Probability

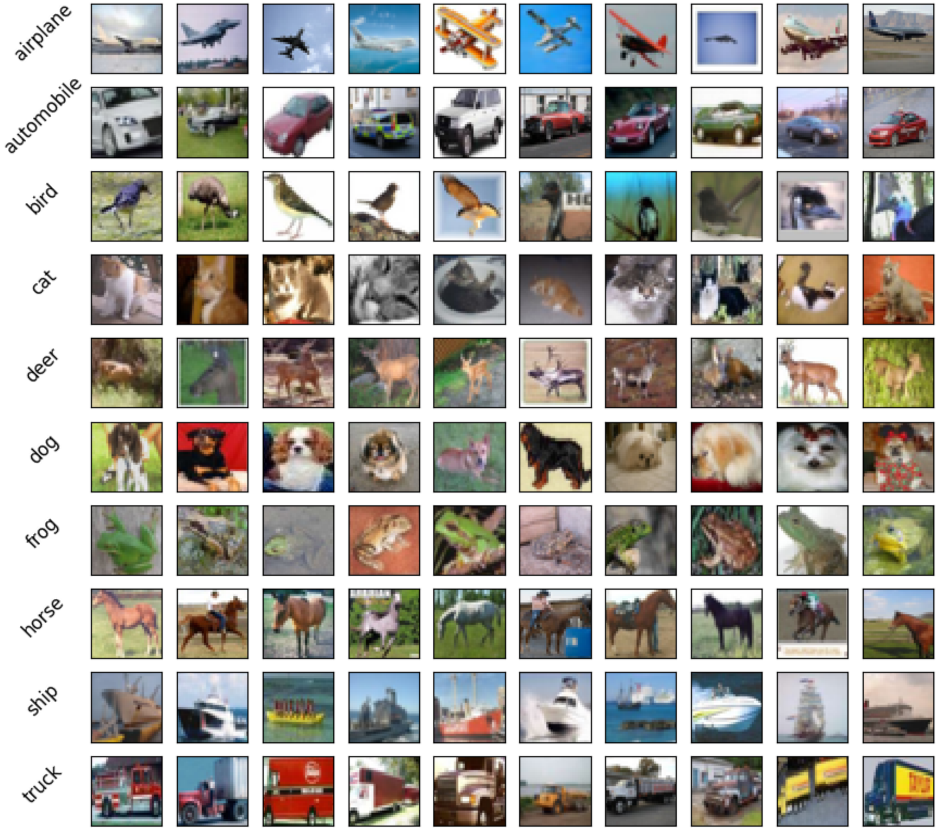
CoE197M/EE298M (Foundations of Machine Learning)

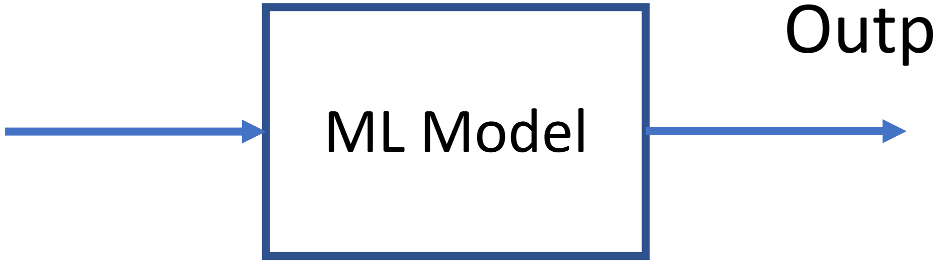
Rowel Atienza, Ph.D.

rowel@eee.upd.edu.ph

Reference: "Mathematics for Machine Learning". Copyright 2020 by Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. Published by Cambridge University Press.

Probability: Language to Describe ML Models





ML Model

Output: $p(\boldsymbol{y}|\boldsymbol{x})$

Parameters : $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$

Input : $p(\boldsymbol{x})$

$(\Omega, \mathcal{A}, P)$: Model of a Real-World Phenomenon

Sample space Ω : set of all possible outcomes

Event space $\mathcal{A}$: set of potential results of the experiment

Probability P : an event $A \in \mathcal{A}$, probability assigns a number that measures the likelihood of its occurrence $P(A) \in [0., 1.]$

$P(\Omega) = 1.$

Random Variable (RV)

Target space $\mathcal{T} \subseteq \Omega$: set of outcomes or states that we are interested in

Random variable $X : \Omega \to \mathcal{T}$: mapping or association from Ω to $\mathcal{T}$

For example, suppose in a roll of 2 dice:

$\Omega = \{(x_1, x_2) \ni x_1, x_2 \in [1,6]\}$

Interested in the sum of the results: $\mathcal{T} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

