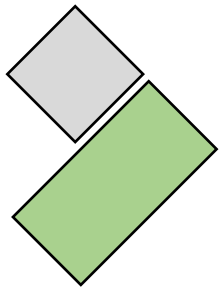
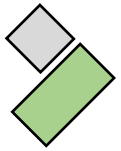


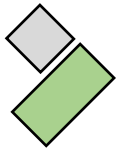
Probability & Distributions



Introduction

- ML often deals with *quantify uncertainties* (in the data, in the ML model, and in the predictions produced by the model)

- Quantifying uncertainty requires the idea of a *random variable*, and *probability distribution* associated with the random variable that measures the probability that a particular outcome will occur



Random Variable

- *Target space* function $X : \Omega \rightarrow T$

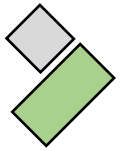
$$X(\omega) = x$$

$\omega \in \Omega$: an outcome,

$x \in T$: a value

- This association/mapping from Ω to T is called a *random variable*

- *Two types of X: discrete and continuous*



Univariate and Multivariate Distributions

- *Univariate* distribution: distribution of a *single* random variable
- *Bivariate* distribution: Distributions of *two* random variables
- *Multivariate* distributions: Distributions of *more than one* random variable, or a *vector* of random variables
  - Joint distribution
  - Marginal distributions
  - Conditional distributions

Example – Bivariate variable

Joint distribution example

- N = 100 students
- X, Y: marks on quiz 1, quiz 2 (r.v.)

$P(X = x, Y = y) = n_{xy}/N$

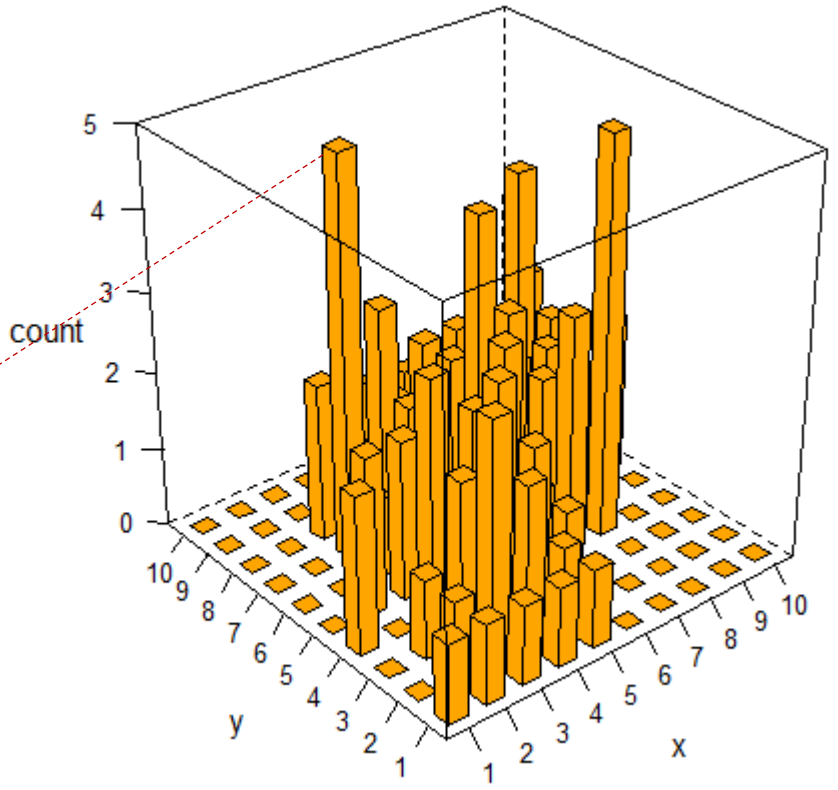
Ex.

$P(X = 3, Y = 7) = 0.05$

$P(X = 7, Y = 8) = 0.02$

$p(x, y) := P(X = x, Y = y)$
$= P(X = x \text{ and } Y = y)$

x \ y	1	2	3	4	5	6	7	8	9	10
1	1	0	0	2	0	0	0	0	0	0
2	1	1	1	0	2	0	0	0	0	0
3	1	3	2	3	2	1	5	2	0	0
4	1	2	3	1	1	2	3	1	0	0
5	1	1	2	3	2	0	0	0	0	0
6	0	0	1	1	3	4	2	2	1	1
7	0	0	0	3	2	2	0	2	1	1
8	0	0	0	5	0	2	4	1	1	1
9	0	0	0	0	0	1	2	0	1	0
10	0	0	0	0	0	2	1	0	2	1



Joint distribution

	1	2	3	4	5	6	7	8	9	10	ysum
1	0.03	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.08
2	0.01	0.01	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.07
3	0.03	0.00	0.02	0.01	0.01	0.01	0.02	0.04	0.00	0.00	0.14
4	0.00	0.01	0.01	0.02	0.01	0.00	0.02	0.02	0.00	0.00	0.09
5	0.00	0.00	0.03	0.04	0.01	0.00	0.01	0.00	0.00	0.00	0.09
6	0.00	0.00	0.01	0.01	0.01	0.03	0.03	0.03	0.03	0.00	0.15
7	0.00	0.00	0.01	0.03	0.03	0.02	0.02	0.03	0.02	0.00	0.16
8	0.00	0.00	0.00	0.03	0.01	0.02	0.02	0.03	0.00	0.00	0.11
9	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.02	0.02	0.08
10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.03
xsum	0.07	0.02	0.11	0.19	0.10	0.11	0.13	0.16	0.08	0.03	1.00

