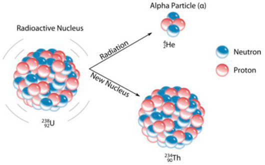
STATS 100A: Advanced

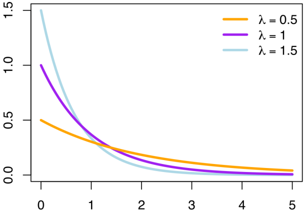
Ying Nian Wu

Department of Statistics
University of California, Los Angeles

Some pictures are taken from the internet.
Credits belong to original authors.

Particle decay





T: time until decay.
$T \sim \text{Exponential}(\lambda)$.
$P(T \in (t, t + \Delta t)) = f(t)\Delta t = \lambda e^{-\lambda t} \Delta t$.

Make a movie

Divide the time into small intervals of length Δt (e.g., 1/24 second, or 1/100 second).

Show a picture at 0, Δt, $2\Delta t$, ...
Give an illusion of continuous time process as $\Delta t \to 0$.

Bank account

Divide $[0, t]$ into n small intervals, $\Delta t = t/n$.
Interest rate $= r$.
Time 0: $\$1$.
Time Δt: $\$(1 + r\Delta t)$.
Time $2\Delta t$: $\$(1 + r\Delta t)^2$.
Time $3\Delta t$: $\$(1 + r\Delta t)^3$.
...
Time $t = n\Delta t$: $\$(1 + r\Delta t)^n$.

$$\left(1 + r\frac{t}{n}\right)^n \to e^{rt},$$

as $n \to \infty$ or $\Delta t \to 0$.

Exponential

Bank account

Divide $[0, t]$ into n small intervals, $\Delta t = t/n$.
Interest rate $= r$.

$$\left(1 + \frac{1}{n}\right)^n \to e.$$

$$1 + \frac{1}{n} \doteq e^{1/n}.$$

$$1 + \Delta x \doteq e^{\Delta x}.$$

$$\left(1 + r\frac{t}{n}\right)^n \to e^{rt}.$$

$$(1 + r\Delta t)^{t/\Delta t} \doteq \left(e^{r\Delta t}\right)^{t/\Delta t} = e^{rt}.$$

Poisson process

Flip a coin within each interval.
$p = \lambda \Delta t$ (e.g., $\Delta t = 1$ hour. $\lambda =$ once every 10 year.
$\lambda \Delta t = 1/3650 \times 1/24$).
Geometric waiting time

$$
\begin{aligned}
P(T \in (t, t+\Delta t)) &= (1-\lambda\Delta t)^{t/\Delta t}\lambda\Delta t \\
&\doteq \left(e^{-\lambda\Delta t}\right)^{t/\Delta t}\lambda\Delta t = e^{-\lambda t}\lambda\Delta t.
\end{aligned}
$$

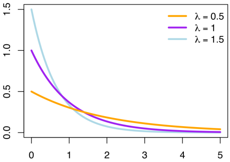
Exponential distribution

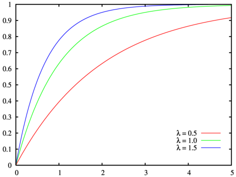
Flip a coin within each interval.
$p = \lambda \Delta t$ (e.g., $\Delta t = .001$ second. $\lambda =$ once every minute.
$\lambda \Delta t = 1/60 \times .001$).

Exponential waiting time

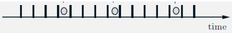
$$\frac{P(T \in (t, t + \Delta t))}{\Delta t} = \lambda e^{-\lambda t}.$$

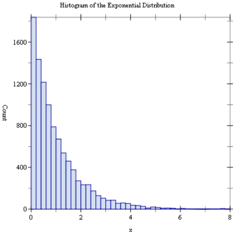
$$P(T > t) = (1 - \lambda \Delta t)^{t/\Delta t} \doteq (e^{-\lambda \Delta t})^{t/\Delta t} = e^{-\lambda t}.$$





Exponential = geometric





1 million particles decay in different period. Each small period is a bin.