$X = x_1 + x_2$

$P(X = 2) = \frac{1}{36}$

Discrete Random Variable X

$\mathcal{T}$ is discrete

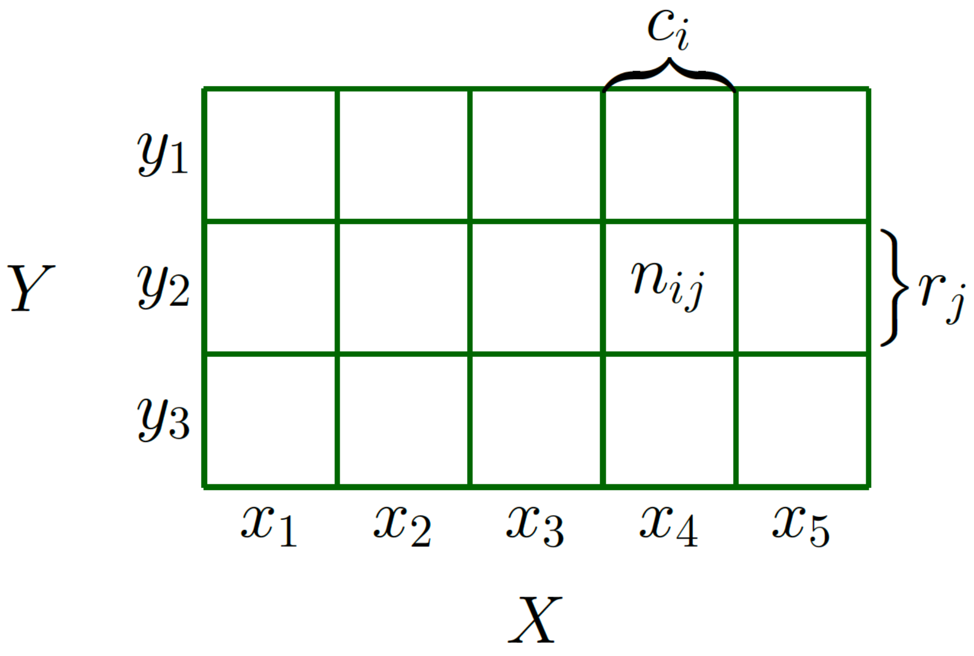
$\forall x \in \mathcal{T}, P(X = x)$ is the probability that X has a value x

$P(X = x)$ is also known as probability mass function (*pmf*)

Example (2 dice): $P(X = 10) = \frac{3}{36}$

$\forall x_i \in \mathcal{T}, \sum_i P(X = x_i) = 1.$

Discrete Random Variables X and Y



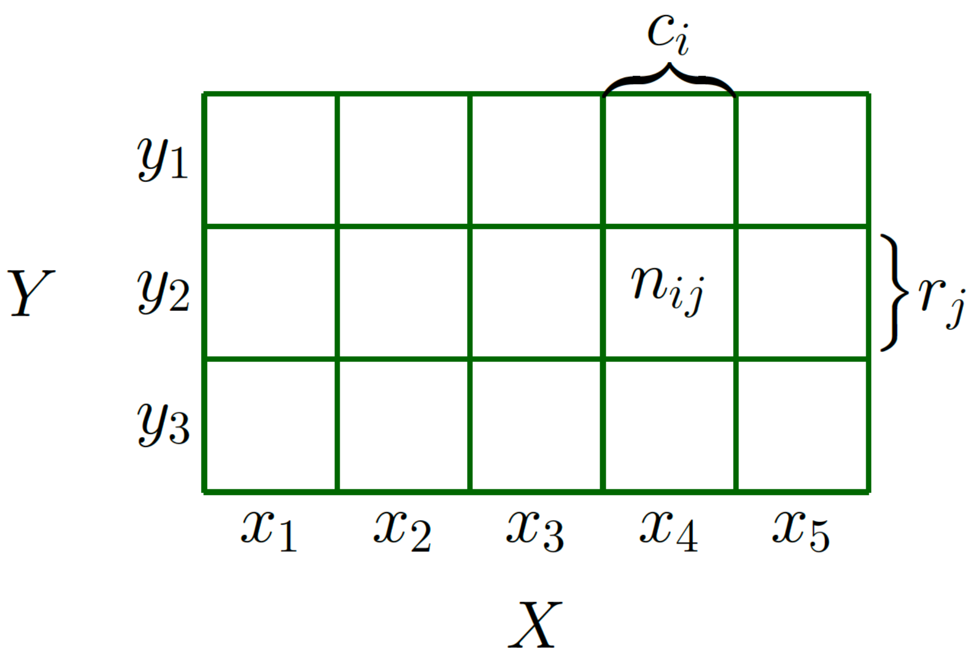
$$P(X = x_i, Y = y_i) = \frac{n_{ij}}{N}$$

n_{ij}: number of events with state x_i and y_j

N: total number of events

Joint probability:
$P(X = x_i, Y = y_i) =$
$P(X = x_i \cap Y = y_i) = p(x, y)$

Discrete Random Variables X and Y



$p(x)$: probability of x irrespective of y (marginal)

$p(y)$: probability of y irrespective of x (marginal)

$p(y|x)$: probability of y given that we know x

$p(x|y)$: probability of x given that we know y

Example

Continuous Random Variables

$\mathcal{T}$ is continuous.

$\forall x \in \mathcal{T}, P(X \leq x)$ is the probability that X has a value less than or equal to x.

$P(X \leq x)$ is also known as cumulative distribution function (*cdf*).

$P(X = x) = 0.$

$$P(a \leq X \leq b)$$
$$= P(X \leq b) - P(X \leq a)$$
$$\ni a, b \in \mathbb{R}\ and\ a < b$$

Continuous Random Variables

Definition: A function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ is a Probability Density Function (*pdf*) if

1. $\forall \boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) \geq 0$
2. $\int_{\mathbb{R}^D} f(\boldsymbol{x})\, d\boldsymbol{x} = 1.$

$$\therefore P(a \leq X \leq b) = \int_a^b f(\boldsymbol{x})\, d\boldsymbol{x} \leq 1.$$

Continuous Random Variables

Definition: The cumulative distribution function (*cdf*) of a multi-variate random variable X with state $\boldsymbol{x} \in \mathbb{R}^D$:

$$F_X(\boldsymbol{x}) = P(X_1 \leq x_1, \dots, X_D \leq x_D)$$

$$\therefore F_X(\boldsymbol{x}) == \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_D} f(x_1, \dots, x_D)\, dx_1 \dots dx_D \leq 1.$$

