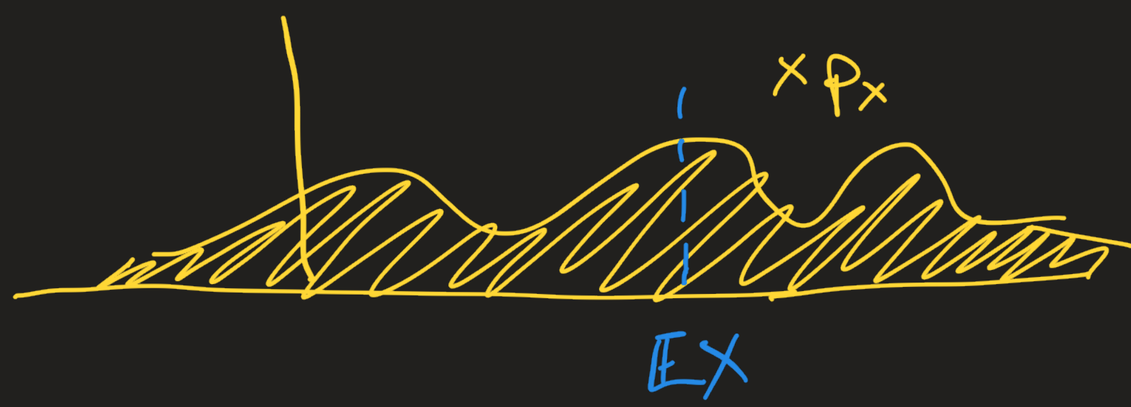
Mean & variance

Two most used quantities to summarize behavior of r.v.s.

Expectation / mean / average

$$\mu = \mathbb{E}X = \int p(x)\, x \, dx$$

$$\mathbb{E}X = \sum_{x \in \Omega_X} x p(x)$$



weighted average: what value we expect to see on averge

think of this as a center of mass

called the first moment of X

Ex $X \sim \text{Bern}(p) \Rightarrow \mathbb{E}X = 0 \cdot (1-p) + 1 \cdot p = p$

$X \sim \text{Bin}(n,p) \Rightarrow \mathbb{E}X = \sum_{i=0}^{n} i \cdot p(i)$

$p(i) = \binom{n}{i} p^i (1-p)^{n-i}$

$= \sum_{i=1}^{n} i \binom{n}{i} p^i (1-p)^{n-i}$

$= \cdots = np$

$X \sim \mathcal{N}(\mu, \sigma^2)$ $\quad \mathbb{E}X = \int_{\mathbb{R}} p_X(x) \, x \, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot x \, dx$

$= \cdots = \mu$

Key property of $\mathbb{E}$: linearity

if $c \in \mathbb{R}$ and X & Y are r.v.s. then

$$\mathbb{E}[cX + Y] = c\mathbb{E}X + \mathbb{E}Y$$

regardless of any relationship b/w X & Y

always true

Ex $\varphi = X_1 + \cdots + X_n$ where $X_i \sim \text{Bern}(p)$

then $\varphi \sim \text{Bin}(n, p)$

$\Rightarrow \mathbb{E}\varphi = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = np$

Variance: how good of an estimate of the typical behavior is the mean of the r.v.?

— on average, how far is the r.v. from its expectation?

$$\boxed{\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2}$$

$$= \mathbb{E}\left(X^2 - 2X\,\mathbb{E}X + (\mathbb{E}X)^2\right)$$

$$= \mathbb{E}(X^2) - 2\,\mathbb{E}(X \cdot \mathbb{E}X) + \mathbb{E}\left((\mathbb{E}X)^2\right)$$