Geometric waiting time

We can write $T = \tilde{T}\Delta t$, where $\tilde{T} \sim \text{Geometric}(p = \lambda\Delta t)$. Then

$$\mathbb{E}(T) = \mathbb{E}(\tilde{T})\Delta t = \frac{1}{p}\Delta t = \frac{1}{\lambda\Delta t}\Delta t = 1/\lambda.$$

Poisson distribution

Flip a coin within each interval.
Let X be the number of heads within $[0, t]$, then
$X \sim \text{Binomial}(n = t/\Delta t, p = \lambda \Delta t)$.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \to \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

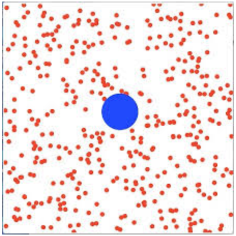
$\mathbb{E}(X) = np = (t/\Delta t)(\lambda \Delta t) = \lambda t$.
$\lambda = \mathbb{E}(X)/t$, rate or intensity.

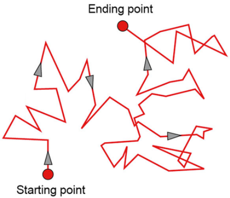
Poisson distribution

$$
\begin{aligned}
P(X=k) &= \frac{n(n-1)...(n-k+1)}{k!}p^k(1-p)^{n-k} \\
&= \frac{t/\Delta t(t/\Delta t-1)...(t/\Delta t-k+1)}{k!} \\
&\times (\lambda\Delta t)^k(1-\lambda\Delta t)^{t/\Delta t-k} \\
&= \frac{t(t-\Delta t)(t-2\Delta t)...(t-(k-1)\Delta t)}{k!} \\
&\times \lambda^k(1-\lambda\Delta t)^{t/\Delta t}(1-\lambda\Delta t)^{-k} \\
&\to \frac{t^k}{k!}\lambda^k(e^{-\lambda\Delta t})^{t/\Delta t} = \frac{(\lambda t)^k}{k!}e^{-\lambda t}.
\end{aligned}
$$

Diffusion or Brownian motion

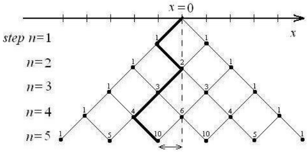
Dust particle in water

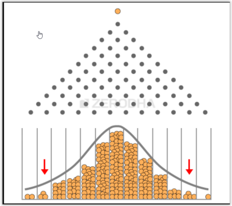




Recall random walk

Either go forward or backward by flipping a fair coin.





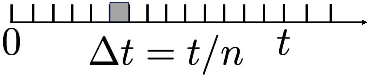
Number of heads $Y \sim \text{Binomial}(n, 1/2)$, then random walk ends up at X,

$$X = Y - (n - Y) = 2Y - n.$$

$$X = \epsilon_1 + \epsilon_2 + ... + \epsilon_n.$$

$\epsilon_i = 1$ or -1 with probability 1/2 each.

Discretize time and space

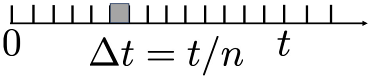


(1) Time: Divide $[0, t]$ into n intervals, $\Delta t = t/n$ (time unit).
(2) Space: Within each small time interval, move forward or backward by Δx (space unit).
$P(\epsilon_i = 1) = P(\epsilon_i = -1) = 1/2$. ϵ_i are independent.

$$X = \sum_{i=1}^{n} \epsilon_i \Delta x = (Y - (n - Y))\Delta x = (2Y - n)\Delta x.$$

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(\epsilon_i)\Delta x = \mathbb{E}(2Y - n)\Delta x = 0.$$

Diffusion or Brownian motion



$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(\epsilon_i)\Delta x^2 = n\Delta x^2 = \frac{t}{\Delta t}\Delta x^2.$$

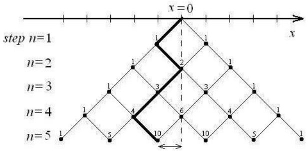
$$\mathrm{Var}(X) = \mathrm{Var}((2Y-n)\Delta x) = 4\mathrm{Var}(Y)\Delta x^2 = n\Delta x^2.$$