Example

Sum Rule (Marginalization)

$$p(\boldsymbol{x}) = \begin{cases} \displaystyle\sum_{\boldsymbol{y}\in\mathcal{Y}} p(\boldsymbol{x},\boldsymbol{y}) & discrete \\ \displaystyle\int_{\boldsymbol{y}\in\mathcal{Y}} p(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{y} & continuous \end{cases}$$

$\mathcal{Y}$ is the target space of random variable Y

Product Rule (Factorization)

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})$$

Bayes Theorem

$$\underbrace{p(\boldsymbol{x}|\boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y}|\boldsymbol{x})}^{\text{likelihood}}\overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

Bayes Theorem (Inverse)

$$\underbrace{p(\boldsymbol{y}|\boldsymbol{x})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{x}|\boldsymbol{y})}^{\text{likelihood}}\overbrace{p(\boldsymbol{y})}^{\text{prior}}}{\underbrace{p(\boldsymbol{x})}_{\text{evidence}}}$$

Summary Statistics

Expectation

Expected Value: The expected value of $g(x) : \mathbb{R} \to \mathbb{R}$, where $X \sim p(x)$:

$$\mathbb{E}_X[g(x)] = \begin{cases} \int_{\mathcal{X}} g(x)p(x)\, dx & continous \\ \sum_{x \in \mathcal{X}} g(x)p(x) & discrete \end{cases}$$

Expectation

For a multivariate random variable X,

$$\mathbb{E}_X[g(\boldsymbol{x})] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{bmatrix} \in \mathbb{R}^D$$

Mean

For the special case $g(x) = x$:

$$\mathbb{E}_X[x] = \begin{cases} \int_{\mathcal{X}} xp(x)\,dx & continous \\ \sum_{x \in \mathcal{X}} xp(x) & discrete \end{cases}$$

Mean

For a multivariate random variable X,

$$\mathbb{E}_X[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D$$

$$\mathbb{E}_{X_d}[x_d] = \begin{cases} \int_{\mathcal{X}} x_d p(x_d)\, dx_d & continous \\ \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) & discrete \end{cases}$$

Linearity of Expectation

For $f(\boldsymbol{x}) = \alpha g(\boldsymbol{x}) + \beta h(\boldsymbol{x})$:

$$\mathbb{E}_X[f(\boldsymbol{x})] = \alpha \mathbb{E}_X[g(\boldsymbol{x})] + \beta \mathbb{E}_X[h(\boldsymbol{x})]$$

Covariance

$$Cov_{X,Y}[x,y] := \mathbb{E}_{X,Y}[(x-\mathbb{E}_X[x])(y-\mathbb{E}_Y[y])]$$

$$Cov_{X,Y}[x,y] := \mathbb{E}_{X,Y}[xy]-\mathbb{E}_X[x]\mathbb{E}_Y[y]$$

Variance and Standard Deviation

Variance:

$$\mathbb{V}_x[x] = Cov_{X,X}[x, x] := \mathbb{E}_{X,X}[x^2] - (\mathbb{E}_X[x])^2$$

Standard Deviation:

$$\sigma(x) = \sqrt{\mathbb{V}_x[x]}$$

Covariance

$X\,:\,\boldsymbol{x} \in \mathbb{R}^D$ and $Y\,:\,\boldsymbol{y} \in \mathbb{R}^E$

$$Cov[\boldsymbol{x},\boldsymbol{y}] := \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}]^T = Cov[\boldsymbol{y},\boldsymbol{x}]^T \in \mathbb{R}^{D\times E}$$

$Cov[\boldsymbol{x},\boldsymbol{y}]$ is symmetric, positive definite and linear

Variance

$X \,:\, \boldsymbol{x} \in \mathbb{R}^D$ with mean $\boldsymbol{\mu} \in \mathbb{R}^D$

$$\mathbb{V}_x[\boldsymbol{x}] = Cov_X[\boldsymbol{x}, \boldsymbol{x}] := \mathbb{E}_X\ [(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T] = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T$$

$$\mathbb{V}_x[\boldsymbol{x}] = \begin{bmatrix} Cov_X[x_1, x_1\ ] & \cdots & Cov_X[x_1, x_D\ ] \\ \vdots & \ddots & \vdots \\ Cov_X[x_D, x_1\ ] & \cdots & Cov_X[x_D, x_D\ ] \end{bmatrix}$$

$\mathbb{V}_x[\boldsymbol{x}]$: Covariance matrix of multi-variable random variable X

