Ng Xuan Jun

Useful Distributions:

Distribution	**Bernoulli**
Description	The random variable only has 2 possible outcomes. Probability of one of them is p
Notation	$X \sim Bernoulli(p)$
PMF	$P(X=k) = \begin{cases} p, & k=1 \\ 1-p, & k=0 \end{cases}$
Expectation	$E(X) = p$
Variance	$Var(x) = p(1-p)$
Properties	Indicator Function is a Bernoulli Random Variable $1_A = \begin{cases} 1, if\ A\ happens \\ 0, if\ A\ doesnt\ happen \end{cases}$
Distribution	**Binomial**
Description	Number of successes in n Bernoulli trials
Notation	$X \sim Bin(n,p)$
PMF	$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$, $for\ k = 0, 1, \cdots, n$
Expectation	$E(X) = np$
Variance	$Var(X) = np(1-p)$
Properties	If $X_1, \cdots, X_n$ are i.i.d. with distribution $Bernoulli(p)$, then $X_1 + \cdots + X_n \sim Bin(n,p)$
Distribution	**Geometric**
Description	Number of Bernoulli trials to obtain the **first** success
Notation	$X \sim Geometric(p)$
PMF	$P(X=k) = p(1-p)^{k-1}$ $for\ k = 1, 2, 3, \cdots$
Expectation	$E(X) = \frac{1}{p}$
Variance	$Var(X) = \frac{1-p}{p^2}$
Distribution	**Poisson**
Description	The number of events occurring in a fixed time interval or region of opportunity. Number of events per single unit of time
Notation	$X \sim Poi(\lambda)$
PMF	$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $k = 0, 1, 2, \cdots, \lambda > 0$
Expectation	$E(X) = \lambda$
Variance	$Var(X) = \lambda$
Properties	When n is large and p is small, np is moderate, $Bin(n,\ p) \rightarrow Poisson\ (np)$
Distribution	**Negative Binomial**
Description	Number of Bernoulli trials to obtain r successes
Notation	$X \sim NB(r,p)$
PMF	$P(X=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$ $for\ k = 1, 2, 3, \cdots$
Expectation	$E(X) = \frac{r(1-p)}{p}$
Variance	$Var(X) = \frac{pr}{(1-p)^2}$
Distribution	**Multinomial Distribution**
Description	n – Number of trials k – Number of mutually exclusive events
Notation	$X \sim NB(r,p)$
PMF	$\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$
Expectation	$E(X_i) = np_i$
Variance	$Var(X_i) = np_i(1-p_i)$ $Cov(X_i, X_j) = -np_ip_j (i \neq j)$
Distribution	**Uniform**
Notation	$X \sim Uniform(a,b)$
PDF	$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & otherwise \end{cases}$
CDF	$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x < b \\ 1, & b \le x \end{cases}$
Expectation	$E(X) = \frac{a+b}{2}$
Variance	$Var(X) = \frac{(b-a)^2}{12}$
Properties	$U(0,1) \equiv Beta(1,1)$ **Transform to Uniform(a, b) from U(0,1)**: $Y = (b-a)X + a$
Distribution	**Normal**
Description	$X \sim N(\mu, \sigma^2)$
PDF	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, $-\infty < x < \infty$
CDF	$F(x) = \int_{-\infty}^{\infty} f(x)dx$, $-\infty < x < \infty$
Expectation	$E(X) = \mu$
Variance	$Var(X) = \sigma^2$
Properties	If $Z \sim N(0,\ 1)$ then $\mu + \sigma Z \sim N(\mu, \sigma^2)$ Can complete the square if we have single exp then we see the distribution of it and the normalising constant is straightforward
Distribution	**Exponential**
Description	$X \sim Exp(\lambda)$
PDF	$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \le 0 \end{cases}$ Note that $\lambda > 0$
CDF	$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \le 0 \end{cases}$
Expectation	$E(X) = \frac{1}{\lambda}$
Variance	$Var(X) = \frac{1}{\lambda^2}$
Properties	For any $X \sim Exp(\lambda)$ $P(X > s + t \mid X > s) = P(X > t)$
Distribution	**Cauchy**
Description	$X \sim Cauchy(x_0, \gamma)$
Notation	$f(x) = \frac{1}{\pi\gamma}\left(\frac{1}{1 + \left(\frac{x - x_0}{\gamma}\right)^2}\right)$, $-\infty < x < \infty$
PDF	$F(x) = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}$, $-\infty < x < \infty$
Inverse PDF	$F^{-1}(u) = \gamma \tan[\pi(u - 0.5)] + x_0$, $u \in [0,1]$
Distribution	**Standard Cauchy**
Description	$X \sim Cauchy(0,1)$
Notation	$f(x) = \frac{1}{\pi}\left(\frac{1}{1+x^2}\right)$, $-\infty < x < \infty$
PDF	$F(x) = \frac{1}{\pi} \arctan x + \frac{1}{2}$, $-\infty < x < \infty$
Inverse PDF	$F^{-1}(u) = \tan[\pi(u - 0.5)]$, $u \in [0,1]$
Distribution	**Gamma**
Notation	$X \sim Gamma(a,b)$
PDF	$g(x) = \begin{cases} \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda t}, & x \ge 0 \\ 0, & t < 0 \end{cases}$
CDF	$G(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha,\ \lambda X)$
Expectation	$E(X) = \frac{\alpha}{\lambda}$
Variance	$Var(X) = \frac{\alpha}{\lambda^2}$
Properties	**Gamma Function**: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ **Recursion Property**: $\Gamma(a+1) = a\Gamma(a)$ **Gamma Function Computation**: $\Gamma(x) = (x-1)!$ **Sum of Gamma Random Variables**: Let $X_1, \cdots, X_k$ be independent random variables, assume $X_i \sim Gamma(a_i, b)$ for each i. Then $X_1 + \cdots + X_k \sim Gamma\ (a_1 + \cdots + a_k, b)$
	Connection with Standard Normal: If $Z \sim N(0,1)$ then $Z^2 \sim Gamma\left(\frac{1}{2}, \frac{1}{2}\right) \sim \chi^2(1)$ **Connection with Chi Squared**: Assume $Z_1, \cdots, Z_k$ are i.i.d N(0, 1) random variables. Then $Z_1^2 + \cdots + Z_k^2 \sim Gamma\left(\frac{k}{2}, \frac{1}{2}\right) \sim \chi^2(k)$ **Connection with Exponential Distribution**: If $X_1, \cdots, X_n$ i.i.d $Exp(\lambda) = Gamma(1, \lambda)$. Then $X_1 + \cdots + X_n \sim Gamma(n, \lambda)$ **Scaling**: If $X \sim Gamma(a,b)$ then $\lambda X \sim Gamma\left(a, \frac{b}{\lambda}\right)$
Distribution	**Beta**
Notation	$X \sim Beta(a,b)$
PDF	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{b-1}$, $0 \le x \le 1, \quad a > 0, b > 0$
CDF	$G(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha,\ \lambda X)$
Expectation	$E(X) = \frac{a}{a+b}$
Variance	$Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$
Properties	**Swap of parameters**: If $X \sim Beta(a,b)$, then $1 - X \sim Beta(b,a)$ If $X \sim Gamma(a, \beta)$, $Y \sim Gamma(b, \beta)$ and X, Y are independent, then $\frac{X}{X+Y} \sim Beta(a,b)$ **Order Statistics**: If $X_1, \cdots, X_n$ are i.i.d from Uniform(0,1) and $X_{(1)} \le \cdots \le X_{(n)}$ are their order statistics, then for $k = 1, \cdots, n$ $X_{(k)} \sim Beta(k, n+1-k)$ • Useful to know when we want to generate Beta distribution, we can just draw iid uniform and order them then pick the kth one