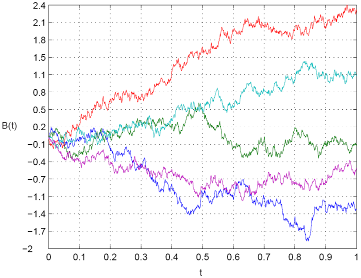
$$\Delta x^2/\Delta t = \sigma^2;\ \Delta x = \sigma\sqrt{\Delta t};\ \mathrm{Var}(X) = \sigma^2 t.$$

$$\text{velocity} = \Delta x/\Delta t = \sigma/\sqrt{\Delta t} \to \infty.$$

Einstein, σ related to the size of molecules.

Diffusion or Brownian motion





Brownian motion:

$$X_{t+\Delta t} = X_t + \sigma\sqrt{\Delta t}\epsilon_t,$$

where $\mathbb{E}(\epsilon_t) = 0$, $\mathrm{Var}(\epsilon_t) = 1$, and ϵ_t are iid.
Nowhere differentiable.
σ: volatility of stock price, basis for option pricing.
A drop of milk (millions of particles) diffuses in coffee.

Normal approximation

100A

Ying Nian Wu

Process

Transformation

Entropy

Central limit theorem

$P(\epsilon_i = 1) = P(\epsilon_i = -1) = 1/2$. ϵ_i are independent.

$$X = \sum_{i=1}^{n} \epsilon_i \Delta x = (2Y - n)\Delta x \sim \mathrm{N}(0, \sigma^2 t),$$

as $n \to 0$.

Sum of independent random variables $\sim$ Normal distribution.

Normal approximation

$X \sim \text{Binomial}(n, 1/2)$. $\mu = \mathbb{E}(X) = n/2$,
$\sigma^2 = \text{Var}(X) = n/4$, $\sigma = SD(X) = \sqrt{n}/2$.
Let

$$Z = \frac{X - \mu}{\sigma} = \frac{X - n/2}{\sqrt{n}/2},$$

then $\mathbb{E}(Z) = 0$, $\text{Var}(Z) = 1$, no matter what n is.
Z takes discrete values, with spacing $\Delta z = 1/\sigma = 2/\sqrt{n}$.

$$P(Z \in (a, b)) = \sum_{z \in (a,b)} p(z) \doteq \sum_{z \in (a,b)} f(z)\Delta z \to \int_a^b f(z)dz,$$

where $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is the density of $\text{N}(0, 1)$.

$$p(z)/\Delta z \to f(z).$$

Proof

Step 1:

$$p(0) \doteq \frac{1}{\sqrt{2\pi}} \Delta z.$$

Step 2:

$$\frac{p(z)}{p(0)} \doteq e^{-z^2/2}.$$

$X = \mu + Z\sigma = n/2 + Z\sqrt{n}/2.$

$$p(0) = P(X = n/2).$$

$$\frac{p(z)}{p(0)} = \frac{P(X = n/2 + z\sqrt{n}/2)}{P(X = n/2)} = \frac{P(X = n/2 + d)}{P(X = n/2)}.$$

Proof

$$P(X=k) = \frac{\binom{n}{k}}{2^n} = \frac{n!}{k!(n-k)!2^n},$$

For big n,

$$n! \sim \sqrt{2\pi n} n^n e^{-n},$$

$$\begin{aligned}
P(X=n/2) &\sim \frac{n!}{(n/2)!^2 2^n} \\
&\sim \frac{\sqrt{2\pi n} n^n e^{-n}}{(\sqrt{2\pi(n/2)}(n/2)^{n/2})^2 2^n} \\
&\sim \frac{1}{\sqrt{2\pi}} \frac{2}{\sqrt{n}}.
\end{aligned}$$

Proof

Let $k = \mu + z\sigma = n/2 + z\sqrt{n}/2 = n/2 + d$.