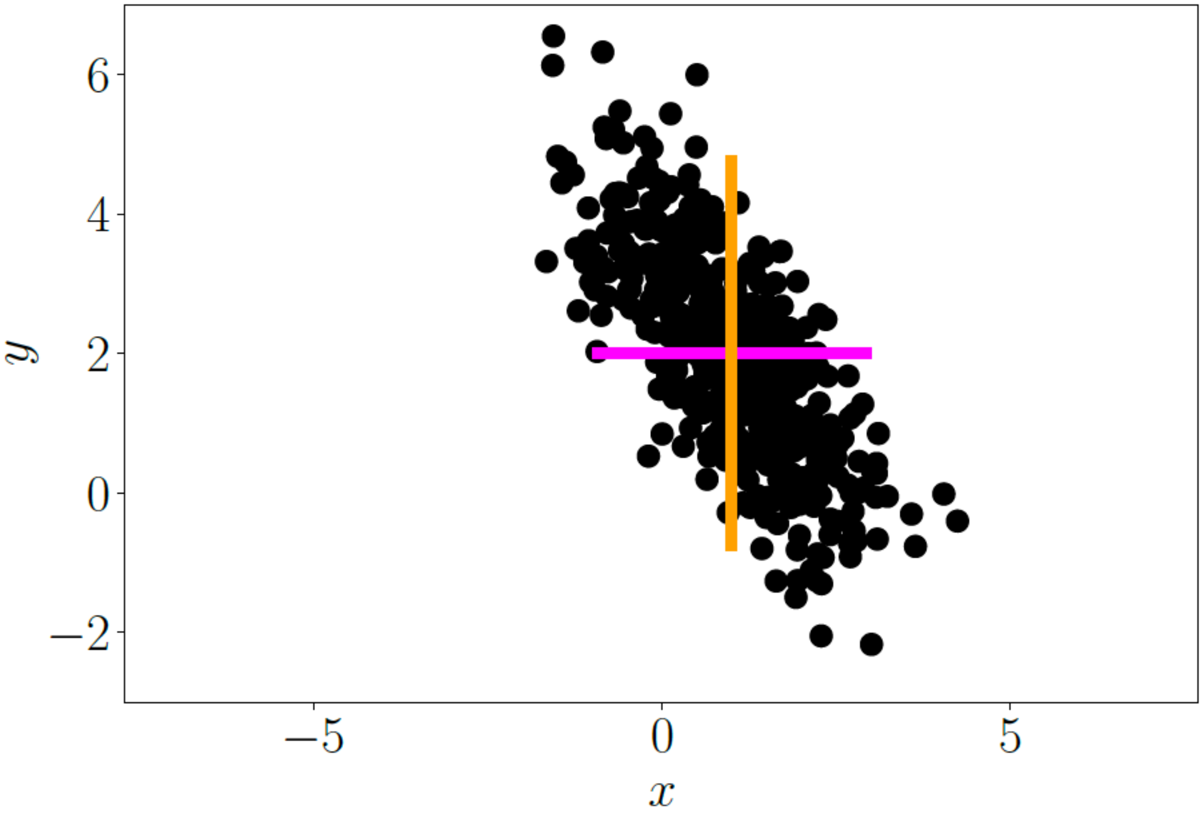
Correlation

Normalized covariance:

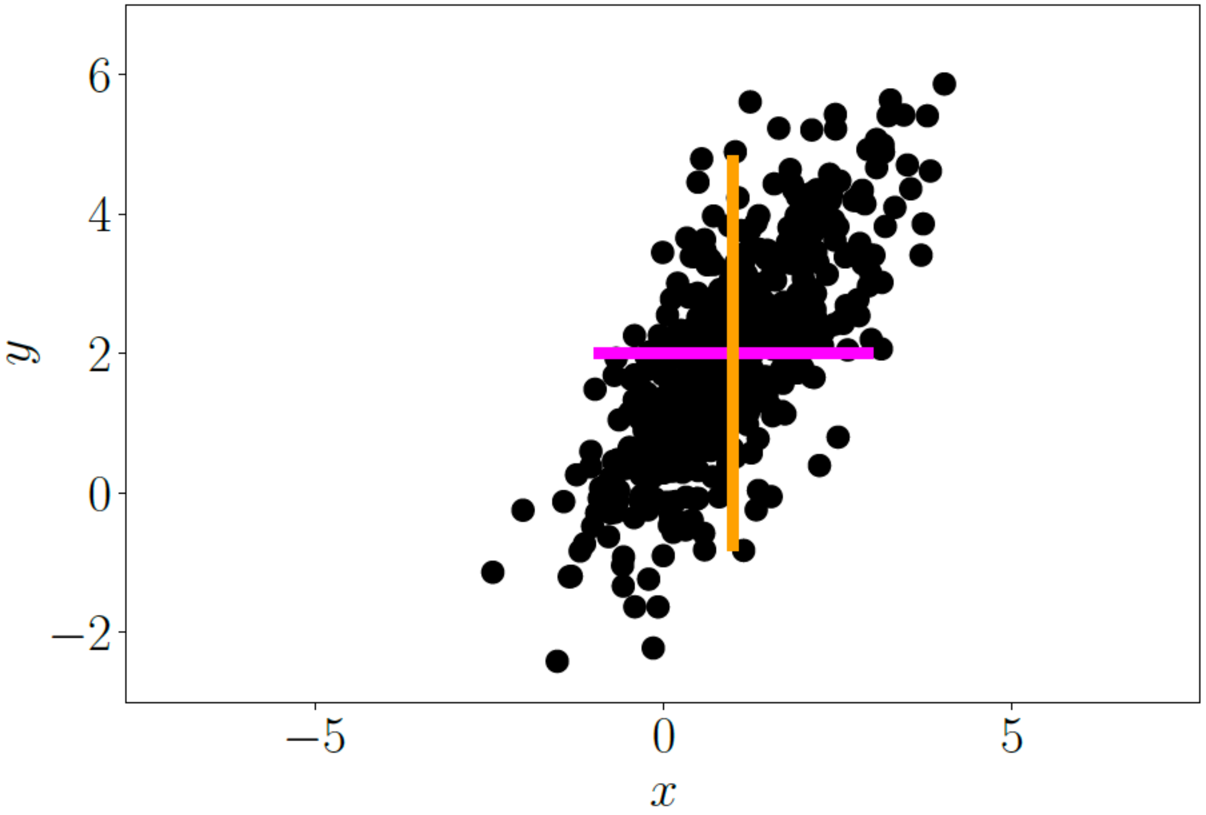
$$corr[x, y] = \frac{Cov_{X,Y}[x, y]}{\sqrt{\mathbb{V}_x[x]\mathbb{V}_y[y]}} \in [-1., 1.]$$

MML Book 2020

Interpretation



(a) x and y are negatively correlated.