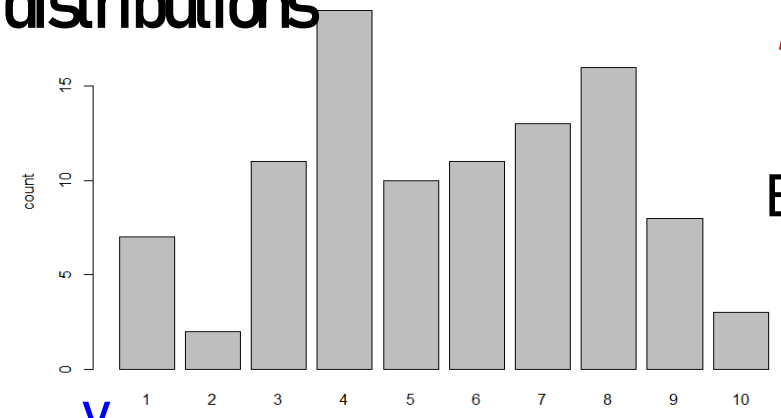
$\Sigma_x \Sigma_y p(x, y) = 1$

→ Tổng các giá trị = 1

Marginal distributions

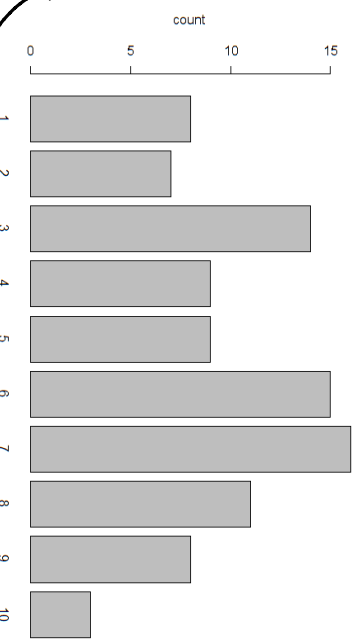
x \ y	1	2	3	4	5	6	7	8	9	10	ysum
1	3	0	1	3	1	0	0	0	0	0	8
2	1	1	2	2	1	0	0	0	0	0	7
3	3	0	2	1	1	1	2	4	0	0	14
4	0	1	1	2	1	0	2	2	0	0	9
5	0	0	3	4	1	0	1	0	0	0	9
6	0	0	1	1	1	3	3	3	3	0	15
7	0	0	1	3	3	2	2	3	2	0	16
8	0	0	0	3	1	2	2	3	0	0	11
9	0	0	0	0	0	2	1	1	2	2	8
10	0	0	0	0	0	1	0	0	1	1	3
xsum	7	2	11	19	10	11	13	16	8	3	100

Marginal distributions



x \ y	1	2	3	4	5	6	7	8	9	10	ysum
1	3	0	1	3	1	0	0	0	0	0	8
2	1	1	2	2	1	0	0	0	0	0	7
3	3	0	2	1	1	1	2	4	0	0	14
4	0	1	1	2	1	0	2	2	0	0	9
5	0	0	3	4	1	0	1	0	0	0	9
6	0	0	1	1	1	3	3	3	3	0	15
7	0	0	1	3	3	2	2	3	2	0	16
8	0	0	0	3	1	2	2	3	0	0	11
9	0	0	0	0	0	2	1	1	2	2	8
10	0	0	0	0	0	1	0	0	1	1	3
xsum	7	2	11	19	10	11	13	16	8	3	100

phân phối biên của x



chia % để tính xác suất

Marginal distribution $Y \sim p(y)$

$$p(y) = \Sigma_x p(x, y)$$

Ex. $p(Y = 6) = \Sigma_x p(x, 6) = 0.11$

Marginal distribution

$X \sim p(x)$,

$$p(x) = \Sigma_y p(x, y)$$

Ex. $p(X = 5) = \Sigma_y p(5, y) = 0.09$

Marginal distributions

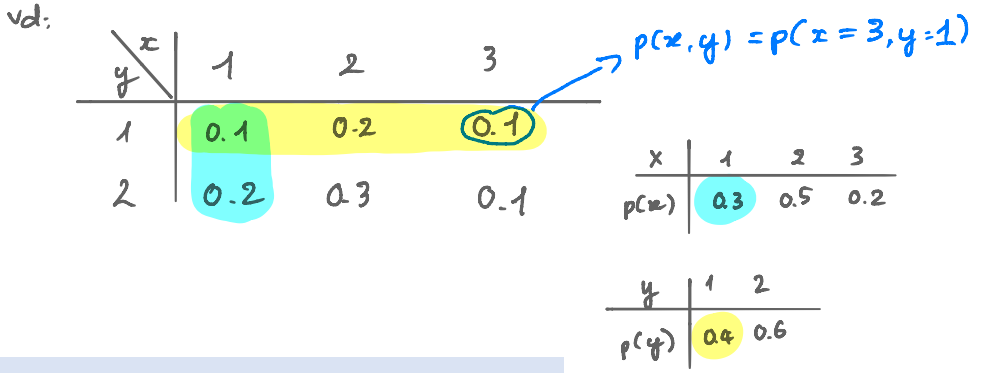
p(x)

		1	2	3	4	5	6	7	8	9	10	ysum
	1	0.03	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.08
	2	0.01	0.01	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.07
	3	0.03	0.00	0.02	0.01	0.01	0.01	0.02	0.04	0.00	0.00	0.14
	4	0.00	0.01	0.01	0.02	0.01	0.00	0.02	0.02	0.00	0.00	0.09
	5	0.00	0.00	0.03	0.04	0.01	0.00	0.01	0.00	0.00	0.00	0.09
	6	0.00	0.00	0.01	0.01	0.01	0.03	0.03	0.03	0.03	0.00	0.15
	7	0.00	0.00	0.01	0.03	0.03	0.02	0.02	0.03	0.02	0.00	0.16
	8	0.00	0.00	0.00	0.03	0.01	0.02	0.02	0.03	0.00	0.00	0.11
	9	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.02	0.02	0.08
	10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.03
p(y)	xsum	0.07	0.02	0.11	0.19	0.10	0.11	0.13	0.16	0.08	0.03	1.00

pmf of X
p(x)

X	1	2	3	4	5	6	7	8	9	10
P(X = x)	0.08	0.07	0.14	0.09	0.09	0.15	0.16	0.11	0.08	0.03

Sum Rule for marginal distributions



- $p(y) = \Sigma_x p(x, y)$
- $p(x) = \Sigma_y p(x, y)$

The sum rule relates the joint distribution to a marginal distribution

The joint probability distribution of two discrete random variables X and Y is partly given in the following table. Complete the table

b \ a	0	1	2	P(Y = b)
-1	?	?	?	1/2
1	?	1/2	?	1/2
P(X = a)	?	3/4	1/8	1

Sum Rule – Example

- $p(y) = \Sigma_x p(x, y)$
- $p(x) = \Sigma_y p(x, y)$

b / a	0	1	2	P(Y = b)
-1	?	1/4 ❶	?	1/2
1	0 ❷	1/2	0 ❸	1/2
P(X = a)	?	3/4	1/8	1

❶: p(Y = 1) = p(X = -1, Y = 1) + p(X = 1, Y = 1)
⬅➡ ¾ = ½ + ❶ ➔ ❶ = 1/4

❷: p(X = 1) = p(Y = 0, X = 1) + p(Y = 1, X = 1) + p(Y = 2, X = 1)
⬅➡ ½ = ❷ + ½ + ❸? ➔ ❷ = ❸ = 0