Transformation of Random Variables

Scaling and Shifting of Random Variables: Suppose that X is a continuous random variable with pdf $f(x)$

- **Shift**: If a is a real number, then pdf of $X + a$ is $f(x - a)$
- **Scale**: If b is a positive number, then the pdf of bX is $b^{-1} f\left(\frac{x}{b}\right)$

Change of variable formula: Suppose U and V are functions of X and Y, $u = g_1(x,\ y)\ \ v = g_2(x,\ y), J(x,\ y) \neq 0$

Multivariable Joint Density of U and V:

$$f_{UV}(u,\ v) = f_{XY}(h_1(u,\ v),\ h_2(u,\ v))|J(x,\ y)|$$

Note that $h_1(u,\ v)$ is x represented by u, v only. $h_2(u,\ v)$ is y represented by u, v only.

Jacobian: $J(x,\ y) = det \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} = \left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial v}{\partial y}\right) - \left(\frac{\partial v}{\partial x}\right)\left(\frac{\partial u}{\partial y}\right)$,

Rows – Functions, Columns, Variables

Single Variable: Suppose $g(x)$ is a one-to-one differentiable function. If X has pdf $f_X(x)$ and $Y = g(X)$ then pdf of Y is:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{dg^{-1}(y)}{dy}\right|$$

$g^{-1}(y)$ is just x in terms of y and we substitute it into f_x

Note that if $g(x, y)$ is not one-to-one, we break it into intervals such that it is one to one and we just add up the distribution on the range where they are one-to-one

Sample Estimators:

Let $X_1, X_2, \cdots$ be a sequence of iid random variables with mean μ and variance σ^2