(b) x and y are positively correlated.

Empirical Mean

In ML, we have a finite dataset of size N

For each observation or sample $\boldsymbol{x}_n \in \mathbb{R}^D$:

$$\overline{\boldsymbol{x}} := \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n$$

Empirical Covariance

For each observation or sample $\boldsymbol{x}_n \in \mathbb{R}^D$:

$$\Sigma := \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \overline{\boldsymbol{x}})(\boldsymbol{x}_n - \overline{\boldsymbol{x}})^T$$

Statistical Independence

Statistical Independence

2 random variables X and Y are statistically independent if and only if:

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$$

Properties of Statistical Independence

$$p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y})$$

$$p(\boldsymbol{x}|\boldsymbol{y}) = p(\boldsymbol{x})$$

$$\mathbb{V}_{X,Y}[\boldsymbol{x}+\boldsymbol{y}] = \mathbb{V}_{X}[\boldsymbol{x}] + \mathbb{V}_{Y}[\boldsymbol{y}]$$

$$Cov[\boldsymbol{x},\boldsymbol{y}] = \boldsymbol{0}$$

i.i.d.

Independent and Identical Distributed random variables $X_1, X_2, \ldots, X_N$

Independent

Identically Distributed (from the same distribution)

Conditional Independence

Two random variables X and Y are conditionally independent if and only if:

$$p(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{z}) = p(\boldsymbol{x} | \boldsymbol{z}) p(\boldsymbol{y} | \boldsymbol{z})$$

$\forall \boldsymbol{z} \in \mathcal{Z}, \mathcal{Z}$ is a state of random variable Z

Also written as $X \perp Y | Z$.

Independence is a special case: $X \perp Y | \emptyset$

Note that by product rule:

$$p(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{z}) = p(\boldsymbol{x} | \boldsymbol{y}, \boldsymbol{z}) p(\boldsymbol{y} | \boldsymbol{z})$$

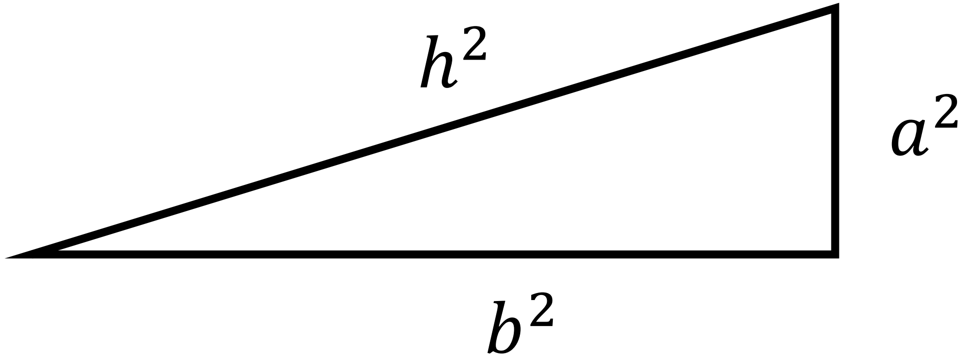
$\therefore p(\boldsymbol{x} | \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x} | \boldsymbol{z})$ if $X \perp Y | Z$

Inner Products of Random Variables

Inner Product of Uncorrelated Random Variables X and Y

$$\mathbb{V}[x + y] = \mathbb{V}[x] + \mathbb{V}[y]$$

Similar to Pythagorean Theorem: $h^2 = a^2 + b^2$



Inner Product of Random Variables X and Y

$$\langle X, Y \rangle := Cov[x, y]$$

Inner product if X and Y have zero mean

Length of a random variable:

$$\|X\| = \sqrt{Cov[x, x]} = \sqrt{\mathbb{V}[x]} = \sigma(x)$$

Or the standard deviation (e.g zero means deterministic)

Angle Between Random Variables X and Y

$$\cos\theta = \frac{\langle X, Y\rangle}{\|X\|\|Y\|} = \frac{Cov[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} = Corr[x, y]$$

Gaussian Distribution

Gaussian or Normal Distribution

To model likelihood and prior in linear regression

Used in Gaussian Mixture Models (GMM)

Used in Reinforcement Learning

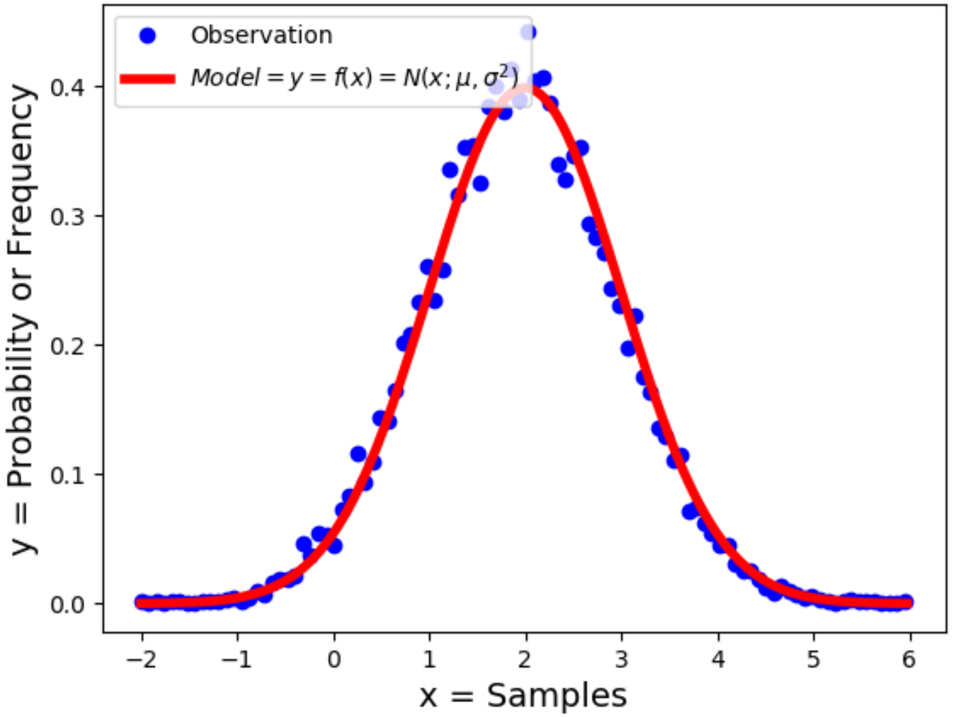
Initialization of weights in deep neural networks

