Analysis and Probability

[Integrate Max/Min] $E[\max\{0,Y\}] = E[YI(Y>0)]$ and $E[\min\{0,Y\}] = E[YI(Y<0)]$
[Variance] $Var(X) = E(X^2) + (EX)^2$ and $Var(X|Y) = Var(E[X|Y]) + E[Var(X|Y)]$
[Finding joint and conditional density] Suppose $X = \epsilon_1$, $Y = X + \epsilon_2$, $Z = X + Y + \epsilon_3$, $\epsilon_i \sim^{iid} N(\mu, \sigma^2)$
Note $f_{X|Y,Z}(x|y,z) \propto f_{X,Y,Z}(x,y,z)$
Method ① $f_{X,Y,Z}(x,y,z) = det(\nabla_J) f_{\epsilon_1,\epsilon_2,\epsilon_3}(x, y-x, z-x-y) \propto f_{\epsilon_1}(x) f_{\epsilon_2}(y-x) f_{\epsilon_3}(z-x-y)$
Method ② $f_{X,Y,Z}(x,y,z) = f_{Z|X,Y}(z|x,y) f_Y(y|x) f_X(x)$ and $Z|X,Y \sim N(x+y+\mu, \sigma^2)$, $Y|X \sim N(x+\mu, \sigma^2)$
[Conditional Density Example] $X_i \sim^{iid} N(\mu, \sigma^2)$ and $Y_i = X_i + X_{i+2}$.
$f_{X_1|Y}(x|Y) \propto f_X(x, y_1 - x, y_2 - y_1 + x, y_3 - y_2 + y_1 - x, y_4 - y_3 + y_2 - y_1 + x) \sim N\left(\frac{1}{5}\left[4Y_1 - 3Y_2 + 2Y_3 - Y_4 + \mu\right], \frac{\sigma^2}{5}\right)$
$f_{X_2|Y}(x|Y) \sim N\left(\frac{1}{5}\left[Y_1 + 3Y_2 - 2Y_3 + Y_4 - \mu\right], \frac{\sigma^2}{5}\right)$ $f_{X_3|Y}(x|Y) \sim N\left(\frac{1}{5}\left[-Y_1 + 2Y_2 + 2Y_3 - Y_4 + \mu\right], \frac{\sigma^2}{5}\right)$
$f_{X_4|Y}(x|Y) \sim N\left(\frac{1}{5}\left[Y_1 - 2Y_2 + 3Y_3 + Y_4 - \mu\right], \frac{\sigma^2}{5}\right)$ $f_{X_5|Y}(x|Y) \sim N\left(\frac{1}{5}\left[-Y_1 + 2Y_2 - 3Y_3 + 4Y_4 + \mu\right], \frac{\sigma^2}{5}\right)$
[RV transformation] $Y = h(X)$ $g_Y(y) = f_X(h^{-1}(Y)) \left|det\left(\frac{dh^{-1}}{dY}\right)\right|$
[KL Divergence] $D_{KL}(g|f) = E_g\left(\log\frac{g(x)}{f(x)}\right) \geq 0$
[Tail of $Exp(0,a)$] $E(Y_j - c_j | Y_j > c_j) = E(Y_j)$
[Series summation] $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ and $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$ and $\sum_{i=1}^n i^3 = \left(\frac{n(n+1)}{2}\right)^2$

L1 Review

[Big $O(\cdot)$] $f(z) = O(g(z))$ as $z \to z_0 \in \mathcal{R}$ if for some $M > 0$, and for all z in neighborhood of z_0.

$$\left|\frac{f(z)}{g(z)}\right| \leq M$$

If $z \to \infty$, then there exists $C > 0$ s.t. statement holds for all $z > C$
E.g. $f(n) = h(n) + \frac{n+1}{3n^2}$, since $\lim_{n\to\infty}\left\{\frac{n+1}{3n^2}/n^{-1}\right\} = 1/3 < \infty$, $\Rightarrow f(n) = h(n) + O(n^{-1})$
[Small $o(\cdot)$] $f(z) = o(g(z))$ as $z \to z_0 \in \mathcal{R}$ if

$$\lim_{z\to z_0}\frac{f(z)}{g(z)} = 0$$

E.g. since $\lim_{n\to\infty}\frac{n+1}{3n^2} = 0$ $f(n) = h(n) - o(1)$ as $n \to \infty$
[Taylor's Expansion] Let $f(\cdot)$ defined on $[a,b]$ s.t. it has continuous $(n+1)$th order derivatives. Then for all x, x_0 in $[a,b]$

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \cdots + \frac{(x-x_0)^n}{n!}f^{(n)}(x_0) + R_n$$

where

$$R_n = \frac{(x-x_0)^{n+1}}{(n+1)!}f^{(n+1)}(\xi) = O(|x-x_0|^{n+1})$$

for some $\xi \in (x, x_0)$ or (x_0, x)
[Alternate Taylor]
Since $f^{(n+1)}(\cdot)$ is bounded based on theorem condition