Sample Mean: The mean of the sample that we are currently looking at

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad E(\bar{X}_n) = \mu$$

Sample Variance: Variance of the sample data that we are currently looking at

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad E(S_n^2) = \sigma^2$$

Variance of Sample Mean: Variation of the sample means that we will get over the n samples

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

Central Limit Theorem (CLT iid version): Suppose that the random variable X has finite second moment (i.e. $E[X^2] < \infty$), then the following **convergence in distribution** holds

$$\lim_{n\to\infty} \sqrt{n}(\bar{X}_n - \mu) = N(0, \sigma^2)$$

- $\bar{X}_n - \mu$ converges to 0 in order of $n^{-\frac{1}{2}}$
- In the multivariate case, replace σ^2 by the covariance matrix

Covariance:

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$
$$= E(XY) - E(X)E(Y)$$
$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Discrete Random Variable Generation:

Compute the probability for each of the possible values for the pmf. Then we just generate a uniform distribution to check which of the probability range it lies within.

Algorithm:

1. Generate $U \sim Uniform(0,1)$
2. If $U < p_0$, set $X = x_0$ and stop
3. If $U < p_0 + p_1$, set $X = x_1$ and stop
4. ⋮
5. Otherwise, set $X = x_n$

Inversion Method (Continuous Random Variable):

For a given random variable X, if we want to generate it, we can do the following. If we are given the pdf of the random variable X, $f(x)$:

1. Integrate $f(x)$ over the entire range to get the CDF, $F(x)$
2. Let $U \sim Uniform(0,1)$. Set $U = F(x)$ and find the inverse of the cdf $F^{-1}(u) = X$
3. Once we have found the inverse CDF, we can just generate a uniform distribution and put inside the inverse CDF to get one X

Algorithm:

1. Generate $U \sim Uniform(0,1)$
2. Set $X = F^{-1}(U)$

Note:

1. For $Exp(\lambda)$, we can use the following to generate the random variable:
$$X = -\frac{1}{\lambda}\log U$$
2. For Gamma(n, λ), we can use the following to generate the random variable.
Note that we are making use of the fact that Gamma(n, λ) is the sum of n Exp(λ)
$$X = -\frac{1}{\lambda}\log(U_1 \cdots U_n)$$

Rejection Sampling:

Quick ways to check that $\sup_x \frac{f(x)}{g(x)} < +\infty$ (But still need to rigorously show, this can give a brief idea)

1. Domain of $g(x)$ should cover the domain of $f(x)$
2. The tails of the proposal $g(x)$ should be heavier than the tails of $f(x)$

Rigorous ways to check:

1. Differentiate the ratio of $\frac{f(x)}{g(x)}$ and find the maximum value that the ratio can attain
2. Try to observe what will happen to the ratio $\frac{f(x)}{g(x)}$ when $x \to +\infty$. Look at what kind of function it will look like and make the conclusion from there

Theoretical number of simulations required to get 1 acceptance: $M = \sup_x \frac{f(x)}{g(x)}$

Logical steps to do (When we are computing):

1. Try to imagine the shape of $g(x)$ and $f(x)$, when one increases, the other should increase also, vice versa
2. Find the value of $M = \sup_x \frac{f(x)}{g(x)}$. Check that it exists and state the value where we can compute the maximum value using the rigorous way to check
3. Specify the rejection function of $\frac{f(x)}{Mg(x)}$ and our U needs to be within the rejection function range else it will be rejected
4. Generate Y using some kind of method (normally inversion)

Algorithm:

1. Generate $Y \sim g$
2. Generate $U \sim Uniform(0,1)$
3. If $U \le \frac{f(Y)}{Mg(Y)}$, then accept: set $X = Y$ and stop.
Otherwise, reject and return to step 1

Unknown Normalising Constant:

If we only know $f(x)$ up till a certain normalising constant, it will work the same, just take the ratio and supremum to be:

$$\frac{\tilde{f}(x)}{g(x)}, \qquad \sup_x \frac{\tilde{f}(x)}{\tilde{M}g(x)}$$

Polar Method for Bivariate Normal:

$$S = R^2 = X^2 + Y^2, \tan\theta = \frac{Y}{X}, X = Rcos(\theta), Y = Rsin(\theta)$$

Change of variable from (X,Y) to (S,θ)

$$f(s,\theta) = \frac{1}{2}e^{-\frac{s}{2}}\frac{1}{2\pi}, \qquad 0 < s < \infty, 0 < \theta < 2\pi$$

$$S = R^2 \sim Exp\left(\frac{1}{2}\right)\ \ and\ \ \theta \sim Uniform(0, 2\pi)$$

Box-Muller Algorithm v1:

1. Generate random numbers $U_1 \sim Uniform(0,1)$ and $U_2 \sim Uniform(0,1)$
2. Set:
$$X = \sqrt{-2\log U_1}\cos(2\pi U_2)$$
$$Y = \sqrt{-2\log U_1}\sin(2\pi U_2)$$