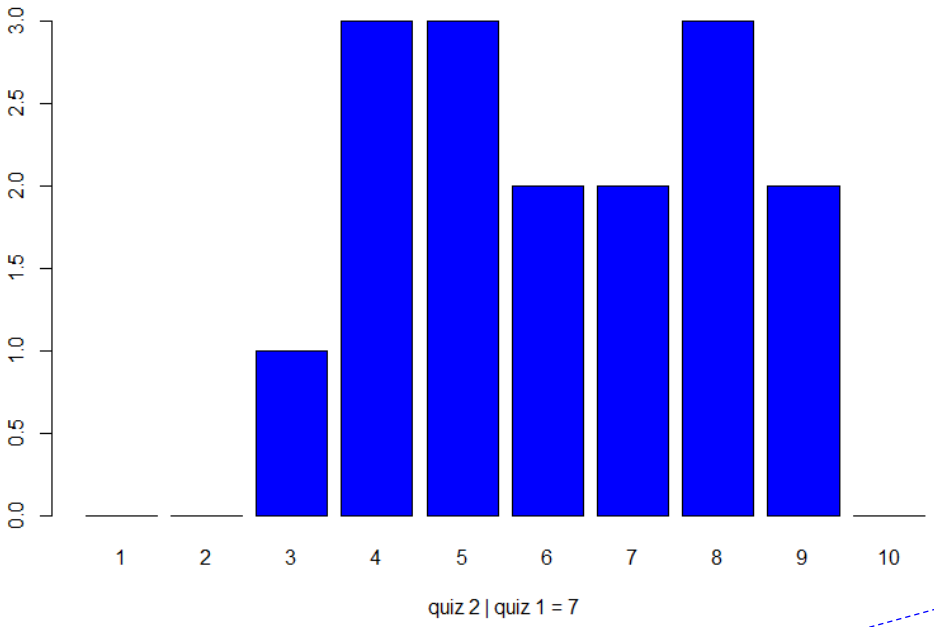
Conditional Distribution

- Collect all students who get 7 on quiz 1 (x = 7)

dùng bảng ↑ vd: $P(x=2 \mid y=1)$
ví dụ phía trên $= \frac{0.2}{0.4} = 0.5$

$P(x=1 \mid y=1) = 0.25$

$P(x=3 \mid y=1) = 0$



1	2	3	4	5	6	7	8	9	10	ysum
0	0	1	3	3	2	2	3	2	0	16

$$P(Y=y \mid X=x) = \frac{P(X=x, Y=y)}{P(X=x)}$$

$$P(Y=5 \mid X=7) = \frac{P(X=7, Y=5)}{P(X=7)}$$

$$= \frac{\frac{3}{100}}{\frac{16}{100}} = \frac{3}{16}$$

$$P(X=7 \mid Y=5) = \frac{P(X=7, Y=5)}{P(Y=5)}$$

$$= \frac{3}{10}$$

Vd:

y \ x	1	2	3
1	0.2	0.1	0.3
2	0.1	0.1	0.2

$E(X)$, $E(Y)$, $E(XY)$, $E(X, Y=2)$, $E(X+Y)$

↓ $p(x)$ ↓ $p(y)$ ↓ $p(x,y)$ ↓ $p(x \mid y=2)$

x	1	2	3
p(x)	0.3	0.2	0.5

$E(X) = 0.3 + 0.2 \times 2 + 0.5 \times 3$
$= 2.2$

$E(Y) = 1 \times 0.6 + 2 \times 0.4$
$= 1.4$

$E(XY) = 2.2 + 1.4 = 3.6$

$E(X+Y) = E(X) + E(Y)$
$= 2.2 + 1.4$
$= 3.6$
or
$= (1+1) \times 0.2 + (1+2) \times 0.1 + (1+3) \times 0.3$
$+ (2+1) \times 0.1 + (2+2) \times 0.1 + (2+3) \times 0.2 = 3.6$

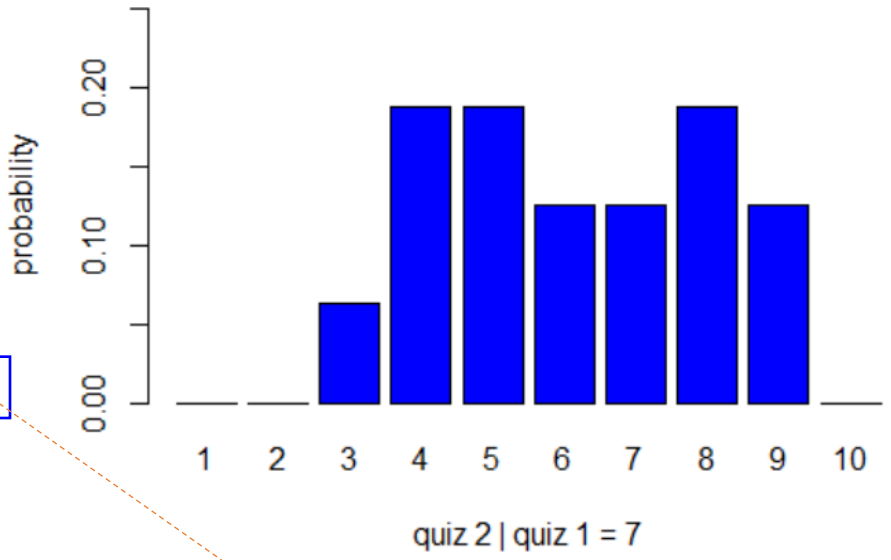
x	1	2	3
p(Y=2)	0.25	0.25	0.5

↳ $\frac{0.1}{0.4} = 0.25$

$P(X \mid Y=2) = 0.25 + 2 \times 0.25 + 3 \times 0.5$
$= 2,25$

Conditional Distribution

	1	2	3	4	5	6	7	8	9	10	ysum
1	0.03	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.08
2	0.01	0.01	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.07
3	0.03	0.00	0.02	0.01	0.01	0.01	0.02	0.04	0.00	0.00	0.14
4	0.00	0.01	0.01	0.02	0.01	0.00	0.02	0.02	0.00	0.00	0.09
5	0.00	0.00	0.03	0.04	0.01	0.00	0.01	0.00	0.00	0.00	0.09
6	0.00	0.00	0.01	0.01	0.01	0.03	0.03	0.03	0.03	0.00	0.15
7	0.00	0.00	0.01	0.03	0.03	0.02	0.02	0.03	0.02	0.00	0.16
8	0.00	0.00	0.00	0.03	0.01	0.02	0.02	0.03	0.00	0.00	0.11
9	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.02	0.02	0.08
10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.03
xsum	0.07	0.02	0.11	0.19	0.10	0.11	0.13	0.16	0.08	0.03	1.00

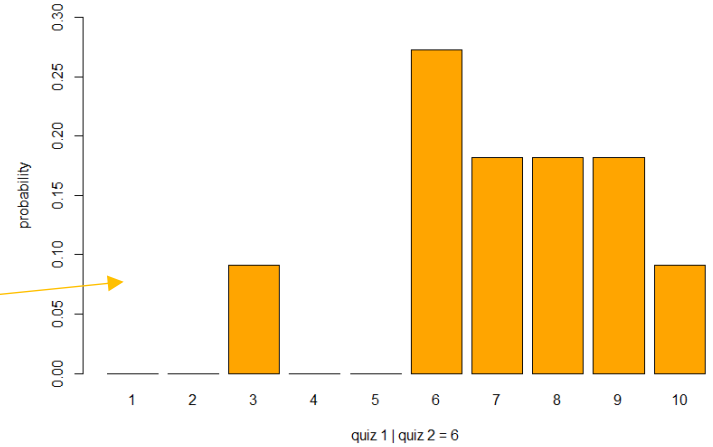


$$p(y \mid x) = \frac{p(x,y)}{p(x)} \Leftrightarrow P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