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \cdots + \frac{(x-x_0)^n}{n!}f^{(n)}(x_0) + O(|x-x_0|^{n+1})$$

as $x \to x_0$
[Multivariate Taylor expansion]
Let $x = (x_1, x_2)^T, y = (y_1, y_2)^T$

$$f(x+y) = f(x) + y_1 f_1(x) + y_2 f_2(x) + R$$

$$R = \frac{1}{2}y_1^2 f_{11}(\xi) + y_1 y_2 f_{12}(\xi) + \frac{1}{2}y_2^2 f_{22}(\xi) = O(||y||^2)$$

and $\xi = \alpha x + (1-\alpha)(x+y)$ for some $\alpha \in [0,1]$
[Likelihood Inference]
$X_1, \cdots, X_n$ be iid with $f(x|\theta)$, then likelihood of $X_1 = x_1, \cdots, X_n = x_n$ is

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta)$$

Likelihood principle find θ that maximises $L(\theta)$. Log-likelihood $= \ell(\theta) = \log L(\theta)$. Score function $s(\theta) = \ell'(\theta)$
[Asymptotic Normality of MLEs]
[Convergence Order]
A root-finding method has convergence order β (≥ 1) if
(a) $\lim_{t\to\infty}\epsilon_t = 0$
(b) $\lim_{t\to\infty}\frac{|\epsilon_{t+1}|^{\beta}}{\epsilon_t} = c$ for some $c > 0$
When $\beta = 1$, we require $c < 1$
[Matrix Digression]
Given y, z not orthogonal to each other, find symmetric matrix M s.t. $y = Mz$

[[Solution 1]] $y^T z$ is scalar, $M = \frac{yy^T}{y^T z}$
[[Solution 2]] Given any symmetric matrix M_0, let $v = y - M_0 z$. $M = M_0 + \frac{vv^T}{v^T z}$
[[Solution 3]] $M = M_0 - \frac{(M_0 z)(M_0 z)^T}{z^T M_0 z} + \frac{yy^T}{y^T z}$

Optimisation

[Optimisation in Uni-variate: find x^* s.t. $g'(x^*) = 0$]

[Bisection]

Condition: $g'(a) > 0, g'(b) < 0$, $g'(x)$ exist and continuous for all $x \in (a, b)$
Let $x_0 = (a+b)/2$, set $\tilde{a} = a$, $\tilde{b} = b$, $t = 0$
(1.1) If $g'(x_{t-1}) > 0$, $X_t = (x_{t-1} + \tilde{b})/2$, $\tilde{a} = x_{t-1}$
(1.2) If $g'(x_{t-1}) < 0$, $X_t = (\tilde{a} + x_{t-1})/2$, $\tilde{b} = x_{t-1}$
(2) $t = t+1$, terminate when $|x_t - x_{t-1}| < \epsilon$

[Modified Bisection]

Instead of choosing the mid-point, we can choose

$$x_t = \frac{|g'(b)|}{|g'(a)| + |g'(b)|} a + \frac{|g'(a)|}{|g'(a)| + |g'(b)|} b$$

[Newton's Method]

$$x_{t+1} = x_t - \frac{g'(x_t)}{g''(x_t)}$$

[Fisher Scoring]

Replace Hessian $\ell''(\theta_t)$ in Newton method by $-I(\theta_t)$

$$-I(\theta) = nE\left\{\frac{d^2}{d\theta^2} \log f(X|\theta)\right\} = \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f(x_i|\theta)$$

$$\theta_{t+1} = \theta_t + \frac{\ell'(\theta_t)}{I(\theta_t)}$$

[Secant Method]

Approximate Hessian $g''(x) = \lim_{y \to x} \frac{g'(y) - g'(x)}{y - x}$, assuming update is small, i.e. $|x_{t-1} - x_t| < \epsilon$

$$g''(x_t) \approx \frac{g'(x_{t-1} - g'(x_t)}{x_{t-1} - x_t}$$

$$x_{t+1} = x_t - g'(x_t) \frac{x_t - x_{t-1}}{g'(x_t) - g'(x_{t-1})}$$

[Fixed-point Iteration]

Let $g'(a) > 0, g'(b) < 0$. Assume $\exists x^* \in [a, b], \epsilon \in (0, \frac{1}{2})$ s.t.
$(1-\epsilon)(x^* - x) \geq g'(x) \geq \epsilon(x^* - x)$ for $x < x^*$
$(1-\epsilon)(x^* - x) \leq g'(x) \leq \epsilon(x^* - x)$ for $x > x^*$
Then $x_{t+1} = x_t + g'(x_t)$ converges to x^*

[Optimisation in Multivariate]

[Newton's Method, Fisher scoring]

Similar to single variable method, with $g' = \nabla g$, $g'' = \nabla^2 g$

[Newton-like method]

General form with $-M_t$ a positive definite matrix

$$x_{t+1} = x_t - \alpha_t [M_t]^{-1} g'(x_t)$$

[Ascent Algorithm: Bracketing]

Ascent algo: Control for α_t s.t. $g(x_{t+1}) \geq g(x_t)$

$$x_{t+1} = x_t + \alpha_t g'(x_t)$$

Bracketing:
(1) start with $\alpha_t = 1$, compute x_{t+1}
(2) if $g(x_{t+1}) < g(x_t)$, α_t is too large and update $\alpha_t = 1/2$

[Discrete Newton]

Approximate Hessian g'' by discrete version, with $e_1 = (1,0)^T, e_2 = (0,1)^T$, some small $h_{ij} > 0$

$$M_{ij}^{(t)} = \frac{g_i(x_t + h_{ij} e_j) - g_i(x_t)}{h_{ij}}$$

To ensure symmetry, consider

$$N_{ij}^{(t)} = \frac{M_{ij}^{(t)} + M_{ji}^{(t)}}{2}$$

[Quasi-Newton]

Estimate Hessian with $g'(x_t) - g'(x_{t-1}) = M_t(x_t - x_{t-1})$.