Box-Muller Algorithm v2: Suppose that (V_1, V_2) is uniformly distributed in the disk centered at $(0,0)$ with radius 1 and the random angle is $\theta \sim Uniform(0, 2\pi)$

Algorithm:

1. Generate random numbers $U_1 \sim Uniform(0,1)$ and $U_2 \sim Uniform(0,1)$
2. Set $V_1 = 2U_1 - 1, V_2 = 2U_2 - 1, S = V_1^2 + V_2^2$ $(V_1, V_2$ are just X and Y coordinates sampled from Uniform(-1, 1)
3. If $S > 1$, return to Step 1 (S is the radius squared for a unit disk so it should be ≤ 1)
4. Return the independent unit normals
$$X = \sqrt{-\frac{2\log S}{S}}V_1, \qquad Y = \sqrt{-\frac{2\log S}{S}}V_2$$

Simple Sampling:

Sample $X_1, X_2, \cdots, X_n$ independently from f, we can estimate the true parameter shown below by

$$\theta = E[\varphi(X)] = \int_S \varphi(X)f(x)\,dx$$

Simple Sampling Estimator: $\hat{\theta}_{SS} = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)$

Simple Sampling Exact Variance of $\hat{\theta}$ (Variance of Sample Mean): $Var(\hat{\theta}) = \frac{Var[\varphi(X)]}{n} = \frac{\int_S \varphi^2(x)f(x)dx - \theta^2}{n}$

Simple Sampling Asymptotic Variance of $\hat{\theta}$:

$$\sigma^2 \equiv Var[\varphi(X)] = \int_S \varphi^2(x)f(x)\,dx - \theta^2$$

Simple Sampling Estimated Asymptotic Variance of $\hat{\theta}$:

Note that this is not an unbiased estimate of σ^2

$$\hat{\sigma}_{SS}^2 = \frac{1}{n}\sum_{i=1}^{n}\varphi^2(X_i) - \hat{\theta}_{SS}^2$$

Simple Sampling Estimated Variance of $\hat{\theta}$ (Sample Variance): $\widehat{Var}(\hat{\theta}) = \frac{\hat{\sigma}_{SS}^2}{n}$

Simple Sampling Asymptotic Confidence Interval for θ:

$$\left[\hat\theta - 1.96\frac{\hat\sigma_{SS}}{\sqrt{n}}, \hat\theta + 1.96\frac{\hat\sigma_{SS}}{\sqrt{n}}\right]$$

Importance Sampling:

Sample $X_1, X_2, \cdots, X_n$ independently from g, we can estimate the true parameter shown below by

$$\theta = E_f[\varphi(X)] = \int_S \varphi(x) f(x)\,dx$$

$$\theta = \int_S \frac{\varphi(x)f(x)}{g(x)} g(x)\,dx = E_g\left[\frac{\varphi(X)f(X)}{g(X)}\right] = E_g[\varphi(Y)w(Y)]$$

Weighting Function: $w(y) = \frac{f(y)}{g(y)}$

Importance Sampling Estimator: Unbiased estimator of θ

$$\hat\theta_{IS} = \frac{1}{n}\sum_{i=1}^n \frac{\varphi(x_i)f(x_i)}{g(x_i)} = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)w(x_i)$$

Importance Sampling Exact Variance of $\hat\theta$ (Variance of Sample Mean):

$$Var(\hat\theta) = \frac{Var[\varphi(X)w(X)]}{n} = \frac{\int_S \frac{\varphi^2(x)f^2(x)}{g(x)}\,dx - \theta^2}{n}$$

Importance Sampling Asymptotic Variance of $\hat\theta$:

$$\sigma_{IS}^2 \equiv Var[\varphi(X)w(X)] = \int_S \frac{\varphi^2(x)f^2(x)}{g(x)}\,dx - \theta^2$$

Importance Sampling Estimated Asymptotic Variance of $\hat\theta$: Note that this is not an unbiased estimate of σ^2

$$\hat\sigma_{IS}{}^2 = \frac{1}{n}\sum_{i=1}^n \frac{\varphi^2(x)f^2(x)}{g(x)} - \hat\theta_{IS}{}^2$$

Importance Sampling Estimated Variance of $\hat\theta$ (Sample Variance): $\widehat{Var}(\hat\theta) = \frac{\hat\sigma_{IS}{}^2}{n}$

Importance Sampling Asymptotic Confidence Interval for θ: $\left[\hat\theta - 1.96\frac{\hat\sigma_{IS}}{\sqrt{n}}, \hat\theta + 1.96\frac{\hat\sigma_{IS}}{\sqrt{n}}\right]$