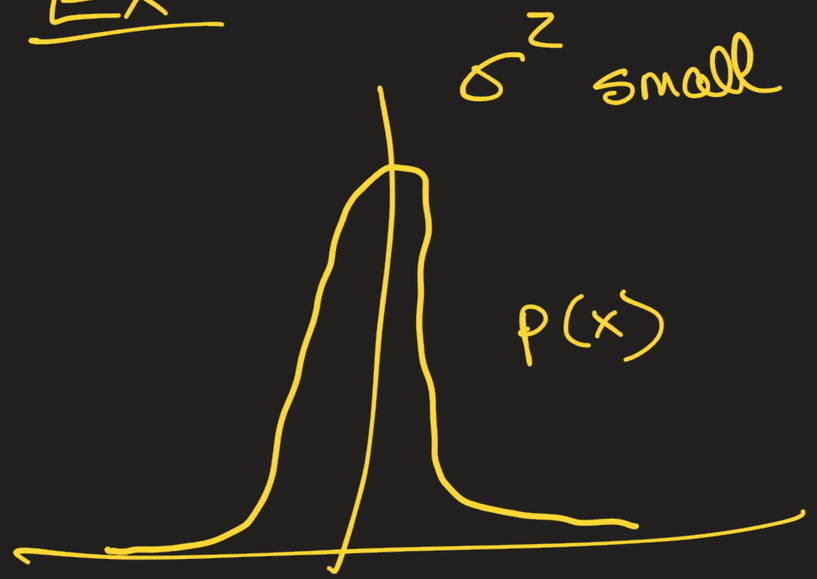
($\mathbb{E}X$: constant; $(\mathbb{E}X)^2$: constant)

$$= \mathbb{E}(X^2) - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2\,\mathbb{E}1$$

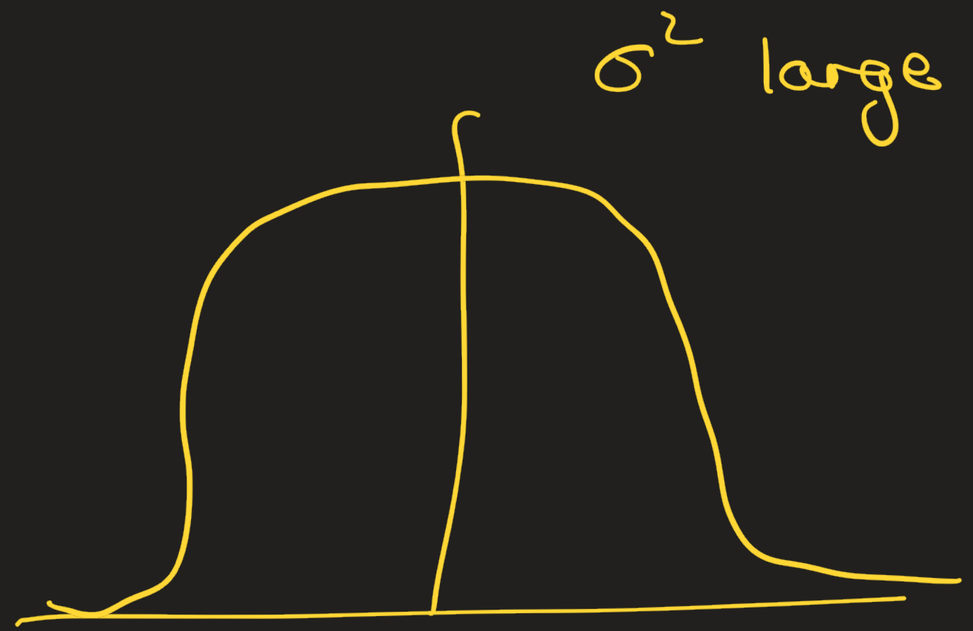
$$\sigma^2 = \boxed{\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2}$$

second moment — first moment

Ex



$\mathbb{E}X = 0$



$\mathbb{E}X = 0$

Ex $X \sim \text{Bern}(p)$

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = (0-p)^2 \cdot (1-p)$$
$$+ (1-p)^2 p = (1-p) \cdot p$$

Interpretation: when p is close to 0 or to 1, the Bern(p) value stays close to its mean p w/ high probability.