Consider $y = g'(x_t) - g'(x_{t-1})$, $z = x_t - x_{t-1}$, $M_t = M_{t-1} + \frac{v^T}{v^T z}$
If $1/(v^T z) \leq 0$, $-M_0 \succ 0 \Rightarrow -M \succ 0$
If $1/(v^T z) > 0$, $M_t = M_{t-1} + \alpha_t v v^T$ with $\alpha_t > 0$ s.t. $-M \succ 0$
[Gaussian-Newton]
Model $y_i = f(z_i, \theta) + \epsilon_i$, $\epsilon_i \sim N(0, \tau)$ iid, then $\theta = (Z^T Z)^{-1} Z^T y$ (linear) else $\theta_{t+1} = \theta_t + [A_t^T A_t]^{-1} A_t^T x_t$
[Nonlinear Gauss-Seidel]
Restrict update to one co-ordinate at a time, find x_1^*, x_2^* s.t. $g_1(x_1^*, x_2^*) = 0$, $g_2(x_1^*, x_2^*) = 0$
Iterate with $g_1(x_1^{(t+1)}, x_2^{(t)}) = 0$ $g_2(x_1^{(t+1)}, x_2^{(t+1)}) = 0$

L2: EM Optimization

[EM]
Want to solve $\hat{\theta} = arg\max \ell_X(\theta)$ with some missing data Z.
Therefore, consider $Y = (X, Z)$ complete data instead. $\ell_Y(\theta) = \ell_X(\theta) + \ell_{Z|X}(\theta)$.
Solve for

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}\left[\ell_Y(\theta)|X\right]$$

with (1) E-step: Compute $Q(\theta|\theta^{(t)})$ (2) M-step: Maximise Q with respect to θ and set $\theta^{(t+1)} = \theta^*$
Only requires: $\ell_X(\theta^{(t+1)}) > \ell_X(\theta^{(t)})$ (generalised EM)
[EM for Canonical Exp Fam]
Canonical Exp Fam has log-likelihood linear in missing data Z and observed data X. Check before solving (1) impute Z (2) estimate $\theta^{(t+1)}$

$$\ell_Y(\theta) = c(Y) + d(\theta) + \sum_{j=1} p\theta_j Y_j$$

$$Q(\theta|\theta^{(t)}) = c(Y) + d(\theta) + \sum_{j=1}^{p} \theta_j E_{\theta^{(t)}}(Y_j|X)$$

[Var estimate of $\hat{\theta}$] Note that variance estimate $\hat{\theta}$ is wrt to i_X
Fisher information for NEF $I(\theta) = E_\theta[-\ell''_X(\theta)] = var_\theta(\ell'_X(\theta))$
MLE asymptotic dist $I(\theta)^{-1/2}(\hat{\theta} - \theta_0) \to N(0, I_K)$
Fisher info for complete data $i_Y(\theta) = i_X(\theta) + i_{Z|X}(\theta) \Rightarrow i_X = i_Y - i_{Z|X}$ (need to compute both i_Y and $i_Z|X$ to get i_X)
BS-MC estimate $\hat{i}_Y(\theta) = -\frac{1}{m}\sum_{i=1}^m \ell''_{Y^{(k)}}(\theta)$, $\hat{i}_{Z|X}(\theta) = -\frac{1}{m}\sum_{i=1}^m \ell''_{Z^{(k)}}(\theta)$

Extended EM

[MC-EM]
Instead of calculating $Q(\theta|\theta^{(t)})$ via integration, use MC instead.
[Expected Conditional Max]
Instead of maximising $\theta = (a, b)$ at once, maximise them sequentially
(a) $\max_a Q(a, b^{(t)}|\theta^{(t)})$ (b) $\max_b Q(a^{(t+1)}, b|\theta^{(t)})$ (c) $\theta^{(t+1)} = (a^{(t+1)}, b^{(t+1)})$
[EM Gradient]
Instead of solving maximisation analytically, use gradient-based methods (e.g. Newton). $\theta^{(t+1)} = \theta^{(t)} - Q''(\theta|\theta^t)^{-1}|_{\theta=\theta^t} \times Q'(\theta|\theta^t)|_{\theta=\theta^{(t)}}$

EM Acceleration Methods

[Convergence rate]
EM est $\hat{\theta}$ converge to θ at linear rate, depending on fraction of observed information $\rho(\theta) = \frac{i_X(\theta)}{i_Y(\theta)}$
[Aitken Acceleration]
Use Newton method for optim (Quad rate) and estimate $\ell_X(\theta)$ using EM with $\rho(\theta) = \frac{i_X(\theta)}{i_Y(\theta)} = 1 - \frac{i_{Z|X}(\theta)}{i_Y(\theta)}$

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\theta_{EM}^{(t)} - \theta^{(t)}}{\rho(\theta^{(t)})}$$

[Quasi-Newton Acceleration]
Avoid estimating $\rho(\theta)$, $\rho(\theta) \approx 1 - \frac{\theta_{EM}^{(t)} - \theta_{EM}^{(t-1)}}{\theta^{(t)} - \theta^{(t-1)}}$

$$\theta^{(t+1)} = \theta^{(t)} + (I - M^{(t)})^{-1}(\theta_{EM}^{(t)} - \theta^{(t)})$$

L3: Numerical Integration Efficient method for lower dimension.

