i = F^{-1}(u_i)$ are the random samples from F.

ex. Distribution function of exponential RV is

$$F(x) = 1 - e^{\lambda x}$$

thus

$$x_i = -\log(1 - u_i)/\lambda, \qquad u_i \overset{iid}{\sim} U(0,1),\ i = 1, \cdots, n$$

are the random samples from $\mathrm{Exp}(\lambda)$.

Monte Carlo Integration IV

- Normal Generator:
- (CLT) Suppose $u_1, \cdots, u_{12} \overset{iid}{\sim} U(-1/2, 1/2)$, then

$$\sum_{i=1}^{12} u_i \sim N(0,1)$$

- (Box-Muller) Suppose $u_1, u_2 \overset{iid}{\sim} U(0,1)$, then

$$z_1 = \sqrt{-2\log(u_1)}\cos(2\pi u_2)$$
$$z_2 = \sqrt{-2\log(u_1)}\sin(2\pi u_2)$$

  follows $N(0,1)$.
- Multivariate Normal RV $\mathbf{x} \sim N(\mu, \mathbf{\Sigma})$ directly follows after

$$\mathbf{x} = \mu + \mathbf{\Sigma}^{1/2}\mathbf{z}$$

  with $\mathbf{z} = (z_1, \cdots, z_p)$ and $z_i \overset{iid}{\sim} N(0,1)$.

Monte Carlo Integration V

- We like to generate sample X from f (target density).
- Suppose there exists a simpler density g (proposal density) s.t.
  - f and g have compatible supports, i.e., $g(x) \geq 0$ when $f(x) \geq 0$.
  - There is a constant M with $f(x)/g(x) \leq M$ for all x.

- Rejection Sampling:
  1. Generate $y_i \sim g$ and $u_i \sim U(0,1)$.
  2. Accept $x_i = y_i$ if
  $$u_i \leq \frac{1}{M} \frac{f(y_i)}{g(y_i)}$$
  3. Return to 1 otherwise.

Monte Carlo Integration VI

- Justification:

$$
\begin{aligned}
P\left(Y \le x \,\middle|\, U \le \frac{f(Y)}{Mg(Y)}\right) &= \frac{P\left(Y \le x,\ U \le \frac{f(Y)}{Mg(Y)}\right)}{P\left(U \le \frac{f(Y)}{Mg(Y)}\right)} \\
&= \frac{\int_{-\infty}^{x} \int_0^{f(y)/\{Mg(y)\}} du g(y) dy}{\int_{-\infty}^{\infty} \int_0^{f(y)/\{Mg(y)\}} du g(y) dy} \\
&= \frac{\int_{-\infty}^{x} f(y)/\{Mg(y)\} g(y) dy}{\int_{-\infty}^{\infty} f(y)/\{Mg(y)\} g(y) dy} \\
&= \frac{\int_{-\infty}^{x} f(y) dy}{\int_{-\infty}^{\infty} f(y) dy} \\
&= P(X \le x)
\end{aligned}
$$

Monte Carlo Integration VII

- Goal is to compute $E\{h(X)\}$ where $X \sim f$.
- Suppose direct sampling from f is not possible, but there exists a simpler density g s.t. $g(x)$ is strictly positive when $h(x) \times f(x)$ is.
- Now we have

$$E_f\{h(X)\} = \int h(x)\frac{f(x)}{g(x)}dx = E_g\left[\frac{h(X)f(X)}{g(X)}\right]$$

- Suppose we have random sample $x_1, \cdots, x_n \overset{iid}{\sim} g(x)$, then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{g(x_i)}h(x_i) \quad \rightarrow \quad E_f\{h(X)\}$$

- This is known as Importance Sampling.

Reference

- Robert, & Casella (2010) Introducing Monte Carlo Methods with R; , Springer.