Optimal g: $g(x) \propto |\varphi(x)| \cdot f(x)$

- Asymptotic variance of $\hat\theta_{IS}$ with the proposal density g is exactly 0 if $\varphi(x) \geq 0$ for all $X \in S$

To find $g(x)$:

1. Let $h(x) = c|\varphi(x)|f(x)$
2. Let $1 = \int_S h(x) = \int_S c|\varphi(x)|f(x)$ and solve for c (Note that if we have the value for $I = \int_S |\varphi(x)|f(x)\,dx \Rightarrow c = \frac{1}{I}$
3. $g(x) = ch(x)$

Self-Normalizing Importance Sampling:

We only know the distribution of f and g up to a normalising constant ($Z_f > 0, Z_g > 0$)

$$f(x) = \frac{\tilde f(x)}{Z_f}, \quad g(x) = \frac{\tilde g(x)}{Z_g}$$

Generalised weights: $\tilde w(x) = \frac{\tilde f(x)}{\tilde g(x)}$, for all $x \in S$

Self-normalised importance sampling estimator of θ $= E_f[\varphi(X)] = \int_S \varphi(x)f(x)\,dx$:

$$\hat\theta_{SIS} = \frac{\sum_{i=1}^n \varphi(X_i)\tilde w(X_i)}{\sum_{i=1}^n \tilde w(X_i)}$$

Asymptotic variance of $\hat\theta_{SIS}$: Normally larger than the IS version because of the random denominator

$$\sigma_{SIS}^2 = E_g(w^2(X)\cdot[\varphi(X)-\theta]^2)$$

Where $w(x) = f(x)/g(x)$, it is the true weight

Exact Variance for $\hat\theta_{SIS}$: No closed form

- Note that $Var[\hat\theta_{SIS}] \neq \frac{\sigma_{SIS}^2}{n}$ (Not the same as simple sampling and importance sampling)

Estimator of the Variance of $\hat\theta_{SIS}$

$$\widehat{Var}(\hat\theta_{SIS}) = \frac{\hat\sigma_{SIS}^2}{n} = \frac{\sum_{i=1}^n\left\{\tilde w^2(X_i)[\varphi(X_i) - \hat\theta_{SIS}]^2\right\}}{\{\sum_{i=1}^n \tilde w(X_i)\}^2}$$

95% Asymptotic Confidence Interval

$$\left[\hat\theta - 1.96\sqrt{\frac{\hat\sigma_{SIS}^2}{n}}, \hat\theta + 1.96\sqrt{\frac{\hat\sigma_{SIS}^2}{n}}\right]$$

Calculus Results:

1. $\int_1^{+\infty}\frac{1}{x^p} = \frac{1}{-p+1}x^{-p+1}|_1^{+\infty} = \begin{cases} < +\infty \text{ if } p > 1 \\ +\infty \ \ if\ p \leq 1\end{cases}$
2. $\int_0^1 \frac{1}{x^p} = \frac{1}{-p+1}x^{-p+1}|_0^1 = \begin{cases} < +\infty \text{ if } p < 1 \\ +\infty \ \ if\ p \geq 1\end{cases}$
3. $\int_0^1 \frac{e^x}{x^p} = \begin{cases} < +\infty \text{ if } p < 1 \\ +\infty \ \ if\ p \geq 1\end{cases}$

Rare Event Estimation:

When the p^* we want to estimate is small

Relative Standard Deviation $= \frac{asymptotic\ s.d}{p^*}$

- Checks the magnitude of the asymptotic sd of our estimator as compared to the actual value → If it is large, it means that the magnitude of the sd of our estimator is larger than the actual value and it will give a very bad estimate

For a Bernoulli RV → $Relative\ s.d. = \frac{\sqrt{p(1-p)}}{p} = \sqrt{\frac{1-p}{p}}$

(Therefore, if the probability of it happening is low then the sd is high)

- Consider using a density centered at the point where we need more points so that the probability is higher and lowering the relative sd
- Remember that we can take the φ as the indicator function to indicate $P(X_i > 4)$ for example.

Control Variates Method:

Using $\hat h$ correlated with $\hat\theta$

$$\tilde\theta = \hat\theta + \beta\{\hat h - E_f[h(X)]\}$$

$$Var(\tilde\theta) = Var(\hat\theta) + \beta^2 Var(\hat h) + 2\beta Cov(\hat\theta, \hat h)$$

$$\beta^* = -\frac{Cov(\hat\theta,\hat h)}{Var(\hat h)}, Var^*(\tilde\theta) = \left(1 - Cor(\hat\theta,\hat h)^2\right)Var(\hat\theta)$$

$$Cor(\hat\theta,\hat h) = Cov(\hat\theta,\hat h)/\sqrt{Var(\hat\theta)Var(\hat h)}$$