[Integration] Objective: approximate $\int_a^b f(x)dx$ numerically
Naive method: Divide $[a, b]$ into n sub-intervals, x_i^* is the middle point of ith subinterval.

$$\int_a^b f(x)dx \approx \frac{b-a}{n}\sum_{i=1}^{n} f(x_i^*)$$

Improvement: for each of the sub-interval $[x_i, x_{i+1}]$ add $(m+1)$ nodes
[Newton-Cotes Quadrature] General class that approximate $I = \frac{\int_{x_i}^{x_{i+1}} f(x)dx}{x_{i+1} - x_i}$ with $\hat{I}_m = \sum_{j=0}^m c_j f(x_j^*)$
and $x_i = x_0^* < x_2^* < \cdots < x_m^* = x_{i+1}$ equally spaced in $[x_i, x_{i+1}]$
[Trapezoidal Rule] Choose 2 nodes $(m = 1)$ in $[x_i, x_{i+1}]$. To approximate height $I = \frac{\int_{x_i}^{x_{i+1}} f(x)dx}{x_{i+1} - x_i}$. Area $= (x_{i+1} - x_i) \times I$

$$\hat{I}_1 = \frac{f(x_0^*) + f(x_1^*)}{2}$$

Total area $\int_a^b f(x)dx$, with $h = (b - a)/n$

$$\hat{T}(n) = h\sum_{i=1}^{n}\frac{f(x_i)+f(x_{i+1})}{2}$$

$\hat{T}(n) - \int_a^b f(x)dx = O(n^{-2})$
[Simpson Rule] Choose 3 nodes ($m=2$). Approximate height I

$$\hat{I}_2 = \frac{1}{6}f(X_0^*) + \frac{4}{6}f(x_1^*) + \frac{1}{6}f(x_2^*)$$

Total area $\int_a^b f(x)dx$, with $h=(b-a)/n$, $x_i^* = (x_i + x_{i+1})/2$

$$\hat{S}(n) = h\sum_{i=1}^{n}\left\{\frac{f(x_i)}{6} + \frac{4f(x_i^*)}{6} + \frac{f(x_{i+1})}{6}\right\}$$

$\hat{S}(n) - \int_a^b f(x)dx = O(n^{-4})$, can generalised to other polynomial order m
To prove the coefficients are as such, show either linear system solution of $I = \int_0^1 f(x)dx = a_0 + \frac{1}{2}a_1 + \frac{1}{3}a_2$ and $\hat{I}_2 = c_0 f(0) + c_1 f(0.5) + c_2 f(1) = (c_0+c_1+c_2)a_0 + (0.5c_1 + c_2)a_1 + (0.25c_1 + c_2)a_2$ assume $I = \hat{I}_2$
[Gaussian Quadrature] Remove Newton-Cotes restriction of equally spaced nodes and $x_0^* = x_i$, $x_m^* = x_{i+1}$, perfect est for polynomial order $2m+1$ and below (or fn close enough) using $2m+2$ points (x_m, x_0, c_m, c_0). Focus on a segment $[a,b]$.

$$I = \int_a^b w(x)f(x)dx \approx \sum_{j=0}^{m} c_j f(x_j)$$

when a,b finite, $w(x)=1$; when $a=0, b=\infty$, $w(x)=e^{-x}$; when $a=-\infty, b=\infty$, $w(x)=e^{-x^2/2}$
[Gaussian Quadrature: Construct $p_m(x)$ and x_m]
1. construct polynomial of degreee $m+1$ denoted by $p_m(x)$ s.t.

$$\int_a^b w(x)x^k p_m(x)dx = 0, k=0,\cdots,m$$

2. construct $x_0,\cdots,x_m$ as roots to $p_m=0$
3. construct $c_0,\cdots,c_m$ as solutions to $\int r(x)dx = \sum_{j=1}^m c_j r(x_j)$ where $r(x) = a_0 + a_1 x + \cdots + a_m x^m$

$$\int_a^b w(x)r(x)dx = \sum_{j=0}^{m} W_j a_j,\ \ W_j = \int_a^b w(x)x^j dx$$

and

$$\sum_{j=0}^{m} c_j r(x_j) = \sum_{j=0}^{m} UJa_j,\ \ U_j = \sum_{i=0}^{m} c_i x_i^j$$

matching coefficients of a_j or $U_j = W_j$, and solve linear system of $m+1$ equations with $m+1$ unknowns: $c_0,\cdots,c_m$.
4. Estimate

$$\int_a^b w(x)f(x)dx \approx \sum_{j=1}^{m} c_j f(x_j)$$

[Forming $p_m(x)$]
1. Form $p_0 := x + a_0$ s.t. $\int w(x)p_0 dx = 0$
2. $p_1 := x^2 + b_1 x + b_0$ s.t. $\int w(x)p_1 dx = \int w(x)xp_1 dx = 0$
3. $p_m = xp_{m-1} + a_m p_{m-1} + b_m p_{m-2}$ s.t. $\int w(x)x^m p_m dx = \int w(x)x^{m-1}p_m dx = 0$