$$
\begin{aligned}
\frac{P(X = n/2 + d)}{P(X = n/2)} &= \frac{\binom{n}{n/2+d}}{\binom{n}{n/2}} \\
&= \frac{n!/[(n/2+d)!(n/2-d)!]}{n!/[(n/2)!(n/2)!]} \\
&= \frac{(n/2)!(n/2)!}{(n/2+d)!(n/2-d)!} \\
&= \frac{(n/2)(n/2-1)...(n/2-(d-1))}{(n/2+1)(n/2+2)...(n/2+d)} \\
&= \frac{1(1-2/n)(1-2\times 2/n)...(1-(d-1)\times 2/n)}{(1+2/n)(1+2\times 2/n)...(1+d\times 2/n)} \\
&= \frac{(1-\delta)(1-2\delta)...(1-(d-1)\delta)}{(1+\delta)(1+2\delta)...(1+d\delta)}
\end{aligned}
$$

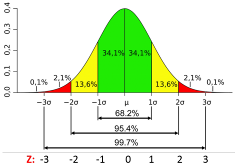
Proof

$$
\begin{aligned}
&\to \frac{e^{-\delta}e^{-2\delta}...e^{-(d-1)\delta}}{e^{\delta}e^{2\delta}...e^{d\delta}} \\
&= \frac{e^{-(1+2+...+(d-1))\delta}}{e^{(1+2+...+d)\delta}} \\
&= \frac{e^{-d(d-1)\delta/2}}{e^{d(d+1)\delta/2}} \\
&= e^{-[d(d-1)/2+d(d+1)/2]\delta} = e^{-d^2\delta} \\
&= e^{-(z\sqrt{n}/2)^2(2/n)} = e^{-\frac{z^2}{2}},
\end{aligned}
$$

where $\delta = 2/n$, and $d = z\sqrt{n}/2$.

Normal approximation

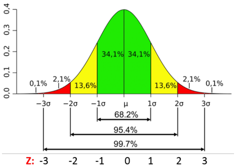
Let $X \sim \text{Binomial}(n, p)$, sum of independent Bernoulli.
$\mathbb{E}(X) = np$, $\text{Var}(X) = np(1-p)$.
$\mathbb{E}(X/n) = p$, $\text{Var}(X/n) = p(1-p)/n$.
Approximately,
$X \sim \text{N}(np, np(1-p))$.
$X/n \sim \text{N}(p, p(1-p)/n)$.
e.g., $n = 100$, $p = 1/2$. $X \sim \text{N}(50, 25)$.
$P(X \in [50 - 2 \times 5, 50 + 2 \times 5]) = P(X \in [40, 60]) = 95\%$.



Recall $\sum_{k=40}^{60} \binom{100}{k} / 2^{100} \to$ integral.

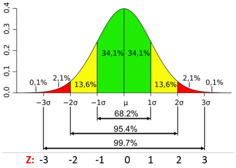
Normal approximation

Let $X \sim \mathrm{Binomial}(n, p)$, sum of independent Bernoulli.
$\mathbb{E}(X) = np$, $\mathrm{Var}(X) = np(1-p)$.
$\mathbb{E}(X/n) = p$, $\mathrm{Var}(X/n) = p(1-p)/n$.
Approximately,
$X \sim \mathrm{N}(np, np(1-p))$.
$X/n \sim \mathrm{N}(p, p(1-p)/n)$.
e.g., Polling $n = 100$, $p = .2$. $X/n \sim \mathrm{N}(.2, .04^2)$.
$P(X/n \in [.2 - 2 \times .04, .2 + 2 \times .04]) = P(X/n \in [.12, .28]) = 95\%$.



Normal approximation

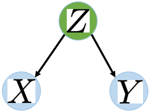
Let $X \sim \text{Binomial}(n, p)$, sum of independent Bernoulli.
$\mathbb{E}(X) = np$, $\text{Var}(X) = np(1-p)$.
$\mathbb{E}(X/n) = p$, $\text{Var}(X/n) = p(1-p)/n$.
Approximately,
$X \sim \text{N}(np, np(1-p))$.
$X/n \sim \text{N}(p, p(1-p)/n)$.
e.g., Monte Carlo $n = 10000$, $p = \pi/4$.
$4m/n \sim \text{N}(\pi, \pi(4-\pi)/10000)$.