Antithetic Variates Method

Useful Facts:

1) If X is generated from an inversion method from a cdf F, i.e. $X = F^{-1}(U)$ with $U \sim Uniform(0,1)$, then $X' = F^{-1}(1-U)$ also follows distribution F. Note that F^{-1} is a monotonic function
2) If we want to compare any other estimator with the Antithetic Estimator, remember to take sample size as $2n$
3) For **$U(0,1)$ U and $1-U$** are perfectly negatively correlated. For **$N(0,1)$, X and $-X$** are perfectly negatively correlated. We just need to find a monotonic function, either the inverse CDF or the function already given to us like e^x. Then it will be $h(X)$ and $h(1-X)$ or whatever that is perfectly negatively correlated.

Antithetic Estimator:

$$\hat I_{A_n} = \frac{1}{2n}\sum_{i=1}^n (h(U_i) + h(1-U_i))$$

Variance of Antithetic Estimator:

$$Var(\hat I_{A_n}) = \frac{1}{2n}[Var(h(U)) + Cov(h(U), h(1-U))]$$

Expectation-Minimization (EM) Algorithm:

Jensen Inequality: $f(E[X]) \leq E[f(x)]$

Important Result: For any positive function $\varphi(x)$ and a density $q(x)$

$$\log\left(\int \varphi(x)q(x)dx\right) \geq \int \log[\varphi(x)]\,q(x)\,dx$$

Likelihood Function: $L(Y|\theta)$ where Y is the observed data and θ are the parameters

Log-likelihood Function: $\ell(Y|\theta) = \log L(Y|\theta)$

Complete data log-likelihood:

$$\ell^c(Y,Z|\theta) = \log p(Y,Z|\theta) = \sum_{i=1}^n \log p(y_i, z_i|\theta)$$

Steps:

1) **Expectation:** Given θ_k, calculate the function $Q(\theta|\theta_k) = E_{z\in Z}[\ell^c(Y,Z|\theta)|Y,\theta_k]$
   - Note that we are taking expectation with respect to the latent variable z
2) **Maximisation:** Calculate the next iterate of the parameters θ $\theta_{k+1} = \arg\max_{\theta\in\Theta} Q(\theta|\theta_k)$
   - Differentiate the $Q(\theta|\theta_k)$ function to get the maximiser

Algorithm:

1. Initialize $\theta_0 = (\mu_1^{(0)}, \mu_2^{(0)}, \cdots)$ and $\epsilon > 0$
2. **E-Step:** In the kth iteration (given $\theta_k = (\mu_1^{(k)}, \mu_2^{(k)}, \cdots)$) calculate $\alpha_i^{(k,j)}$ for $i = 1, \cdots, n$ $j = 1, \cdots, m$. Where i is the ith observation and j is the jth parameter
3. **M-Step:** Calculate $\theta_{k+1} = (\mu_1^{(k+1)}, \mu_2^{(k+1)}, \cdots)$ using (2)
4. Iterate between the E-step and M-step until convergence. ($|\theta_{k+1} - \theta_k| < \epsilon$)

Tips:

1. For mixture distribution where we do not know the relative number of samples in each group, we can make use of an indicator $I_{\{Z_i=1\}}$ to indicate which density to use.

Note that once we look at the Q function, this will be $E_z[I(z_i = 1)|Y, \theta_k]$ and it becomes

$$\alpha_i^{(k,1)} = p(z_i = 1|y_i, \theta_k)$$

$$= \frac{p^{(k)}p(y_i|z_i = 1, \theta_k)}{p^{(k)}p(y_i|z_i = 1, \theta_k) + (1-p^{(k)})p(y_i|z_i = 2, \theta_k)}$$

$$\alpha_i^{(k,2)} = 1 - \alpha_i^{(k,2)}$$

If there is more than 1 latent variable, then it will still be the same idea. Just that we cannot take $1 - \alpha_i^{(k,2)}$ and denominator, we will sum over all possible combinations of z_i using the idea of Law of Total Probability

2. If we know the number of relative samples in each group, like the truncated Poisson, we may just need to take expectation over those z_i terms and for truncated Poisson it will be a *negative binomial distribution*. (Can look at the mean of it when we are computing expectation of $E[z|Y, \lambda_k]$

Markov Chains

Stationary Distribution: $\pi P = \pi$ where π is the stationary distribution and P is the transition matrix

- When finding the stationary distribution, we can consider the π of those states that are recurrent. Those that are transient will have probability 0 at the stationary distribution.

Transition Matrix P: Tells us the probability of going from a location to the next. Every row is the starting position and the columns are the position that we are going next.

Nice Properties:

1) For nice Markov Chains (Irreducible, Positive Recurrent, Aperiodic), in the long run, it will converge to its stationary distribution.
2) The empirical distribution of $\{X_1, X_2, \cdots, X_T\}$ as $T \to \infty$ (Noting that they are dependent samples since it is a Markov Chain) is close to the stationary distribution.