L4: Bootstrap

[Nonparametric] Re-sample with replacement and estimate $E(f(X))$ with $\frac{1}{B}\sum_{b=1}^{B} f(X^{(b)})$
[Parametric] First estimate $\hat{\theta}$ (e.g. with MLE) then generate samples from $F_{\hat{\theta}}(x)$. require assumption on parametric form.
[BS techniques] Paired BS: generate BS samples by pairing $Z_i = (x_i, y_i)$
BS residual: generate est y_i^* by bootstrapping $\hat{\epsilon}_i^*$
Bias correction: $\hat{\text{bias}} = \frac{1}{B}\sum_{k=1}^{B}(\hat{\theta}_k^* - \hat{\theta})$, correct estimate with $\hat{\theta} - \hat{\text{bias}}$
[BS Percentile CI] 90% BS CI for $\theta = \left(\hat{\theta}^*_{(5)}, \hat{\theta}^*_{(95)}\right)$
Only works well if $\hat{\theta} - \theta$ does not depend on θ and is symetric about 0
[BS t CI] Consider $\frac{\hat{\theta}-\theta}{\hat{\theta}}$ instead, let $d_k^* = \frac{\hat{\theta}_k^* - \hat{\theta}}{\hat{\sigma}_k^*}$, 90% CI for θ is $(\hat{\theta} - \hat{\sigma}d^*_{(95)}, \hat{\theta} - \hat{\sigma}d^*_{(5)})$
[Balanced BS] Reduce MC error from some observed X_i are too frequently selected by chance.
(1) Generate every X_i exactly B times. (2) Permute/re-order the samples (3) first n is assigned to first BS sample
[Antithetic BS] Reduce MC error by enforcing data pairing. (1) Generate B data (2) second sample is replacing $X_{(k)}$ with $X_{(n-k+1)}$
[BS as SIS] Proposal density $X^* \sim f(x)$, same as target density $f(x)$ $w(X^*) = \frac{f(x)/f(x)}{\sum f(x)/f(x)} = \frac{1}{n}$
[BS is unbiased estimate] Note $\sum_{j=1}^{b} I(X_i^* = X_j) = 1$

$$E[h(X_i^*)] = \sum_{i=1}^{b}[h(X_i^*)I(X_i^* = X_i)] = \sum_{i=1}^{b} E[E[h(X_i^*)I(X_i^* = X_i)|I(X_i^* = X_i)]]$$

$$= \sum_{i=1}^{b} E[h(X_i^*)I(X_i^* = X_i) = 1]P(X_i^* = X_i) = \sum_{i=1}^{b} E[h(X_i)]\frac{1}{n} = E[h(X_i)]$$

L5: Simulation and MC Integration

[MC integration] Estimate $\mu = E[h(X)]$, generate X_i from $f(x)$ (known)
$\hat{\mu}_{MC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ and $\hat{\sigma}^2_{MC} = \frac{1}{n-1}\sum_{i=1}^{n}[h(X_i) - \hat{\mu}_{MC}]^2$ and MC estimate: $\hat{\mu}_{MC} \pm \hat{\sigma}/\sqrt{n}$
[Extract Simulation] Simulate samples from $f(x)$ directly if $F^{-1}(U)$ exist and known, and is single-varaite
(1) Generate $U \sim Unif(0,1)$ (2) $X = F^{-1}(U)$
Known distributions such as Gaussian, Beta have special algorithm.
[Rejection Sampling] Assume $f(x)$ can be computed easily, find proposal density $Y \sim g$ s.t. $f(x) \le g(x)/\alpha$ for known $\alpha > 0$ If $\alpha f(Y)/g(y)$ is small, then algo is inefficient. To ensure rejection sampling exists, require $\frac{f(x)}{g(x)} \le \frac{1}{\alpha}$, bounded by a constant.
(1) Generate $Y \sim g$
(2) Generate $U \sim unif(0,1)$
(3) If $U \le \alpha f(Y)/g(Y)$, set $X = Y$
(4) Else, repeat (1-3) until succeed
[Deducing Rejection Sampling distribution] $P(X \le x) = P(Y \le y | U \le \alpha f(Y)/g(Y))$
[Rejection Sampling for multivariate] Consider $\mathcal{O} = \{(x,y,z) : x^2+y^2+z^2 \le 1\}$, $\mathcal{D} = \{(x,y) : x^2+y^2 \le 1\}$
area of $\mathcal{D} = \pi$, area of $\mathcal{O} = \frac{4\pi}{3}$
1 generate $\mathcal{D}$ using $X \sim unif(-1,1)$, $Y \sim unif(-1,1)$
(1) Generate $W \sim unif(-1,1)$, $V \sim unif(-1,1)$
(2) If $W^2+V^2 \le 1$ or $(W,V) \in \mathcal{D}$, set $(\tilde{X},\tilde{Y}) = (W,V)$ else repeat (1)
This is rejection sampling with $g(w,v) = I(w \in (-1,1))I(v \in (-1,1))$, $f(x,y) = \frac{1}{\pi}I(x^2+y^2 \le 1)$, $f(x,y)/g(x,y) \le \frac{1}{\pi} \Rightarrow \alpha = \pi$
Since $\alpha f(w,v)/g(w,v) = I(w^2+v^2 \le 1)$, $U \le \alpha f(w,v)/g(w,v) \Rightarrow I(w^2+v^2 \le 1)$ or $(W,V) \in \mathcal{D}$
1 gnerate $\mathcal{O}$ using $Z \sim unif(-1,1)$ and X,Y
(3) Generate $S \sim unif(-1,1)$
(4) If $\tilde{X}^2+\tilde{Y}^2+S^2 \le 1$ or $(\tilde{X},\tilde{Y},S) \in \mathcal{O}$, set $\tilde{Z} = S$ else repeat (3)
Similarly, $g(w,v,s) = I(w \in (-1,1))I(v \in (-1,1))I(s \in (-1,1))$, $f(x,y,z) = \frac{3}{4\pi}I(x^2+y^2+z^2 \le 1)$, $f(x,y,z)/g(x,y,z) \le \frac{3}{4\pi} \Rightarrow \alpha = 4\pi/3$
Since $\alpha f(w,v,s)/g(w,v,s) = I(w^2+v^2+s^2 \le 1)$, $U \le \alpha f(w,v,s)/g(w,v,s) \Rightarrow I(w^2+v^2+s^2 \le 1)$ or $(W,V,S) \in \mathcal{O}$
[SIR]
Sampling Importance Resampling, with envelope function $g(x)$. Note $E[h(X)] = \sum_{i=1}^{n} w_i h(Y_i)$
Generate approximate distribution from $f(x)$ (previous 2 methods are exact).
(1) Sample $Y_i, \cdots, Y_m$ from $g(x)$
(2) Calculate standardised importance weight $w(Y_1), \cdots, w(Y_m)$
$w^*(Y_i) = f(Y_i)/g(Y_i)$ and $w(Y_i) = \frac{w^*(Y_i)}{\sum_{j=1}^{m} W^*(Y_j)}$
(3) Resample X_i from $Y_1, \cdots, Y_m$ with probability $w(Y_1), \cdots, w(Y_m)$
[Finding SIR asymptotic distribution]