Ex $X \sim \mathcal{N}(\mu, \sigma^2)$

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

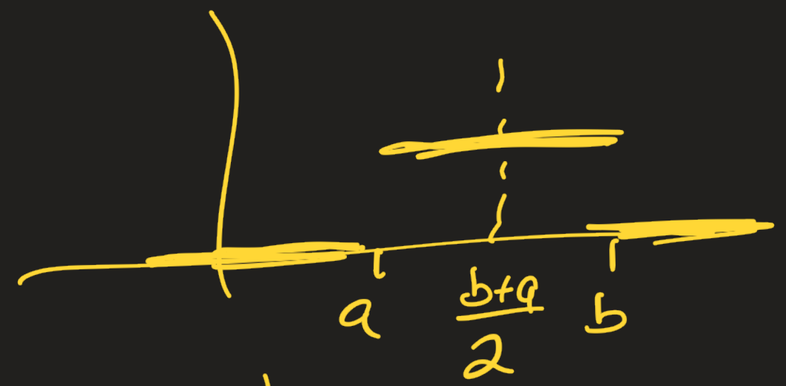
$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 p_X(x) = \cdots = \sigma^2$$

Clear from the form of $p_X(x)$ that when x is slightly moved away from μ:

- if σ^2 is small, then you get a large change in $p_X(\mu+\Delta)$
- if σ^2 is large, then you don't get a large change in $p_X(\mu+\Delta)$

Ex $X \sim U[a,b]$

$$p_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{elsewhere} \end{cases}$$



$$\mathbb{E}X = \int_{\mathbb{R}} x\,p(x)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{x^2}{2}\Big|_a^b \frac{1}{b-a}$$

$$= \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

$$\mathrm{Var}(X) = \mathbb{E}(X-\mathbb{E}X)^2 = \int_a^b \left(x - \left(\frac{b+a}{2}\right)^2 \frac{1}{b-a}\,dx\right.$$

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \int_a^b \frac{x^2}{b-a}\,dx - \left(\frac{b+a}{2}\right)^2$$

$$= \dots = \frac{1}{12}(b-a)^2$$

In ML, we typically model our problems as one of determining a relationship b/w two random quantities:

$X \in \mathbb{R}^d$ features

$y \in \mathbb{R}^k$ targets

e.g. if $k=1$ and y takes values in $\{\pm 1\}$ we get classification.

We assume that there exists a function f so that

$$y \approx f(X)$$

and we try to find f.

Key points

- We need to understand rlns b/w two or more r.v.s.
- We need to understand functions of r.v.s

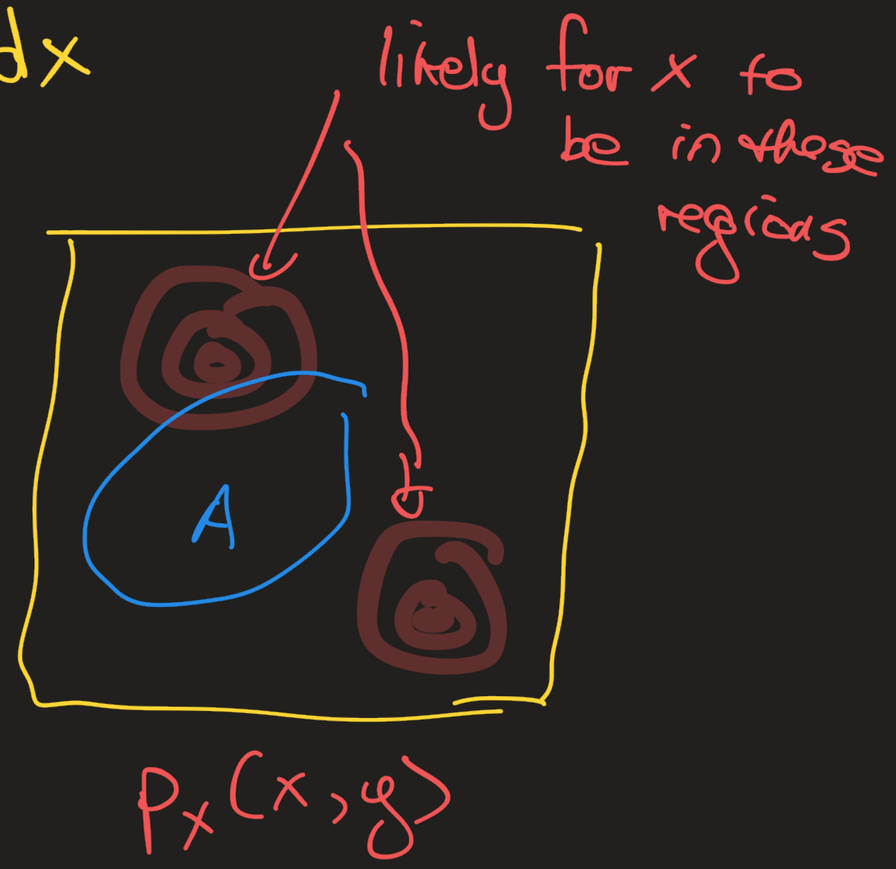
Multivariate r.v.s (random vectors)