Properties of Stationary Distributions:

1) The stationary distribution π exists and is unique, if the Markov chain is irreducible and positive recurrent
2) $\lim_{t\to\infty} P^t = 1\pi$ where $1 = (1, \cdots 1)^T$ holds if the Markov chain is irreducible, positive recurrent and aperiodic
3) To find the stationary distribution π, solve the system of equation $\pi P = \pi$ or use the detailed balance condition.
4) $\lim_{(t\to\infty)} p_{ab}(t) = \pi_b$ where π is the stationary distribution
5) If for each row, the probability is equal for any non-zero entries. Then the stationary distribution for π_i will be the number of neighbours it has / total number of neighbours for everyone. (i.e. ½ for row 1 and 1/3 for row 2, $\pi_1 = 2/5$)

Terminologies:

1) **Irreducible Sets:** All states within this set are accessible from one another
2) **Positive Recurrent:** This happens if $P[\tau_{ii} < \infty] = 1$ for some (and hence all) states i. This means that we can always come back to this state after a finite number of steps. (**Always for finite state space**)
3) **Transient:** If we are not able to come back to a state after a finite number of ways after we leave
4) **Aperiodic:** We should be come back to the same point in irregular moves

Ergodic Theorem: If X is irreducible and positive recurrent (**Only concerned with recurrent states**), and a function $h(x)$ satisfies $E_\pi[|h(x)|] < \infty$, then

$$\frac{1}{N}\sum_{i=1}^N h(X_t) \to E_\pi[h(X)], \quad as\ N \to \infty, with\ probability\ 1$$

Where $E_\pi[h(X)] = \sum_i h(i)\pi_i$, expectation of $h(x)$ with respect to $\pi(.)$

Bayesian Statistics:

Posterior ∝ Likelilood × Prior

$$\pi(\theta|Y) \propto P(Y|\theta)\pi(\theta)$$

Note that the denominator for the usual Bayes Theorem is a constant in terms of θ since we integrate out θ so under the proportionality, we can remove it.

To find the posterior, just find the likelihood first and then multiply with the prior. Under proportionality, we can do that and try to find the form of it

Useful Trick for Splitting:

$$\sum_{i=1}^n (y_i - \theta)^2 = \sum_{i=1}^n (y_i - \bar y + \bar y - \theta)^2$$

$$= \sum_{i=1}^n (y_i - \bar y)^2 + \sum_{i=1}^n (\bar y - \theta)^2 + 2\sum_{i=1}^n (y_i - \bar y)(\bar y - \theta)$$

$$= \sum_{i=1}^n (y_i - \bar y)^2 + n(\bar y - \theta)^2$$

- $2\sum_{i=1}^n (y_i - \bar y)(\bar y - \theta) = 0$ because $\sum_{i=1}^n (y_i - \bar y) = 0$ but not the case for the squared version
- $(\bar y - \theta)^2$ independent of i, therefore, we can just sum over n of them.

Metropolis Hastings Algorithm:

Symmetric Transition Kernels:

1. **Uniform kernel:** $Q(\theta_a, .)$ is the density of Uniform$[\theta_a - \delta, \theta_a + \delta]$, for some user-set $\delta > 0$.
   $$Q(\theta_a, \theta_b) = \frac{1}{2\delta}, \quad for\ |\theta_b - \theta_a| < \delta$$
2. **Normal kernel:** $Q(\theta_a, .)$ is the density of $N(\theta_a, \delta^2)$, for some user-set $\delta^2 > 0$.
   $$Q(\theta_a, \theta_b) = \frac{1}{\sqrt{2\pi\delta}}\exp\left\{-\frac{(\theta_b - \theta_a)^2}{2\delta^2}\right\}$$

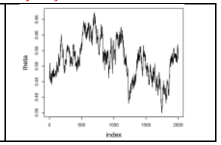
Algorithm:

1. Set $\theta^{(0)}$ to some initial value $\theta^{(0)} \in \Theta$
2. For $t = 1, \cdots, T$
   a. Draw θ_1^* from $N(\theta_1^{(t-1)}, \delta_1^2), \cdots, \theta_n^*$ from $N(\theta_n^{(t-1)}, \delta_n^2)$. Could be common δ^2. Check transition kernel formula
   b. Compute the acceptance probability
   $$\alpha(\theta^{(t)}, \theta^*) = \min\left(1, \frac{p(Y|\theta^*)\pi(\theta^*)Q(\theta^*, \theta^{(t)})}{p(Y|\theta^{(t)})\pi(\theta^{(t)})Q(\theta^{(t)}, \theta^*)}\right)$$
   Note that if the transition kernel is symmetric, then it will reduce to
   $$\alpha(\theta^{(t)}, \theta^*) = \min\left(1, \frac{p(Y|\theta^*)\pi(\theta^*)}{p(Y|\theta^{(t)})\pi(\theta^{(t)})}\right)$$
   c. Generate $U \sim Uniform(0,1)$
   d. If $U < \alpha(\theta^{(t-1)}, \theta^*)$, then accept and set $\theta^{(t)} = \theta^*$
   e. Otherwise reject θ^* set $\theta^{(t)} = \theta^{(t-1)}$
3. Repeat until $t = T$. Return $\{\theta^{(t)}\}_{t=1}^T$