$$P(X_i \in A | Y_1, \cdots, Y_m) = P(\cup_{j=1}^{m}\{X_i = Y_j \text{ and } Y_j \in A\} | Y_1, \cdots, Y_m) = \frac{\sum_{j=1}^{m} I(Y_j \in A)w^*(Y_j)}{\sum_{j=1}^{m} w^*(Y_j)} = \frac{\frac{1}{m}\sum_{j=1}^{m} I(Y_j \in A)w^*(Y_j)}{\frac{1}{m}\sum_{j=1}^{m} w^*(Y_j)}$$

Using LLN with $m \to \infty$

$$\frac{1}{m}\sum_{j=1}^{m} I(Y_j \in A)w^*(Y_j) \to E[I(Y_j \in A)w^*(Y)] = \int_A \frac{f(y)}{g(y)}g(y)dy = \int_A f(y)dy$$

$$\frac{1}{m}\sum_{j=1}^{m} w^*(Y_j) \to E[w^*(Y)] = \int \frac{f(y)}{g(y)}g(y)dy = \int f(y)dy = 1$$

By DCT, $P(X_i \in A) = E[P(X_i \in A | Y_1, \cdots, Y_m)] = \int_A f(y)dy$
[Sequential MC]
Splitting high-dimensional task into sequence of simpler steps, each step updates the previous one. Goal: simulate $X_{1:t}^{(i)}$, $i = 1, \cdots, n$ iid from $f(x_{1:t})$
(1) Sample $X_1 \sim g(x_1)$. Let $w_1 = u_1 = f(x_1)/g(x_1)$. set $t = 2$, $X_{1:t-1} = X_1$
(2) Sample $X_t = g(x_t | X_{1:t-2})$
(3) Append X_t to $X_{1:t-1}$. Obtain $X_{1:t}$
(4) Let $u_t = f(X_t|X_{1:t-1})/g(X_t|X_{1:t-1})$
(5) Let $w_t = w_{t-1}u_t$
(6) Increase t by 1 and return to step (2)
[SISR]
When t increases $w_t^{(i)}$ may have large variability and reduce sampling efficiency.
Effective sample size $\hat{N}_t = \frac{n}{1+cv_t^2}$, $cv_t^2 = \sum_{i=1}^{n}(w_t^{(i)} - \bar{w}_t)^2/(n\bar{w}_t^2)$, $\bar{w}_t = \sum_{i=1}^{n} w_t^{(i)}/n$
(1) When $\hat{N}_t$ is smaller than predetermined threshold, stop SIS
(2) Resample n sequences from $\{X_{1:t}^{(1)}, \cdots, X_{1:t}^{(n)}\}$ with probability $\{w_t^{(1)}, \cdots, w_t^{(n)}\}$, set weight for new resampled seq as $1/n$
(3) Use resample sequences and weights as inputs for next step in SIS algo

Variance Reduction

[Importance Sampling]
$\mu = E[h(X)] = \int h(x)w(x)g(x)dx$, $w(x) = \frac{f(x)}{g(x)}$

$$\hat{\mu}_{IS} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)w(X_i)$$

[Antithetic Sampling]
Find two unbiased estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ that are negatively correlated

$$\hat{\mu}_{AS} = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$$

[Control Variates]
Generate 2 sets of samples $\{(X_i, Y_i)\}$, $\mu = E[h(X)]$, $\theta = E(c(Y))$

$$\hat{\mu}_{CV} = \hat{\mu}_{MC} + \lambda(\hat{\theta}_{MC} - \theta)$$

with $\lambda_{\min} = -\frac{cov(h(X), c(Y)}{var(c(Y)}$
[Rao-Blackwellization]
Remove randomness from some vectors by solving conditional expectation.
Consider $X = (X_1, X_2)$, $\mu = E(h(X)) = E[E(h(X)|X_2)] = E(\tilde{h}(X_2))$

$$\hat{\mu}_{RB} = \frac{1}{n}\sum_{i=1}^{n} \tilde{h}(X_{i2})$$

L6: Markov Chain Monte Carlo
[MCMC] Generate stationary distribution s.t. $X_t \sim f(x) \Rightarrow X_{t+1} \sim f(x)$ using exchangeable transition kernel $R(X_t, Y)$.
Require $P(X_t \le x, X_{t+1} \le x') = P(X_t \le x', X_{t+1} \le x) \Leftrightarrow F(x, x') = F(x', x)$

$$F(x, x') = P\left(X_t \le x, Y \le x', U \le R(X_t, X_{t+1})\right) + P\left(X_t \le x, X_t \le x', U > R(X_t, X_{t+1})\right) = F_1(x, x') + F_2(x, x')$$