Conditional Distribution

Collect all students who get 6 on quiz 2 (Y = 6)

x \ y	1	2	3	4	5	6	7	8	9	10	ysum
1	0.03	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.08
2	0.01	0.01	0.02	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.07
3	0.03	0.00	0.02	0.01	0.01	0.01	0.02	0.04	0.00	0.00	0.14
4	0.00	0.01	0.01	0.02	0.01	0.00	0.02	0.02	0.00	0.00	0.09
5	0.00	0.00	0.03	0.04	0.01	0.00	0.01	0.00	0.00	0.00	0.09
6	0.00	0.00	0.01	0.01	0.01	0.03	0.03	0.03	0.03	0.00	0.15
7	0.00	0.00	0.01	0.03	0.03	0.02	0.02	0.03	0.02	0.00	0.16
8	0.00	0.00	0.00	0.03	0.01	0.02	0.02	0.03	0.00	0.00	0.11
9	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.02	0.02	0.08
10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.03
xsum	0.07	0.02	0.11	0.19	0.10	0.11	0.13	0.16	0.08	0.03	1.00



$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$P(X = 8 \mid Y = 6) = \frac{P(X = 8, Y = 6)}{P(Y = 6)} = \frac{2}{11}$$

$$P(X = x \mid Y = 6) = \frac{P(X = x, Y = 6)}{P(Y = 6)}$$

Conditional probability

Joint distribution

The conditional probability of Y given X is $$p(y|x) = \frac{p(x,y)}{p(x)}$$

Marginal distribution

$$\Rightarrow p(x|y)\,p(y) = p(y|x)\,p(x)$$

Product rule

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

$$p(x,y) = p(x|y)\,p(y) = p(y|x)\,p(x)$$

"Every joint distribution of two random variables can be factorized as a product of two other distributions"

Total probability Rule

$$p(y) = \sum_x p(x, y) \quad \text{[Sum Rule]}$$

$$p(x, y) = p(y \mid x)\, p(x) \quad \text{[Product Rule]}$$

$$\Rightarrow \boxed{p(y) = \sum_x p(y \mid x)\, p(x)}$$

Exercise

Two random variables X and Y have *joint distribution* shown below.

1/ Find marginal distributions p(x), p(y)

2/ Find conditional distributions p(x | Y = 1), p(y | X = -1)

3/ Find conditional probability p(X = 0 | Y = 1)

X \ Y	0	1	2
-1	0.2	0	0.3
0	0	0.1	0
1	0.2	0.1	0.1

1/ Find marginal distributions p(x), p(y)

Use sum rule

p(X = -1)

= p(X = -1, Y = 0) +

+ p(X = -1, Y = 1)

+ p(X = -1, Y = 2)

= 0.2 + 0 + 0.3

= 0.5

p(X = -1, Y = 0)

X \ Y	0	1	2
-1	0.2	0	0.3
0	0	0.1	0
1	0.2	0.1	0.1

x	-1	0	1
p(x)	0.5	0.1	0.4

Marginal distribution p(x), p(y)

y	0	1	2
p(y)	0.4	0.2	0.4

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

2/ Find conditional distributions p(x | Y = 1), p(y | X = -1)

Joint distribution

$$p(y|X=-1) = \frac{p(X=-1,y)}{p(X=-1)}$$

Marginal distribution

X \ Y	0	1	2
-1	0.2	0	0.3
0	0	0.1	0
1	0.2	0.1	0.1

Joint distribution p(x, y)

y	0	1	2
p(y|X = -1)	0.2/0.5	0/0.5	0.3/0.5

Conditional distributions p(y | X = -1)

x	-1	0	1
p(x)	0.5	0.1	0.4

Marginal distribution p(x)

3/ Find conditional probability p(X = 0 | Y = 1)

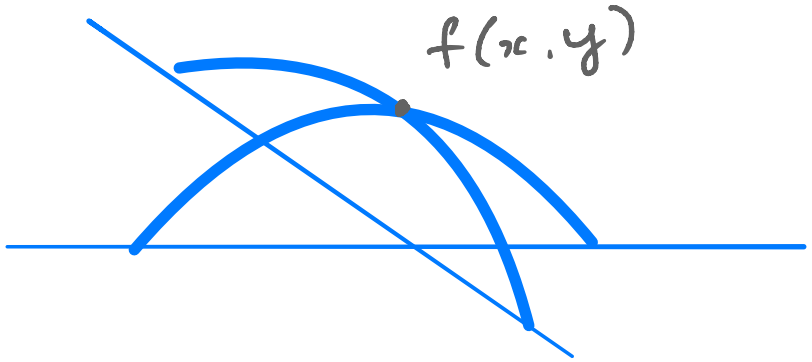
p(X = 0 | Y = 1)

$= \frac{p(X=0,Y=1)}{p(Y=1)} = \frac{0.1}{0.2} = 0.5$

X \ Y	0	1	2
-1	0.2	0	0.3
0	0	0.1	0
1	0.2	0.1	0.1

y	0	1	2
p(y)	0.4	0.2	0.4

(Continuous) Bivariate R.V.



- *pdf f(x,y)* of bivariate random variable

$$(1)\quad f(x,y) \geq 0, \forall (x,y) \in \mathbb{R}^2$$

$$(2)\quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy = 1$$

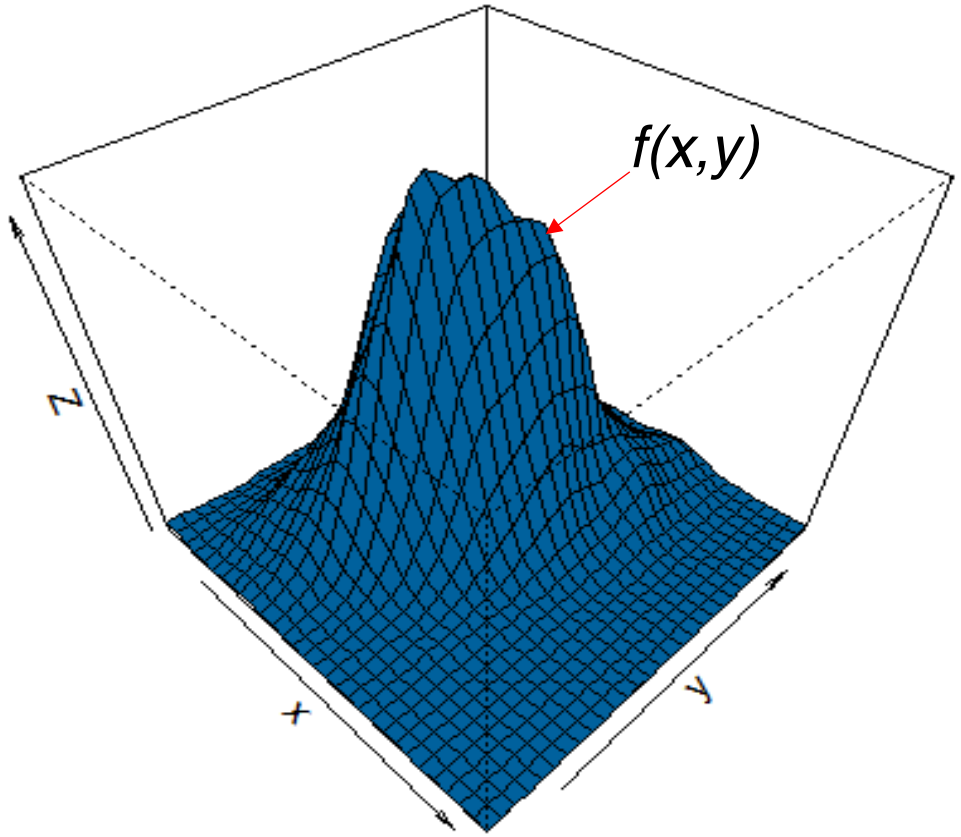
$$f(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$$