A random vector X is a collection of random variables

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

We say X has pdf $p_X(x_1, \dots, x_d)$ if $p_X \geq 0$ and for any $A \in \mathbb{R}^d$

$$P_X(A) = \int_A p_X(x_1, \dots, x_d)\, dx$$



Similarly if x is a discrete r.vector, then p_x is a pmf:

$$1 \geq p_x(x_1, \dots, x_d) \geq 0 \quad \text{everywhere}$$

$$\text{and} \quad \sum p_x = 1$$

Each entry of x has its own <u>marginal distribution</u> that's obtained by integrating out the other coordinates

e.g.

$$p_{x_i}(x) = \int p_x(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_d) \, dx_1 \cdots dx_{i-1} \, dx_{i+1} \cdots dx_d$$

This is a <u>pdf</u>

$$p_{x_i}(z) = P_x\left(\{x \mid x_i = z\}\right)$$

$$= \int_{\{x \mid x_i = z\}} p_x(x)\, dx$$

where $x_i = z$

$$= \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} \cdots \int_{x_{i-1}=-\infty}^{\infty} \int_{x_{i+1}=-\infty}^{\infty} \cdots \int_{x_d=-\infty}^{\infty} p_x(x)\, dx$$

$$= \int p_x(x_1, x_2, \cdots, x_{i-1}, z, x_{i+1}, \cdots, x_d)\, dx_1 dx_2 \cdots dx_d$$

Ex Imagine x_1 & x_2 both take values $\{1,2,3,4\}$ with joint pmf

x \ y	1	2	3	4
1	1/40	1/10	1/10	1/40
2	1/10	1/40	1/40	1/10
3	1/10	1/40	1/40	1/10
4	1/40	1/10	1/10	1/40

x
5/20
5/20
5/20
5/20

y				
	5/20	5/20	5/20	5/20

Compute $\mathbb{E}Z$ where $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$

$$\mathbb{E}Z = \sum_{(x,y) \in \mathcal{X}_Z} p_Z(x,y) \begin{bmatrix} x \\ y \end{bmatrix}$$

$$= \sum_{x \in \{1,2,3,4\}} \sum_{y \in \{1,2,3,4\}} p_Z(x,y) \begin{bmatrix} x \\ y \end{bmatrix}$$

$$= \underbrace{\sum_{y \in \{1,2,3,4\}} p_Z(1,y) \begin{bmatrix} 1 \\ y \end{bmatrix}}_{x=1} + \cdots + \underbrace{\sum_{y \in \{1,2,3,4\}} p_Z(4,y) \begin{bmatrix} 4 \\ y \end{bmatrix}}_{x=4}$$

$$= \underbrace{p_Z(1,1) \begin{bmatrix} 1 \\ 1 \end{bmatrix} + p_Z(1,2) \begin{bmatrix} 1 \\ 2 \end{bmatrix} + p_Z(1,3) \begin{bmatrix} 1 \\ 3 \end{bmatrix} + p_Z(1,4) \begin{bmatrix} 1 \\ 4 \end{bmatrix}}_{x=1} + \cdots$$

$$= \underbrace{\frac{1}{40}\begin{bmatrix}1\\1\end{bmatrix} + \frac{1}{10}\begin{bmatrix}1\\2\end{bmatrix} + \frac{1}{10}\begin{bmatrix}1\\3\end{bmatrix} + \frac{1}{40}\begin{bmatrix}1\\4\end{bmatrix}}_{x=1} + \cdots$$

$$= \cdots \quad \text{(exercise)} \quad = \begin{bmatrix}5/2\\5/2\end{bmatrix}$$