$F_2(x, x')$ is exchangeable as both is about X_t Note $f(x_t, y) = f_{X_t}(x_t)g_Y(y|x_t)$

$$F_1(x, x') = \int_{x_t \le x, y \le x'} \min\{f(x_t)g(y|x_t), R(x_t, y)f(x_t)g(y|x_t)\}dx_t dy = \int_{z \le x, w \le x'} \min\{f(z)g(w|z), R(z, w)f(z)g(w|z)\}dzdw$$
$$F_1(x', x) = \int_{x_t \le x', y \le x} \min\{f(x_t)g(y|x_t), R(x_t, y)f(x_t)g(y|x_t)\}dx_t dy = \int_{z \le x, w \le x'} \min\{f(w)g(z|w), R(w, z)f(w)g(z|w)\}dzdw$$

as X_t, Y are dummy variables. Require

$$\min\left\{f(x_t)g(y|x_t), R(x_t, y)f(x_t)g(y|x_t)\right\} = \min\left\{f(y)g(x_t|y), R(y, x_t)f(y)g(x_t|y)\right\}$$

[Deducing MCMC distribution] $P(X_t \le x) = P(Y_t \le x, U \le R(X_t, Y)) + P(X_t \le x, U > R(X_t, Y))$
$= E\left[I(Y_t \le x)\min\{1, R(X_t, Y)\}\right] + E\left[I(X_t \le x)[1 - \min\{1, R(X_t, Y)\}]\right]$
[Independence Chains] Proposal distribution $g(x)$, $w(x) = f(x)/g(x)$
(1) Generate $X_1 \sim g(x)$, let $t = 1$
(2) Generate $Y \sim g(x)$, $U \sim Unif(0, 1)$
(2.1) If $U \le w(Y)/w(X_t)$, $X_{t+1} = Y$
(2.2) If $U > w(Y)/w(X_t)$, $X_{t+1} = X_t$
(3) Increase t by 1
(4) Repeat steps (2) and (3) to generate $X_1, X_2, \cdots$
Basically,

$$R(X_t, Y) = \frac{f_{X_t}(Y)g_Y(X_t)}{f_{X_t}(X_t)g_Y(Y)}$$

[Metropolis-Hasting]
(1) Generate X_1 from arbitary initial distribution and set $t = 1$
(2) Simulate $Y \sim g(y|X_t)$
(3) Compute MH ratio $R(X_t, Y)$

$$R(X_t, Y) = \frac{f_{X_t}(Y)g_Y(X_t|Y)}{f_{X_t}(X_t)g_Y(Y|X_t)}$$

(4) Generate $U \sim Unif(0, 1)$,
(4.1) If $U \le R(X_t, Y)$, $X_{t+1} = Y$
(4.2) Otherwise, $X_{t+1} = X_t$
(5) Increase t by 1
(6) Repeat steps (2)-(5) t generate MC chain $X_1, X_2, \cdots$
[Metropolis] Initial algorithm proposed by Metropolis require symmetric transition kernel $g(x_t|y) = g(y|x_t)$
[Gibbs Sampling]
(1) Simulate $X_1 = (X_{11}, X_{12})$ from arbitary distribution, set $t = 1$
(2) Simulate $X_{t+1|1} \sim f_1(x_1|X_{t,2})$ and then simulate $X_{t+1,2} \sim f_2(x_2|X_{t+1}, 1)$
(3) Increase t by 1 and repeat (2)
[Gibbs Sampling tricks]
When given mixture density, define latent variable $Z_{ij} \in \{0, 1\}$, and $Z_i = \sum_{j=1}^{k} Z_{ij} = 1$

$$f(X)=\sum_{j=1}^{k}p_jf(x|\theta_j)=\sum_{j=1}^{k}p_jf(x|Z_{ij}=1,\theta_j)$$

$$f(X,Z_i)=\prod_{j=1}^{k}p_jf(x|\theta_j)^{Z_j},\quad f(X|Z_i)=p_jf(x|\theta_j),\quad f(Z_i|X)=\frac{f(X,Z_i)}{f(X)}=\frac{\prod_{j=1}^{k}p_kf(x|\theta_k)^{Z_{ij}}}{\sum_{j=1}^{k}p_jf(x|\theta_j)}$$

L7: Non-parametric Density Estimation

[Measure of Performance]
ISE: Integrated squared error

$$ISE(\hat{f}(x))=\int\left\{\hat{f}(x)-f(x)\right\}^2dx$$

MSE: mean squared error

$$MSE(\hat{f}(x))=E\left[\left\{\hat{f}(x)-f(x)\right\}^2\right]=\text{bias}^2\{\hat{f}(x)\}+\text{var}\{\hat{f}(x)\}$$