Conditional independence

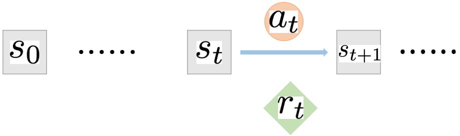
Markov: [future | present, past], [child | parent, grandparent]
$p(y|x,z) = p(y|z)$

Shared cause: [siblings | parent]
$p(x,y|z) = p(x|z)p(y|z)$



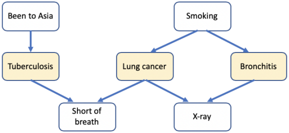
Markov decision process

state s_t, action a_t, reward r_t.



Dynamics: $p(s_{t+1} \mid s_t, a_t)$.
Policy: $\pi(a_t \mid s_t)$.
Reward: $p(r_t | s_t, a_t, s_{t+1})$.
Return: $R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...$.
Value: $V(s) = \mathbb{E}[R_t \mid s_t = s]$, $Q(s, a) = \mathbb{E}[R_t \mid s_t = s, a_t = a]$.
Reinforcement learning: find π to optimize $V(s_0)$.
Imagine 1 million people playing out.

Bayes net



a: Been to Asia; s: Smoking; t: Tuberculosis; l: Lung cancer; b: Bronchitis; d: Short of breath (Dyspnea); x: X-ray.

$$p(a, s, t, l, b, d, x) = p(a)p(s)p(t|a)p(l|s)p(b|s)p(d|t, l)p(x|b, l),$$

$$p(l|a, s, d, x) = \frac{p(l, a, s, d, x)}{p(a, s, d, x)},$$

$$p(l, a, s, d, x) = \sum_{t,b} p(a, s, t, l, b, d, x),$$

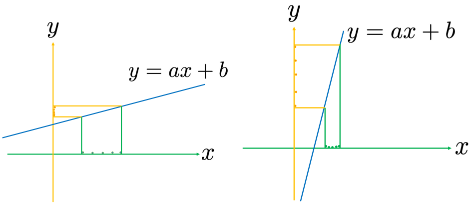
$$p(a, s, d, x) = \sum_{l} p(l, a, s, d, x).$$

Efficient calculation: message passing / belief propagation.

Linear transformation

Change of variable

$X \sim f(x)$, $Y = aX + b$ $(a > 0)$. $Y \sim g(y)$.



$$y = ax + b, \; x = (y - b)/a.$$

$$P(X \in (x, x + \Delta x)) = P(Y \in (y, y + \Delta y)).$$

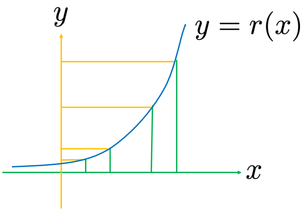
$$f(x)\Delta x = g(y)\Delta y.$$

$$g(y) = f(x)\frac{\Delta x}{\Delta y} = f((y - b)/a))/a.$$

Space warping, stretching or squeezing.

Non-linear transformation

$X \sim f(x)$, $Y = r(X)$, monotone. $Y \sim g(y)$.



$$y = r(x), \; x = r^{-1}(y).$$

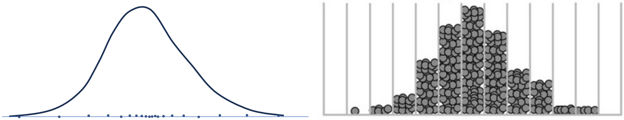
$$P(X \in (x, x + \Delta x)) = P(Y \in (y, y + \Delta y)).$$

$$f(x)\Delta x = g(y)\Delta y.$$

$$\Delta y / \Delta x = r'(x).$$

Locally linear, space warping.