Note

$$\mathbb{E}X = \mathbb{E}_X x = \frac{5}{20}(1+2+3+4) = \frac{5}{2}$$

$$\mathbb{E}Y = \mathbb{E}_Y y = \frac{5}{20}(1+2+3+4) = \frac{5}{2}$$

We also have joint marginals, e.g.

$$p_{X_i, X_j}(x, y) = \int p_X(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_{j-1}, y, x_{j+1}, \dots, x_d)\, dx_1 \cdots dx_d$$

Can define means & variances of the coordinates using their marginals, e.g.

$$\mathbb{E}_{x_i} x_i = \int p_{x_i}(x)\, x\, dx$$

$$\mathrm{Var}(x_i) = \int (x_i - \mathbb{E}_{x_i} x_i)^2\, p_{x_i}(x)\, dx$$

Mean of a random vector

$$\mathbb{E}X = \int_{\mathbb{R}^d} p_X(x)\, x\, dx = \int_{\mathbb{R}^d} p_X(x) \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} dx$$

$$= \begin{bmatrix} \int_{\mathbb{R}^d} p_X(x)\, x_1\, dx \\ \vdots \\ \int_{\mathbb{R}^d} p_X(x)\, x_d\, dx \end{bmatrix} \overset{\text{HW}}{=} \begin{bmatrix} \mathbb{E}_{X_1} X_1 \\ \vdots \\ \mathbb{E}_{X_d} X_d \end{bmatrix} = \begin{bmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_d \end{bmatrix}$$

Variance of x_i is defined as before

$$\mathrm{Var}(x_i) = \mathbb{E}(x_i - \mathbb{E}x_i)^2 = \mathbb{E}_{x_i}(x_i - \mathbb{E}x_i)^2$$

where $x_i \sim p_{x_i}$ the i-th marginal

But we also have <u>covariance</u> for each pair of coordinates: how likely are x_i & x_j to vary <u>in the same direction</u> around their means

$$\mathrm{Cov}(x_i, x_j) = \mathbb{E}_{x_i, x_j}\left[(x_i - \mathbb{E}x_i)(x_j - \mathbb{E}x_j)\right]$$
$$= \mathbb{E}\left[(x_i - \mathbb{E}x_i)(x_j - \mathbb{E}x_j)\right]$$

when x_i and x_j tend to simultaneously be both larger (or smaller) than their means then $(x_i - \mathbb{E}x_i)(x_j - \mathbb{E}x_j) > 0$

so $\mathrm{Cov}(x_i, x_j)$ is large

- If they have opposite behaviors, then

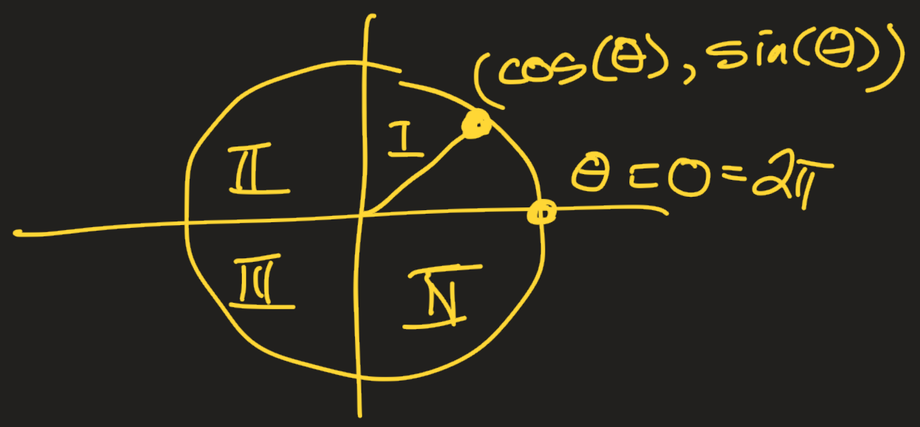
$$(x_i - \mathbb{E}x_i)(x_j - \mathbb{E}x_j) < 0$$

so $Cov(x_i, x_j)$ is large & negative

- If their behaviors aren't really related then

$$Cov(x_i, x_j) = 0$$

Ex let $X = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}$ where $\theta \sim U[0, 2\pi]$