MISE: mean integrated squared error

$$MISE(\hat{f}(x))=E\left\{ISE(\hat{f}(x))\right\}=\int MSE(\hat{f}(x))dx=\int\text{bias}^2\{\hat{f}(x)\}+\int\text{var}\{\hat{f}(x)\}$$

[Naive Estimators] $X\sim f(x)$, $x\in[a,b]$

$$\hat{f}_n(x)=\frac{\hat{F}_2(x+h)-\hat{F}_n(x-h)}{2h}=\frac{1}{2nh}(\#\text{ of }X_1,\cdots,X_n\text{ in }(x-h,x+h])$$

Equivalently,

$$w(x)=I\left(|x|<1\right)\frac{1}{2}$$

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}w\left(\frac{x-X_i}{h}\right)$$

[Histogram moments]
$\hat{f}_n(x)=\frac{1}{2nh}\sum_{i=1}^{n}I(x-h<X_i\le x+h)$, and $2nh\hat{f}_n(x)=\sum_{i=1}^{n}I(x-h<X_i\le x+h):=\sum_{i=1}^{n}Y_i$
where $Y_i\sim Ber(p(x))$, $p(x)=\int_{x-h}^{x+h}f(x)dx$. $2nh\hat{f}_n(x)\sim B(n,p(x))$

$$E(\hat{f}_n(x))=\frac{1}{2nh}E(2nh\hat{f}_n(x))=\frac{1}{2nh}p(x)$$

and $E(\hat{f}_x(x))^2=Var(\hat{f}_n(x))+[E(\hat{f}_n(x))]^2$

$$Var(\hat{f}_n(x))=\frac{1}{(2nh)^2}Var(2nh\hat{f}_n(x))=\frac{1}{(2nh)^2}np(x)[1-p(x)]$$

[Kernel Density Estimators]
h bandwidth - most important hyper-parameter, $K(\cdot)$ kernel function, $K_h(x)=K(y/h)/h$ bandwidth-rescaled kernel function

$$\hat{f}_n(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)=\frac{1}{n}\sum_{i=1}^{n}K_h\left(x-X_i\right):=\frac{1}{n}\sum g(X_i)$$

[Kernel Function]
Non-negative function $K(\cdot)$ with following condition, usually a pdf
(1) $\int_{-\infty}^{\infty}K(x)dx=1$
(2) $\int_{-\infty}^{\infty}xK(x)dx=0$
(3) $\int_{-\infty}^{\infty}x^2K(x)dx=\sigma_k^2>0$
Common kernel:
Uniform: $K(t)=\frac{1}{2}I(|t|<1)$
Gaussian (most popular): $K(t)=\frac{1}{\sqrt{2\pi}}exp(-t^2/2)$
Epanechnikov (most popular): $K(t)=\max(0.75(1-t^2),0)$
Biweight $K(t)=\max(15/16(1-t^2)^2,0)$
[Kernel MSE]
$MSE(\hat{f}(x))=\text{bias}^2\{\hat{f}(x)\}+\text{var}(\hat{f}(x))$
$E\hat{f}_n(x)=Eg(X_1)=\frac{1}{h}EK\left(\frac{x-X_i}{h}\right)=\frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(y)dy=\int K(t)f(x-ht)dt=\int K(t)\left[f(x)-htf'(x)+\frac{(ht)^2}{2}f''(x)+\cdots\right]dt$
$=f(x)+\frac{h^2}{2}f''(x)\int t^2K(t)dt+O(h^3)$

$$\text{bias}(\hat{f}_n(x))=E(\hat{f}_n(x))-f(x)=\frac{h^2}{2}f''(x)\int t^2K(t)dt+O(h^3)$$

$EK^2\left(\frac{x-X_i}{h}\right)=\int K^2\left(\frac{x-y}{h}\right)f(y)dy=h\int K^2(t)f(x-ht)dt=h\int K^2(t)[f(x)-htf'(x)+\frac{(ht)^2}{2}f''(x)+\cdots]dt=hf(x)\int K^2(t)dt+O(h^2)$

$$\mathrm{var}(\hat{f}_n(x)) = \frac{1}{n}\mathrm{var}(g(X_i)) = \frac{1}{nh^2}\left[EK^2\left(\frac{x-X_i}{h}\right) - \left(EK\left(\frac{x-X_i}{h}\right)^2\right)\right] = \frac{1}{nh}f(x)\int K^2(t)dt + O(1/n)$$

$$\mathrm{MSE}(\hat{f}_n(x)) = \frac{1}{nh}f(x)\left(\int K^2(t)dt\right) + \frac{h^4}{4}[f''(x)]^2\left(\int t^2K(t)dt\right)^2 + o\left(\frac{1}{nh} + h^4\right)$$

$$\mathrm{MISE}(\hat{f}_n(x)) = \int \mathrm{MSE}(\hat{f}_n(x))dx = \frac{1}{nh}\int K^2(t)dt + \frac{h^2}{4}\left(\int [f''(x)]^2dx\right)\left(\int t^2K(t)dt\right)^2 + o\left(\frac{1}{nh} + h^4\right)$$

condition required is $h \to 0$, $nh \to \infty$

[Unbiased C-V]

UCV is a better approach than conventional Cross Validation

$$\min_h UCV(h) = \int \hat{f}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i,n}(x_i)$$