- *cdf* of bivariate random variable

$$F_{XY}(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)\,dudv$$

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_{c}^{d}\int_{a}^{b} f(x,y)\,dxdy$$

Plot the probability density function (pdf) of a bivariate random variable



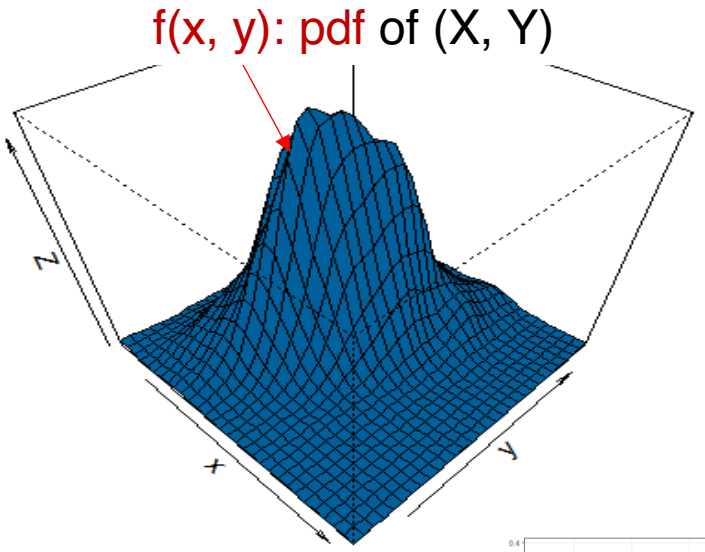
Marginal distributions

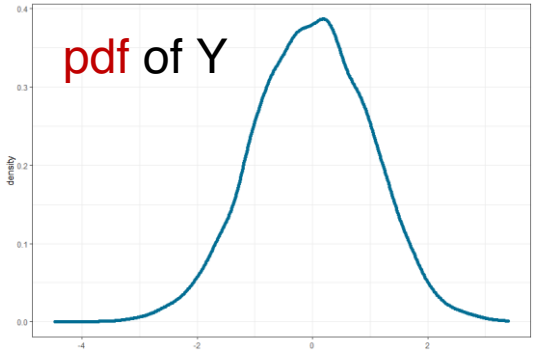
- *Marginal pdf* of X

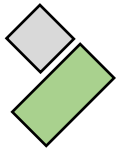
$$f_X(x) = \int_{D_y} f(x,y)\,dy$$

- *Marginal pdf* of Y

$$f_Y(y) = \int_{D_x} f(x,y)\,dx$$



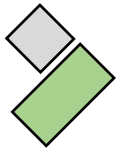




Conditional distributions

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)} = \frac{f(x,y)}{\int_{D_x} f(x,y)\,dx}$$

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{f(x,y)}{\int_{D_y} f(x,y)\,dy}$$



Example

Let the *joint pdf* of (X, Y) be given by $f(x, y) = m(x^2 + y)$, $0 \le x \le 1$, $0 \le y \le 2$.

1/ Find the value of m

2/ Find *marginal pdf* of X, and Y

3/ Compute $P(0 \le X \le ½, 1 \le Y \le 3/2)$

4/ Compute $P(0 \le X \le ½ \mid Y = 1)$

5/ Compute $P(0 \le X \le ½ \mid 1 \le Y \le 3/2)$

1/ $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy = \int_{0}^{2}\int_{0}^{1} m\left(x^2+y\right)dxdy$$

$$= m\int_{0}^{2}\left(\left[\frac{x^3}{3}+yx\right]_0^1\right)dy = m\int_{0}^{2}\left(\frac{1}{3}+y\right)dy = m\left(\frac{1}{3}y+\frac{y^2}{2}\right)\Bigg|_0^2 = \frac{8m}{3}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy = 1 \Leftrightarrow m = \frac{3}{8}$$

2/ Find marginal pdf of X, and Y

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^2 f(x,y)\,dy$$

$$= \int_0^2 \frac{3}{8}\left(x^2 + y\right)dy = \frac{3}{8}\left(x^2 y + \frac{y^2}{2} \right)\Bigg|_0^2 = \frac{3}{8}\left(2x^2 + 2\right)$$

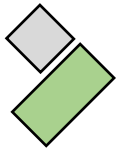
$$\Leftrightarrow f_X(x) = \frac{3}{4}x^2 + \frac{3}{4}$$

3/ Compute P(0 ≤ X ½, 1 ≤ Y ≤ 3/2)

$$P\left(0 \le X \le \frac{1}{2}, 1 \le Y \le \frac{3}{2}\right) = \int_1^{\frac{3}{2}} \int_0^{\frac{1}{2}} f(x,y)\,dxdy = \int_1^{\frac{3}{2}} \int_0^{\frac{1}{2}} \frac{3}{8}\left(x^2 + y\right)dxdy$$

$$= \int_1^{\frac{3}{2}} \left(\frac{x^3}{8} + \frac{3xy}{8} \right)\Bigg|_0^{\frac{1}{2}} dy = \int_1^{\frac{3}{2}} \left(\frac{1}{64} + \frac{3y}{16} \right) dy$$

$$= \left(\frac{y}{64} + \frac{3y^2}{32} \right)\Bigg|_1^{\frac{3}{2}} = \frac{43}{512} \approx 0.084$$



4/ Compute P(0 ≤ X ≤ ½ | Y = 1)

$$f_{X|Y=1} = \frac{f(x,1)}{f_Y(1)} = \frac{\frac{3}{8}\left(x^2+1\right)}{\int_0^1 \frac{3}{8}\left(x^2+1\right)dx} = \frac{x^2+1}{\frac{4}{3}} = \frac{3}{4}\left(x^2+1\right)$$

$$P\left(0 \le X \le \frac{1}{2} \mid Y=1\right) = \int_0^{\frac{1}{2}} f_{X|Y=1}dx = \int_0^{\frac{1}{2}} \frac{3}{4}\left(x^2+1\right)dx = \frac{13}{32}$$

5/ P(0 ≤ X ≤ ½ | 1 ≤ Y ≤ 3/2)

$$P\left(0 \le X \le \frac{1}{2} \mid 1 \le Y \le \frac{3}{2}\right) = \frac{P\left(0 \le X \le \frac{1}{2}, 1 \le Y \le \frac{3}{2}\right)}{P\left(1 \le Y \le \frac{3}{2}\right)}$$

Continuous Probabilities

Definition.