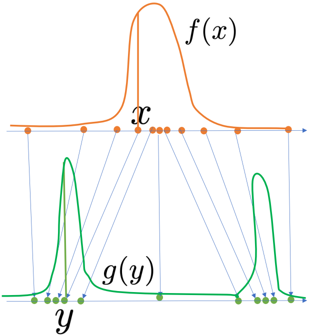
Space warping



Squeezing or stretching the bins $\rightarrow$ changes the density and histogram.

Non-linear transformation

$X \sim f(x)$, $Y = r(X)$, monotone. $Y \sim g(y)$.



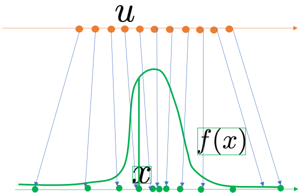
$$y = r(x),\ x = r^{-1}(y).$$

Order preserving mapping:

$$P(X \le x) = P(Y \le y).$$

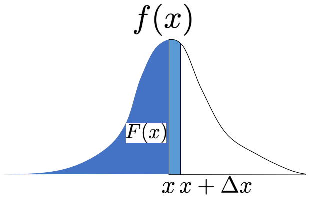
$$F(x) = G(y).$$

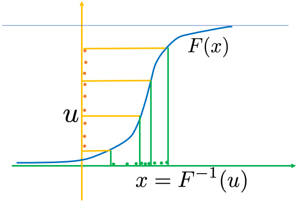
Inversion method



$U \sim \text{Unif}[0, 1]$.
$P(U \leq u) = P(X \leq x)$.
$u = F(x)$, $x = F^{-1}(u)$.
Population: $\{x_1, x_2, ..., x_N\}$ (ordered).
Sample $i \sim \text{Uniform}\{1, 2, ..., N\}$, return x_i.
$P(X \leq x_i) = i/N = F(x_i)$.
$U = i/N \sim \text{Uniform}[0, 1]$, $x_i = F^{-1}(U)$.

Inversion method





$U \sim \text{Unif}[0,1]$. $X = F^{-1}(U)$. Then $f(x) = F'(x)$ is the pdf of X.

$$P(U \in (u, u + \Delta u)) = P(X \in (x, x + \Delta x)).$$

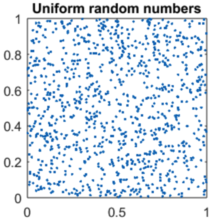
$$\Delta u = f(x)\Delta x.$$

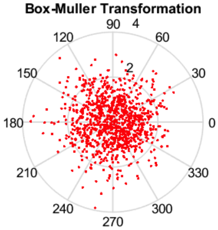
$$f(x) = \frac{\Delta u}{\Delta x} = F'(x).$$

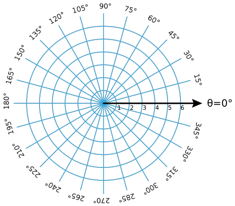
Inversion method

Suppose we want to generate $X \sim \text{Exponential}(1)$.
$F(x) = 1 - e^{-x}$.
$F(x) = u$, i.e., $1 - e^{-x} = u$, $e^{-x} = 1 - u$. $x = -\log(1 - u)$.
Generate $U \sim \text{Unif}[0, 1]$. Return $X = -\log(1 - U)$.

Polar method







Polar method

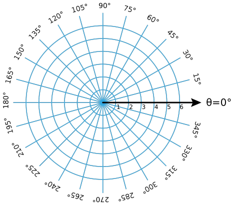
$X \sim \mathrm{N}(0,1)$, $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.
$Y \sim \mathrm{N}(0,1)$, $f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$.
X and Y are independent.