Univariate Gaussian

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ is mean

σ is standard deviation



Multivariate Gaussian

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}}$$

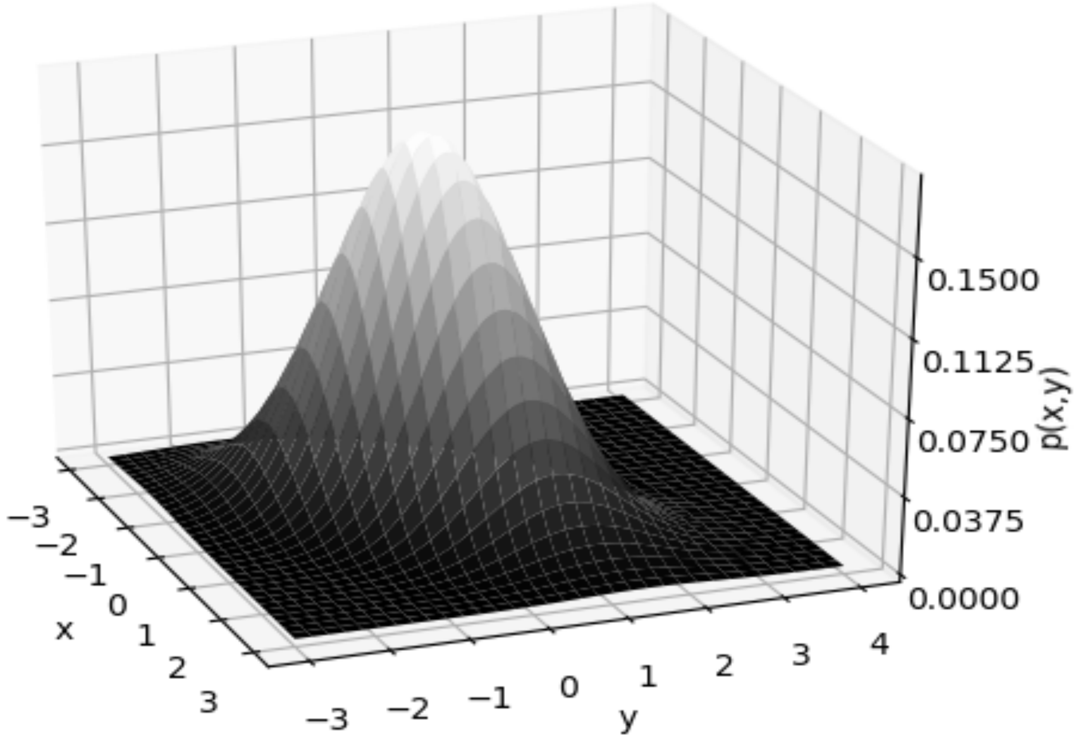
$\boldsymbol{\mu}$ is mean vector

$\boldsymbol{\Sigma}$ is covariance matrix
$\boldsymbol{x} \in \mathbb{R}^D$

$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is also written as $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$

or $X \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$

2D Gaussian



$$\boldsymbol{\mu} = [0 \quad 0]^T$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Conditional Gaussian

For states $[\boldsymbol{x}^T \quad \boldsymbol{y}^T]$:

$$p(\boldsymbol{x},\boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right)$$

$\boldsymbol{\Sigma}_{xx} = Cov[\boldsymbol{x},\boldsymbol{x}], \boldsymbol{\Sigma}_{\boldsymbol{yy}} = Cov[\boldsymbol{y},\boldsymbol{y}]$: marginal covariance

$\boldsymbol{\Sigma}_{x\boldsymbol{y}} = Cov[\boldsymbol{x},\boldsymbol{y}]$: cross covariance

Conditional Gaussian

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{\mu}_{x|y}\middle|\boldsymbol{\Sigma}_{x|y}\right)$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_y\right)$$

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$$

Marginal Gaussians

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y} = \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$$

Product of Gaussians

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{A}\ )$$
$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{b}, \boldsymbol{B}\ )$$

$$p(\boldsymbol{x})p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{a}, \boldsymbol{A}\ )\mathcal{N}(\boldsymbol{x}|\boldsymbol{b}, \boldsymbol{B}\ ) = c\mathcal{N}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{C}\ )$$

$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1}$$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b})$$

$$c = (2\pi)^{-\frac{D}{2}}|\boldsymbol{A} + \boldsymbol{B}|^{-\frac{1}{2}}e^{-\frac{1}{2}(\boldsymbol{a}-\boldsymbol{b})^T(\boldsymbol{A}+\boldsymbol{B})^{-1}(\boldsymbol{a}-\boldsymbol{b})}$$