A function f : $\mathbb{R}^D \rightarrow \mathbb{R}$ is called a *probability density function* (pdf) if

1. $\forall \boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) \geqslant 0$
2. Its integral exists and

$$\int_{\mathbb{R}^D} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1 \,.$$

Vd: joint pdf of x, y:

$$f(x,y) = \begin{cases} c(2x+y^2), & 0<x<1,\ 0<y<1 \\ 0, & \text{otherwise} \end{cases}$$

a) Find the constant c

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_0^1\int_0^1 c(2x+y)^2\,dx\,dy = c\int_0^1 (1+y^2)\,dy$$

$$= c\cdot\frac{4}{3}$$

$$\Rightarrow c = \frac{3}{4}$$

b) $f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$

$$= \begin{cases} \int_0^1 \frac{3}{4}(2x+y^2)\,dy, & 0<x<1 \\ 0, \text{otherwise} \end{cases}$$

$$= \frac{3}{4}\left(2xy + \frac{y^3}{3}\right)\Big|_0^1 = \frac{3}{4}\left(2x+\frac{1}{3}\right)$$

c) $E(X \mid Y=\frac{1}{2})$ → dv trường hợp này phải tính f lại

$$f_{X|Y}(x) = \frac{f(x,y)}{f_y(y)} = \frac{\frac{3}{4}(2x+y^2)}{\frac{3}{4}(1+y^2)} \overset{y=\frac{1}{2}}{=} \frac{(2x+\frac{1}{4})}{1+\frac{1}{4}} = \frac{4}{5}\left(2x+\frac{1}{4}\right)$$

$$E(X|Y=\tfrac{1}{2}) = \int_{-\infty}^{\infty} x\cdot f_{X|Y}(x)\,dx = \frac{4}{5}\int_{-\infty}^{\infty} x\cdot\left(2x+\frac{1}{4}\right)dx = \frac{19}{30} \approx 0.63$$

d) $V(X) = \sigma^2 = E(X^2) - [E(X)]^2$

$$= \int_{-\infty}^{\infty} x^2\cdot f_x(x)\,dx - \left[\int_{-\infty}^{\infty} x\cdot f_x(x)\,dx\right]^2$$

$$= \int_0^1 x^2\cdot\frac{3}{4}\left(2x+\frac{1}{3}\right) - \left[\int_0^1 x\cdot\frac{3}{4}\left(2x+\frac{1}{3}\right)\right]^2$$

$$= 0.067$$

e) $P(0<X\le\frac{1}{2},\ \frac{1}{2}\le Y<1)$

$$= \int_0^{\frac{1}{2}}\int_{\frac{1}{2}}^1 \frac{3}{4}(2x+y^2)\,dy\,dx = \text{(tính tay)}$$

Vd: $f(x,y) = \begin{cases} c \cdot xy^2 & , 0 < x \quad , y < 1 \\ 0 & , \text{otherwise} \end{cases}$

Find $E(x)$, $E(Y)$, $E(XY)$

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_0^1\int_0^1 c(xy^2)\,dx\,dy$$

$$= c\int_0^1 \frac{y^2}{2}\,dy$$

$$= c\left(\frac{y^3}{6}\right)\Big|_0^1 = c \cdot \frac{1}{6} \Rightarrow c = 6$$

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 6xy^2\,dy = 6x\frac{y^3}{3}\Big|_0^1 = 2x$$

$$E(x) = \int_{-\infty}^{\infty} x \cdot f_x(x)\,dx = \int_0^1 x \cdot 2x\,dx = \frac{2}{3}$$

$$f_y(y) = \int_0^1 6xy^2\,dx = 3x^2y^2\Big|_0^1 = 3y^2$$

$$E(Y) = \int_0^1 y \cdot f_y(y)\,dy = \int_0^1 y \cdot 3y^3\,dy = \frac{3}{4}$$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \cdot f(x,y)\,dx\,dy = \int_0^1\int_0^1 xy \cdot 6xy^2\,dx\,dy$$

$$= \frac{1}{2}$$

(phương sai)

Covariance X, Y

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

if cov		
	$= 0$	độc lập
	> 0	đồng biến
	< 0	nghịch biến

Correlation

$$Corr[X,Y] = \frac{Cov[X,Y]}{\sigma_x \cdot \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

correlation matrix

	x	y
x	1	$\frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$
y	$\frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$	1

Practice

y \ x	1	2	3
0	0.1	0.2	0.1
1	0.1	0.1	0.1
2	0.1	0.1	0.1

(?) $E(X)$, $E(Y)$, $E(X \mid Y=2)$

	1	2	3
$P(x)$	0.3	0.4	0.3

$E(X) = 1\times0.3 + 2\times0.4 + 3\times0.3 = 1.7$

	0	1	2
$P(y)$	0.4	0.3	0.3

$E(y) = 0\times0.4 + 1\times0.3 + 2\times0.3 = 0.9$

	0	1	2
$P(y=2)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$E(X \mid Y=2) = 1\times\frac{1}{3} + 2\times\frac{1}{3} + 3\times\frac{1}{3} = 2$

$E(XY) = E(X) + E(Y) = 1.7 + 0.9 = 2.6$

$cov[X,Y] = E(X,Y) - E(X)E(Y)$

$= 1,8 - 2\times0.9 = 0$ (độc lập)

correlation coefficient $corr[x,y] = \frac{cov[X,Y]}{\sigma_x \cdot \sigma_y} = 0$

$X = [X_1 \quad X_2 \quad X_3]^T$ random vector

variance matrix

$$V[X] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \end{array}_{3\times3}$$

correlation matrix

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{5}} \\ \frac{1}{\sqrt{3}} & 1 & \frac{1}{2\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{2\sqrt{5}} & 1 \end{bmatrix} \end{array} \leftarrow V[X] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 3 & 2 & -2 \\ 2 & 4 & 1 \\ -2 & 1 & 5 \end{bmatrix} \end{array}$$

variance matrix

$$corr[x_1, x_2] = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2} = \frac{2}{\sqrt{3}\cdot\sqrt{4}} = \frac{1}{\sqrt{3}}$$

$$corr[x_1, x_3] = \frac{\sigma_{13}}{\sigma_1 \cdot \sigma_3} = \frac{-2}{\sqrt{3}\cdot\sqrt{5}} = \frac{-2}{\sqrt{5}}$$

$X = [X_1, X_2, X_3]^T$ random vector

$E(X) = [2, -1, 1]^T$ mean vector

Let $Y = AX$ where $A = \begin{bmatrix} 1 & 0 & 2 \\ -1 & 1 & 3 \end{bmatrix}_{2\times3}$

Compute variance matrix $V(Y)$

$$V(Y)_{2\times2} = A_{2\times3} \times V(X)_{3\times3} \times A^T_{3\times2}$$

$$V(Y) = \begin{bmatrix} 1 & 0 & 2 \\ -1 & 1 & 3 \end{bmatrix}\begin{bmatrix} 3 & 2 & -2 \\ 2 & 4 & 1 \\ -2 & 1 & 5 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 2 & 3 \end{bmatrix} \quad \sigma_{12}$$

$$= \begin{bmatrix} 15 & 29 \\ 29 & 66 \end{bmatrix}$$

$V(X) = 5$, $V(Y) = 3$, $Cov[x,y] = -3$

Let $Z = 3X - 2Y = [3 \quad -2]\begin{bmatrix} X \\ Y \end{bmatrix}$

Compute $V(Z)$

$$V(Z) = A\begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix}A^T = [3 \quad -2]\begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix}\begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

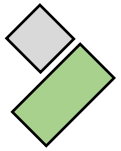
or

$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab\, cor[X,Y]$

$V(Z) = V(3X - 2Y)$

$= 9\times5 - 12(-3) + (-2)^2\times3$

$= 93$



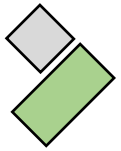
Definition. (Cumulative Distribution Function).