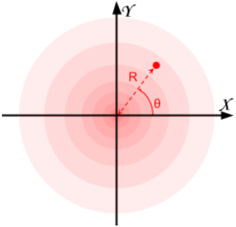
$$\begin{aligned} & P(X \in (x, x+\Delta x), Y \in (y, y+\Delta y)) \\ = \; & P(X \in (x, x+\Delta x)) \times P(Y \in (y, y+\Delta y)). \end{aligned}$$

$$f(x,y)\Delta x \Delta y = f(x)\Delta x \times f(y)\Delta y.$$

$$f(x,y) = \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right).$$

Polar method

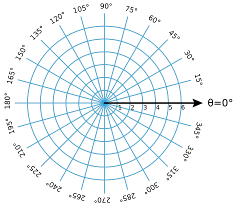


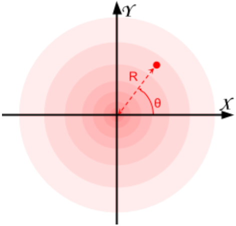


$x = r\cos\theta$, $y = r\sin\theta$.
Area of ring $R \in (r, r+\Delta r)) = 2\pi r \Delta r$.
Count proportion of points in the ring = density $\times$ area.

$$
\begin{aligned}
P(R \in (r, r+\Delta r)) &= \frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right) 2\pi r \Delta r \\
&= \exp\left(-\frac{r^2}{2}\right) r\Delta r = \exp\left(-\frac{r^2}{2}\right) d\frac{r^2}{2}.
\end{aligned}
$$

Polar method





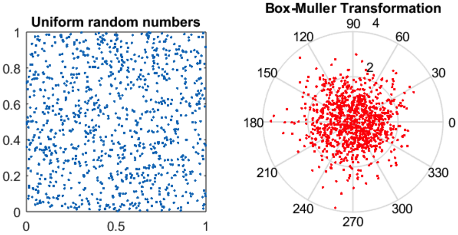
$x = r\cos\theta$, $y = r\sin\theta$.
Let $t = r^2/2$. $\Delta t = r\Delta r$.

$$P(T \in (t, t+\Delta t)) = P(R \in (r, r+\Delta r)).$$

$$f(t)\Delta t = \exp\left(-\frac{r^2}{2}\right) r\Delta r = \exp(-t)\Delta t.$$

$$T \sim \text{Exponential}(1).$$

Polar method



$$T = -\log(1 - U_1).$$
$$R = \sqrt{2T}.$$
$$\theta = 2\pi U_2.$$
$$X = R\cos\theta,\ Y = R\sin\theta.$$
$$(U_1, U_2) \rightarrow (X, Y).$$

Non-linear transformation

$X \sim f(x)$, $Y = r(X)$. $Y \sim g(y)$.
X consists of iid Gaussian $\mathrm{N}(0, 1)$ noises.
r is learned from training examples by neural network (deep learning).

Function

(1) Linear

$$h(x) = ax + b.$$

$$\mathbb{E}[h(X)] = \mathbb{E}(aX + b) = a\mathbb{E}(X) + b = h(\mathbb{E}(X)).$$

(2) Square

$$h(x) = x^2.$$

$$\mathbb{E}[h(X)] = \mathbb{E}(X^2);$$

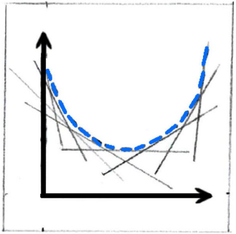
$$h(\mathbb{E}(X)) = [\mathbb{E}(X)]^2.$$

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \geq 0.$$

Question: expectation of function vs function of expectation?

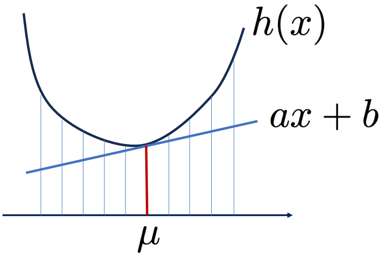
Convex function

Upper envelop and supporting lines



Supporting line at x_0 touches $h(x)$ at x_0, but below $h(x)$ at other places.