Sum of Gaussians

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$$
$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y}\middle|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$$

$$p(\boldsymbol{x}+\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\right)$$

Note that:

$$p(\boldsymbol{x}+\boldsymbol{y}) \neq p(\boldsymbol{x}) + p(\boldsymbol{y})$$

Mixture of 2 Univariate Gaussians

$$p(x) = \alpha p_1(x) + (1-\alpha)p_2(x)$$

$\ni 0 \leq \alpha \leq 1, \mu_1 \neq \mu_2$ and $\sigma_1 \neq \sigma_2$

$$\mathbb{E}[x] = \alpha\mu_1 + (1-\alpha)\mu_2$$

$$\mathbb{V}[x]$$
$$= (\alpha\sigma_1^2 + (1-\alpha)\sigma_1^2) + \left((\alpha\mu_1^2 + (1-\alpha)\mu_1^2) - (\alpha\mu_1 + (1-\alpha)\mu_2)^2\right)$$

Linear Transformation of Gaussians

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{\mu}$$

$$\mathbb{V}[\boldsymbol{y}] = \mathbb{V}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}[\boldsymbol{x}]\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$$

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$$

Reverse Form

Other Distributions

Bernoulli

$x \in \{0,1\}, \mu \in [0., 1.]$

$$p(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

$$\mathbb{E}[x] = \mu$$

$$\mathbb{V}[x] = \mu(1 - \mu)$$

For example, the probability of getting a head, $X = 1$ in a coin-flip experiment

Binomial

m occurrences of $X = 1$ in N samples from Bernoulli with $p(X = 1) = \mu \in [0., 1.]$:

$$Bin(N, \mu) = p(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$\mathbb{E}[m] = N\mu$$

$$\mathbb{V}[m] = N\mu(1 - \mu)$$

For example, the probability of observing m heads in N coin-flip experiments

Beta

Continuous over $\mu \in [0., 1.]$.

$\alpha > 0$ and $\beta > 0$:

$Beta(\alpha, \beta) = p(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$

$\mathbb{E}[\mu] = \frac{\alpha}{\alpha+\beta}$, $\mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

$\Gamma(t) := \int_0^\infty x^{t-1}e^{-x}\,dx$, for $t > 0$
$\Gamma(t+1) = t\Gamma(t)$

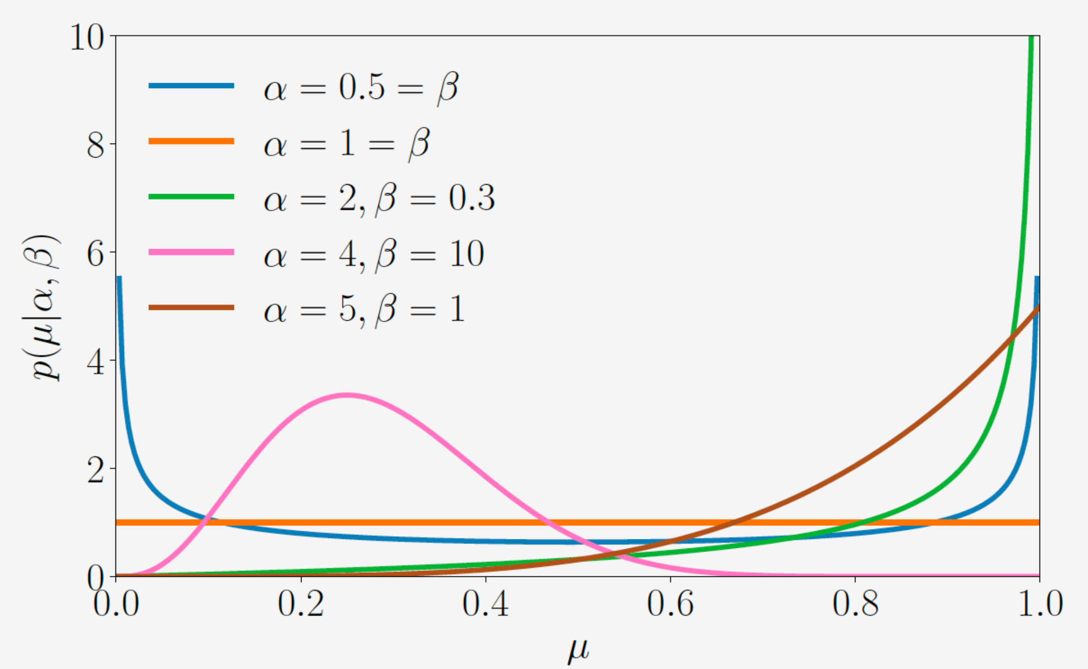
Beta

$\alpha = \beta = 1$: uniform distribution

$\alpha < 1$ and $\beta < 1$: bi-modal spikes

$\alpha > 1$ and $\beta > 1$: unimodal

$\alpha > 1, \beta > 1, \alpha = \beta$, the distribution is unimodal, symmetric, centered in $[0.\,, 1.\,]$ with mean and mode of $\frac{1}{2}$



Recall Bayes Theorem

$$\underbrace{p(\boldsymbol{y}|\boldsymbol{x})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{x}|\boldsymbol{y})}^{\text{likelihood}}\overbrace{p(\boldsymbol{y})}^{\text{prior}}}{\underbrace{p(\boldsymbol{x})}_{\text{evidence}}}$$