A *cumulative distribution function (cdf)* of a multivariate real-valued random variable X with states $x \in R^D$ is given by

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

The *cdf* can be expressed also as the integral of the *probability density function* f(x) so that

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \ldots, z_D) \mathrm{d}z_1 \cdots \mathrm{d}z_D \,.$$



- If the *cdf* $F_X(.)$ is differentiable, then the *pdf* *f(.)* is

$$f(\mathbf{x}) = \frac{\partial^D F_X(\mathbf{x})}{\partial x_1 \partial x_2 \cdots \partial x_D}$$

Marginal distributions

- Let $X_1, X_2, \ldots, X_n$ be n random variables with joint probability $p(x_1, x_2, \ldots, x_n)$ or $f(x_1, x_2, \ldots, x_n)$, then

  - Discrete case:

$$p(x_i) = \sum_{x1}\sum_{x2}\cdots \sum_{x(i-1)}\sum_{x(i+1)}\sum_{xn} p(x_1, x_2, \ldots, x_n)$$

  - Continuous case:

$$f(x_i) = \int \int \ldots \int f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

Conditional distributions

- Let $X_1, X_2, \ldots, X_n$ be n random variables with joint probability

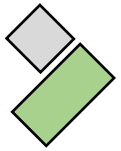
$p(x_1, x_2, \ldots, x_n)$ or $f(x_1, x_2, \ldots, x_n)$, then the conditional joint probability of $X = (X_1, X_2, \ldots, X_q)$ given $Y= (X_{q+1}, X_{q+2}, \ldots, X_n)$ is

  - Discrete case:

$$p_{X|Y}(x_1, x_2, \ldots, x_q \mid x_{q+1}, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, xn)}{p_Y(xq_{+1, \ldots,} xn)}$$

  - Continuous case:

$$f_{X|Y}(x_1, x_2, \ldots, x_q \mid x_{q+1}, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, xn)}{f_Y(xq_{+1, \ldots,} xn)}$$

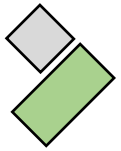


The Sum Rule

- Recall.
  - p(x, y): is the *joint distribution* of the two random variables x, y.
  - p(x) and p(y): *marginal distributions*,
  - and p(y | x): the *conditional distribution* of y given x.
- The *sum rule* states that

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

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



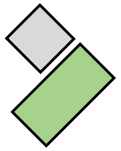
The Sum Rule

- If $x = [x_1, \ldots, x_D]^T$, we obtain

$$p(x_i) = \int p(x_1, \ldots, x_D) \mathrm{d}\boldsymbol{x}_{\setminus i}$$

\i means “all except i”

by repeated application of the sum rule for where we integrate/sum out all random variables except x_i



The Product Rule

- The *product rule* relates the *joint distribution* to the conditional distribution via

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

➔ Every joint distribution of two random variables can be factorized (written as a product) of two other distributions.

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

Total Probability Rule (discrete)

- Sum rule

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

- Product Rule

$$p(x, y) = p(x \mid y)\, p(y)$$

➔ Total probability rule

$$p(x) = \sum_{y} p(x \mid y)\, p(y)$$

Bayes' theorem (Bayes' law)

- In ML and Bayesian statistics, we are often interested in *making inferences* of X (unobserved) given Y (observed).
- Assume
  - p(x): some prior knowledge about x (unobserved)
  - p(y | x): some relationship between x and y (y can be observed)

- *Bayes' theorem.*

  Given y, we can use Bayes' theorem to draw some conclusions about x.

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

Bayes' Law – Example

(Reliability of a test). There exists a test for a certain viral infection. The *probability* that the test correctly identifies someone with the illness as *positive* is 0.99, and the probability that the test correctly identifies someone without the illness as *negative* is 0.95. The incidence of the illness in the general population is 0.0001.

Find:

a) The probability that you get positive in a test.

b) Find the probability that you have the illness given the result of your test is positive.

Let x_1 = “you have the illness”,
x_2 = “you do not have the illness”,
y = “the test signals positive”.

	x_1	x_2
y	0.99	0.05

a) Use *total probability rule* to find p(y)

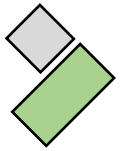
$p(y) = p(y \mid x_1)p(x_1) + p(y \mid x_2)p(x_2)$

$= (0.99)(0.0001) + (0.05)(0.9999) = 0.050094$

b) Use *Bayes' Rule* to compute $p(x \mid y)$

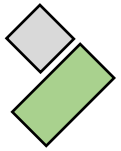
$p(x \mid y) = p(y \mid x)p(x)/p(y) = (0.99)(0.0001)/(0.050094)$

$= 0.001976285 \approx 0.2\%$



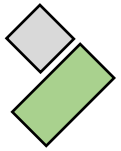
Explanation of prior p(x), likelihood p(y | x), and posterior p(x | y)

- p(x), the prior, encapsulates our subjective prior knowledge of the unobserved (latent) variable x before observing any data.
- The "likelihood of x (given y)" or the "probability of y given x" p(y | x) describes how x and y are related, and in the case of discrete probability distributions, it is the probability of the data y if we were to know the latent variable x.
- The posterior p(x | y) is the quantity of interest in Bayesian statistics because it expresses exactly what we are interested in (i.e., what we know about x after having observed y). The posterior can be used within a decision-making system.



Conditional probability

- $p(x, y, z) = p(x \mid y, z)p(y, z) = p(x \mid y, z)p(y \mid z)p(z)$
- $p(x, y, z, w) = p(x \mid y, z, w)p(y, z, w) = p(x \mid y, z, w)p(y \mid z, w)p(z, w) = p(x \mid y, z, w)p(y \mid z, w)p(z \mid w)p(w)$

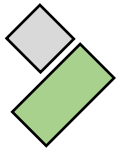


Marginal likelihood/evidence

The *marginal likelihood*

$$p(\boldsymbol{y}) := \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbb{E}_X[p(\boldsymbol{y} \mid \boldsymbol{x})]$$

can also be interpreted as the *expected likelihood* where we take the expectation with respect to the prior p(x).