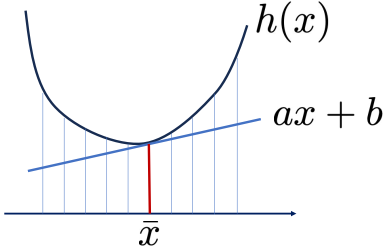
Jensen inequality



$$\mu = \mathbb{E}(X).$$
$$h(\mu) = a\mu + b.$$
$$h(x) \geq ax + b.$$

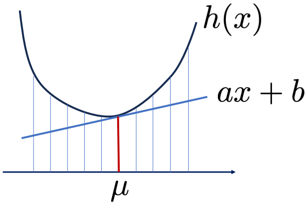
$$\begin{aligned}\mathbb{E}(h(X)) &\geq \mathbb{E}(aX + b) \\ &= a\mathbb{E}(X) + b \\ &= a\mu + b = h(\mu) = h(\mathbb{E}(X)).\end{aligned}$$

Jensen inequality



$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) \geq \frac{1}{n}\sum_{i=1}^{n}(ax_i+b)$$
$$= a\frac{1}{n}\sum_{i=1}^{n} x_i + b$$
$$= a\bar{x} + b = h(\bar{x}).$$

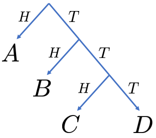
Utility



$\mu = \mathbb{E}(X)$.
Offer 1: Get $\$\mu$ with 100% probability.
Offer 2: Get $\$X \sim f(x)$, with $\mathbb{E}(X) = \mu$.
Utility $=$ perceived value of $\$x = h(x)$.
Convex: $\mathbb{E}[h(X)] \geq h(\mu)$. Prefer Offer 2, risk taking.
Concave: $\mathbb{E}[h(X)] \leq h(\mu)$. Prefer Offer 1, risk averse.
e.g., $h(x) = -x^2$, variance. $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2$.

Entropy

x	A	B	C	D
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
$-\log_2 p(x)$	1	2	3	3
coin	H	TH	TTH	TTT



Entropy $=$ expected number of coin flips

$$\mathbb{H}(p) = \mathbb{E}_p[-\log_2 p(X)] = \sum_x (-\log_2 p(x)) p(x)$$

$$= 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = 1.75 \text{ flips.}$$

Prefix code

x	A	B	C	D
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
$-\log_2 p(x)$	1	2	3	3
coin	H	TH	TTH	TTT
bit	1	01	001	000

e.g, $101100101000 =$ abacbd
Shortest code: sequence of coin flips, completely random sequence, cannot be compressed.
code length of $x = l(x) = -\log_2 p(x)$.
entropy $=$ expected code length: $\mathbb{H}(p) = \mathbb{E}[l(x)]$.

Kullback-Leibler divergence

x	A	B	C	D
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$
$-\log_2 p(x)$	1	2	3	3
$q(x)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$
$-\log_2 q(x)$	3	3	2	1

Coding by $q(x)$:

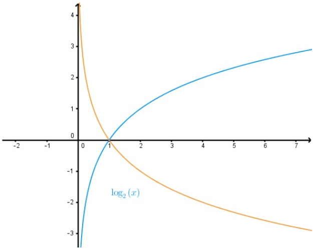
$$\mathbb{E}_p[-\log_2 q(X)] = \sum_x (-\log_2 q(x))p(x)$$
$$= 3 \times \frac{1}{2} + 3 \times \frac{1}{4} + 2 \times \frac{1}{8} + 1 \times \frac{1}{8} = \frac{21}{8} \text{ flips.}$$

Redundancy: Kullback-Leibler divergence.

$$\mathbb{D}_{\text{KL}}(p\|q) = \mathbb{E}_p[-\log q(X)] - \mathbb{E}_p[-\log p(X)]$$
$$= \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] = \sum_x \left(\log \frac{p(x)}{q(x)}\right) p(x) \geq 0.$$

Jensen inequality

$$\mathbb{E}_p\left[\frac{q(X)}{p(X)}\right] = \sum_x \left(\frac{q(x)}{p(x)}\right) p(x) = \sum_x q(x) = 1.$$



$$\mathbb{D}_{\mathrm{KL}}(p\|q) = \mathbb{E}_p\left[-\log\frac{q(X)}{p(X)}\right] \geq -\log\mathbb{E}_p\left[\frac{q(X)}{p(X)}\right] = 0.$$