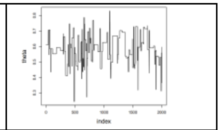
Trace Plots for Corresponding δ^2:

Size is relative to the actual value of δ^2 for the data

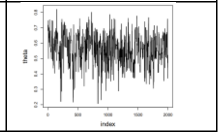
1. **Small δ^2** → High acceptance rate, High serial dependence, bad mixing



2. **Large δ^2** → Low acceptance rate, High serial dependence, bad mixing



3. **Proper choice of δ^2** → Moderate acceptance rate, good mixing



Optimal Acceptance Rate:

- For 3D and above: 0.234
- For 1D and 2D: ~ 30%-50%+

Gibbs Sampler:

Computing the Marginal Conditional Posteriors

$$f(x|y,z) = \frac{f(x,y,z)}{f(y,z)}$$

$$\propto f(x,y,z)$$

- Conditional Marginal is Proportional to the Joint Posterior because the denominator is constant with rest to x.
- Therefore, we just need to take out the multiplicative constants (because of proportionality) to find the marginal conditional posterior densities

We use the most updated values to update the rest.

Algorithm:

1. Initialise $\theta^{(0)} \in \Theta$ where $\theta = (\theta_1, \cdots, \theta_d)$
2. At the step t ($t = 1, \cdots, T$), we do the following sequentially
   a. Sample $\theta_1^{(t)} \sim \pi(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_d^{(t-1)}, Y)$;
   b. Sample $\theta_2^{(t)} \sim \pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_d^{(t-1)}, Y)$;
   c. …
   d. Sample $\theta_d^{(t)} \sim \pi(\theta_d|\theta_1^{(t)}, \theta_3^{(t)}, \cdots, \theta_{d-1}^{(t)}, Y)$;
3. Repeat the steps above until time $t = T$. Output $\{(\theta_1^{(t)}, \cdots, \theta_d^{(t)})\}_{t=1}^T$

Dirichlet Distribution:

$$f(x_1, x_2, \cdots, x_d) \propto x_1^{a_1-1}x_2^{a_2-1}\cdots x_d^{a_d-1}$$

$$0 < x_1 < 1, \cdots, 0 < x_d < 1, \quad x_1 + x_2 + \cdots + x_d = 1$$

- Degenerate distribution because only $d-1$ free points because $x_d = 1 - x_1 - x_2 - \cdots - x_{d-1}$

Property: They are marginally Beta Distributions

If $(X_1, \cdots X_d) \sim Dirichlet(a_1, \cdots, a_d)$

$$f(x_1, \cdots, x_d) \propto x^{a_1-1}\cdots x_d^{a_d-1}$$

Then for every $i = 1, \cdots, d$ $X_i \sim Beta(a_i, \sum_{j\neq i} a_j)$

Trick for Beta and Dirichlet Distribution:

If we have multiple variables and we are unable to get the $\theta(1-\theta)$ form

$$f(x|y,z) \propto x^{a_1}(1-x-y-z)^{a_2}$$

If we see this, just need to divide with $1-y-z$ (Everything other than the variable we are concerned about)

$$f(x|y,z) \propto \left(\frac{x}{1-y-z}\right)^{a_1}\left(1 - \frac{x}{1-y-z}\right)^{a_2}$$

$$U = \frac{X}{1-Y-Z} \sim Beta(a_1 + 1, a_2 + 1)$$

$$X \sim (1-Y-Z)U$$

- Note that because this is a linear transformation, we can just multiply (1-Y-Z) to U or else we need to do the change-of-variable formula

Trick for Binomial

- Try to find the $\frac{1}{z!(n-z)!}(\theta)^z(1-\theta-\eta)^{n-z}$
- Then the first term can multiply by $n!$ Since proportionality, for the power z and power $n-z$ term. Just need to add them together and then divide to both under the denominator.

$$\frac{n!}{z!\,(n-z)!}\left(\frac{\theta}{1-\eta}\right)^z\left(\frac{1-\theta-\eta}{1-\eta}\right)^{n-z} \sim Bin\left(n, \frac{\theta}{1-\eta}\right)$$