$X_1 - \mathbb{E}X_1$ $X_2 - \mathbb{E}X_2$

I: > >

II: < >

III: < <

IV: > <

To figure out the covariance, we need to compute:

- $\mathbb{E}x_1$, $\mathbb{E}x_2$
- $\mathbb{E}\left[(x_1 - \mathbb{E}x_1)(x_2 - \mathbb{E}x_2)\right] = Cov(x_1, x_2)$

to compute $\mathbb{E}x_1, \mathbb{E}x_2$, use

$$\mathbb{E}X = \int_0^{2\pi} p(\theta) \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} d\theta = \frac{1}{2\pi} \int_0^{2\pi} \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} d\theta$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow Cov(X_1, X_2) = \mathbb{E}X_1X_2 = \frac{1}{2\pi} \int_0^{2\pi} \cos\theta \sin\theta \, d\theta = 0$$

We define a covariance matrix for x, C_x, that contains all the variances for the coordinates as well as all the covariances

$$C_x = \mathbb{E}\left[(x - \mathbb{E}x)(x - \mathbb{E}x)^T\right] \in \mathbb{R}^{d \times d}$$

Ex: if $d = 2$

$$C_x = \mathbb{E}\left(\begin{bmatrix} x_1 - \mathbb{E}x_1 \\ x_2 - \mathbb{E}x_2 \end{bmatrix} \begin{bmatrix} x_1 - \mathbb{E}x_1 & x_2 - \mathbb{E}x_2 \end{bmatrix}\right)$$

$$= \mathbb{E}\left(\begin{bmatrix} (x_1 - \mathbb{E}x_1)^2 & (x_1 - \mathbb{E}x_1)(x_2 - \mathbb{E}x_2) \\ (x_2 - \mathbb{E}x_2)(x_1 - \mathbb{E}x_1) & (x_2 - \mathbb{E}x_2)^2 \end{bmatrix}\right)$$

$$= \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{bmatrix}$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$X - \mathbb{E}X = \begin{bmatrix} x_1 - \mathbb{E}x_1 \\ \vdots \\ x_d - \mathbb{E}x_d \end{bmatrix}$$

$$(X - \mathbb{E}X)(X - \mathbb{E}X)^T = \begin{bmatrix} x_1 - \mathbb{E}x_1 \\ \vdots \\ x_d - \mathbb{E}x_d \end{bmatrix} \begin{bmatrix} x_1 - \mathbb{E}x_1 & \cdots & x_d - \mathbb{E}x_d \end{bmatrix}$$

i^{th} — row

j^{th} — column

Ex $x = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$, $\theta \sim U[0, 2\pi]$

$$\mathrm{Var}(x) = \mathbb{E}\,x^2 = \frac{1}{2\pi}\int_0^{2\pi} \cos^2\theta \, d\theta = \frac{1}{2}$$

$$\mathrm{Var}(y) = \mathbb{E}\,y^2 = \frac{1}{2\pi}\int_0^{2\pi} \sin^2\theta \, d\theta = \frac{1}{2}$$

$$\Rightarrow C_x = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \qquad \mu_x = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Ex Multivariate Gaussian

generalizes 1d normal dist $\mathcal{N}(\mu, \sigma^2)$

to random vectors in $\mathbb{R}^d$, $\mathcal{N}(\mu, \Sigma)$

μ: mean vector $\in \mathbb{R}^d$

Σ: covariance matrix in $\mathbb{R}^{d \times d}$

$$p_X(x) = \frac{1}{|\det(2\pi\Sigma)|^{1/2}} \exp\left(\frac{-(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}\right)$$

weighted distance b/w x and the mean vector

when $d=1$, $\Sigma = \sigma^2$, $|\det(2\pi\Sigma)|^{1/2} = \sqrt{2\pi\sigma^2}$, $\Sigma^{-1} = \frac{1}{\sigma^2}$

$$\Rightarrow p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

so this is a direct generalization of the 1d normal dist.

Next class: finish up prob. theory, ERM, examples

HW1 will be posted later today due next Thurs.

& participation 2 " due Friday.