Observations

Posterior is directly proportional to product of likelihood and prior

Prior encapsulates our knowledge about the problem before seeing the data

Evidence provides normalization of the posterior

Posterior is hard to compute analytically

Conjugate Priors

A prior is a conjugate for the likelihood function if the posterior is of the same form/type as prior

Conjugacy allows us to algebraically compute the posterior from prior

Beta-Bernoulli Conjugacy

Likelihood: $x \in \{0,1\}$ be a Bernoulli distribution with parameter $\theta \in [0.,1.]$:

$$p(x|\theta) = \theta^{x}(1-\theta)^{1-x}$$

Prior: parameter θ has a Beta distribution:

$$p(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Posterior:

$$p(\theta|x) \propto p(x|\theta)p(\theta|\alpha,\beta) = \theta^{(x+\alpha)-1}(1-\theta)^{\beta+(x+1)-1}$$

Beta distribution: $p(\theta|x) \propto p(\theta|(x+\alpha),\beta+(x+1))$

Conjugate Priors

Likelihood	Conjugate Prior	Posterior
Bernoulli	Beta	Beta
Binomial	Beta	Beta
Gaussian	Gaussian	Gaussian
Multinomial	Dirichlet	Dirichlet

Sufficient Statistics

Contain all available information about data and distribution it represents

Theorem (Fisher-Neyman): Let X have the probability function $p(x|\theta)$. The statistics $\phi(x)$ are sufficient for θ if and only if $p(x|\theta)$ can be written in the form:

$$p(x|\theta) = h(x)g_{\theta}\big(\phi(x)\big)$$

$h(x)$ is a distribution independent of θ and g_{θ} captures all the dependencies on θ via sufficient statistics $\phi(x)$

Exponential Family

An exponential family is family of probability distributions parameterized by $\boldsymbol{\theta} \in \mathbb{R}^D$:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x})e^{\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\right)}$$

Where $\phi(\boldsymbol{x})$ is a vector of sufficient statistics

$\boldsymbol{\theta}$ are natural parameters

$A(\boldsymbol{\theta})$ ensures that $\int p(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} = 1$

Exponential Family

Intuitively,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto e^{\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x})\right)}$$

Example

Univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$:

$$\phi(\boldsymbol{x}) = \begin{bmatrix} x \\ x^{\mathbf{2}} \end{bmatrix}$$

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto e^{\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x})\right)} = e^{\left([\theta_1 \quad \theta_2]\begin{bmatrix} x \\ x^{\mathbf{2}} \end{bmatrix}\right)} = e^{(\theta_1 x + \theta_2 x^{\mathbf{2}})}$$

$$\boldsymbol{\theta}^T = \begin{bmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{bmatrix}$$

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Every member of exponential family has a conjugate prior:

$$p(\boldsymbol{x}|\gamma) = h_c(\boldsymbol{x})e^{\left(\left\langle \begin{bmatrix}\gamma_1\\ \gamma_1\end{bmatrix}, \begin{bmatrix}\boldsymbol{\theta}\\ -A(\boldsymbol{\theta})\end{bmatrix}\right\rangle - A_c(\gamma)\right)}$$

Change of Variables

Transformation of Distribution

Problem: Given $p(\boldsymbol{x})$, what is $p(\boldsymbol{y})$ if $\boldsymbol{y} = U(\boldsymbol{x})$?

$U(\boldsymbol{x})$ is an invertible function

$$p(\boldsymbol{y}) = p\left(\boldsymbol{x} = U^{-1}(y)\right)$$

Change of Variables

Fundamental equation change of variables:

$$\int f\big(g(x)\big)g'(x)dx = \int f(u)\,du$$

$$du = g'(x)dx$$

Assume:

- A random variable X
- A random variable Y
- An invertible function U such that $Y = U(X)$

By definition of cdf:

$$F_Y(y) = P(Y \leq y) = P(U(X) \leq y)$$

If U is strictly increasing/decreasing, then U^{-1} is also strictly increasing/decreasing. Therefore:

$$P(Y \leq y) = P\left(U^{-1}\big(U(X)\big) \leq U^{-1}(y)\right) = P\big(X \leq U^{-1}(y)\big)$$

Using the definition of cdf:

$$F_Y(y) = P\left(X \leq U^{-1}(y)\right) = \int_{a}^{U^{-1}(y)} f(x)\,dx$$

$$f(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} \int_{a}^{U^{-1}(y)} f(x)\,dx$$

$$f(y) = \frac{d}{dy}\int_{a}^{U^{-1}(y)} f(x)\,dx = \frac{d}{dy}\int_{a}^{U^{-1}(y)} f_x\big(U^{-1}(y)\big)U^{-1\prime}(y)\,dy$$

$$f(y) = f_x\big(U^{-1}(y)\big)\frac{d}{dy}U^{-1}(y)$$

To represent both strictly increasing and strictly decreasing $U(\cdot)$

$$f(y) = f_x\left(U^{-1}(y)\right)\left|\frac{d}{dy}U^{-1}(y)\right|$$

$\left|\frac{d}{dy}U^{-1}(y)\right|$ measures how much a unit volume changes

Multivariate Distribution Transformation

$f(\boldsymbol{x})$ is the probability distribution of a continuous random variable X

Vector valued function $\boldsymbol{y} = U(\boldsymbol{x})$ is differentiable and invertible for all values of the domain $\boldsymbol{x}$, then for the corresponding values of $\boldsymbol{y}$, the probability density of $Y = U(X)$:

$$f(\boldsymbol{y}) = f_x\left(U^{-1}(\boldsymbol{y})\right)\left|det\left(\frac{\partial}{\partial \boldsymbol{y}} U^{-1}(\boldsymbol{y})\right)\right|$$