Means and Covariances

- The concept of the *expected value* is central to ML, and the foundational concepts of probability itself can be derived from the expected value
- **Definition (Expected Value).** The *expected value* of a function $g : R \rightarrow R$ of a univariate random variable $X \sim p(x)$ is given by

(continuous) $$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)\mathrm{d}x$$

where $\mathcal{X}$ is the set of possible outcomes

(discrete) $$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

Expected value

(X is continuous)

- Given the *pdf* of a random variable X,
- Find $E_X[x]$, $E_X(x^2)$

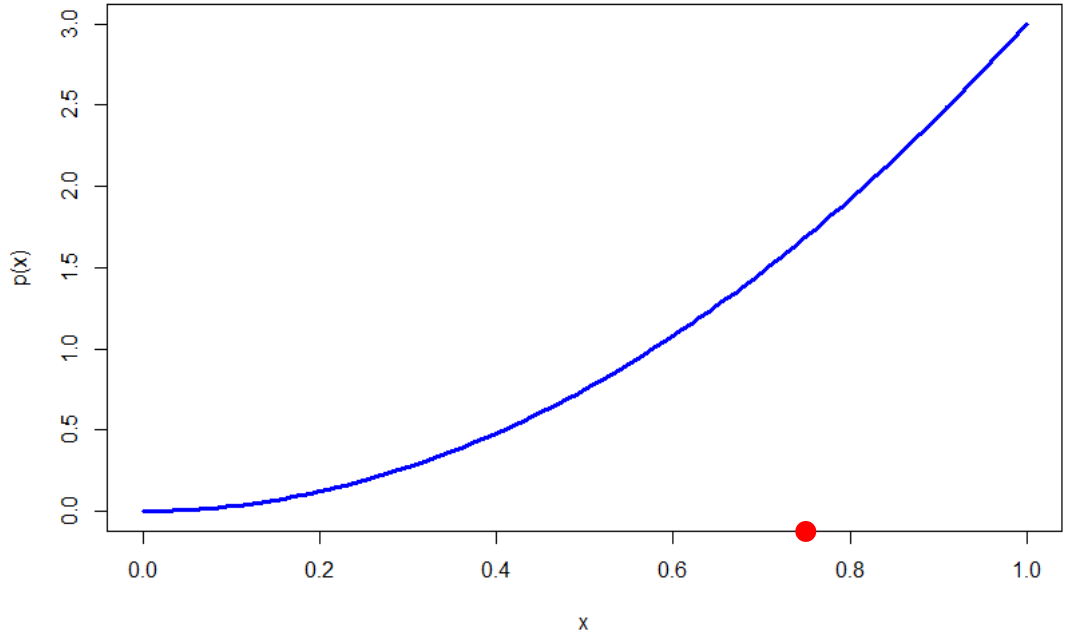
$$p(x) = \begin{cases} 3x^2 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

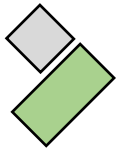
$$E_X[x] = \int_{-\infty}^{\infty} xp(x)\,dx = \int_{0}^{1} xp(x)\,dx$$

$$= \int_{0}^{1} 3x^3\,dx = \frac{3}{4}$$

$$E_X[x^2] = \int_{-\infty}^{\infty} x^2 p(x)\,dx = \int_{0}^{1} x^2 p(x)\,dx$$

$$= \int_{0}^{1} 3x^4\,dx = \frac{3}{5}$$





Expected Value

- For *multivariate random variables* $X = [X_1, X_2, \ldots, X_D]^T$, we define the *expected value* element wise

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

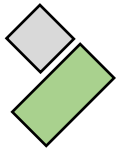
Mean

- ***Definition (Mean).*** The *mean* of a random variable X with states $x \in R^D$ is an *average* and is defined as

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

where

$$\mathbb{E}_{X_d}[x_d] := \begin{cases} \int_{\mathcal{X}} x_d p(x_d) \mathrm{d}x_d \\ \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) \end{cases}$$



Expectation, Mean – Ex 3

- Consider joint distribution of two random variables X_1, X_2

X_1 \ X_2	0	1	2
0	0.1	0.25	0.16
1	0.15	0.22	0.12

Compute $E[x_1]$, $E[x_2]$, $E[x_1 x_2]$

X_1 \ X_2	0	1	2	
0	0.1	0.25	0.16	0.51
1	0.15	0.22	0.12	0.49
	0.25	0.47	0.28	1

- Marginal distributions

X_1	0	1
$p(x_1)$	0.51	0.49

X_2	0	1	2
$p(x_2)$	0.25	0.47	0.28

- $E[x_1x_2] = (0\times0)(0.1) + (0\times1)(0.25) + (0\times2)(0.16) + (1\times0)(0.15) + (1\times1)(0.22) + (1\times2)(0.12) = 0.46$
- $E[x_1] = 0(0.51) + 1(0.49) = 0.49$, $E[x_2] = 1.03$

Expectation, Mean – Continuous Bivariate r. v. – Ex 4

- Given *pdf* of a bivariate random variable

$$p(x, y) = \frac{1}{2}x + \frac{3}{2}y, 0 \le x, y \le 1$$

- Compute E[x], E[y], and E[x^2], E[x | Y = y], E[y | X = ½]

$$E[x] = \int_0^1 x f_X(x)\,dx \qquad E[y] = \int_0^1 y f_Y(y)\,dy$$

$$E_X[x^2] = \int_0^1 x^2 f_X(x)\,dx \qquad E[x \mid Y = y] = \int_0^1 x f_{X|Y}(x \mid y)\,dx$$

$$p(x,y) = \frac{1}{2}x + \frac{3}{2}y, 0 \le x, y \le 1$$

$$f_X(x) = \int_0^1 f(x,y)\,dy = \int_0^1 \left(\frac{x}{2} + \frac{3y}{2}\right) dy$$

$$= \left(\frac{xy}{2} + \frac{3y^2}{4}\right)\Bigg|_0^1 = \frac{x}{2} + \frac{3}{4}$$

$$E[x] = \int_0^1 x f_X(x)\,dx = \int_0^1 x\left(\frac{x}{2} + \frac{3}{4}\right) dx = \left(\frac{x^3}{6} + \frac{3x^2}{8}\right)\Bigg|_0^1 = \frac{13}{24}$$

$$p(x,y) = \frac{1}{2}x + \frac{3}{2}y, 0 \le x, y \le 1$$

$$f_Y(y) = \int_0^1 f(x,y)\,dx = \int_0^1 \left(\frac{x}{2} + \frac{3y}{2}\right)dx$$

$$= \left(\frac{x^2}{4} + \frac{3xy}{2}\right)\Bigg|_0^1 = \frac{1}{4} + \frac{3y}{2}$$

$$E[y] = \int_0^1 y f_Y(y)\,dy = \int_0^1 y\left(\frac{1}{4} + \frac{3y}{2}\right)dy = \left(\frac{y^2}{8} + \frac{3y^3}{8}\right)\Bigg|_0^1 = \frac{1}{2}$$

$$p(x,y) = \frac{1}{2}x + \frac{3}{2}y, 0 \le x, y \le 1$$

$$f_X(x) = \frac{x}{2} + \frac{3}{4}$$

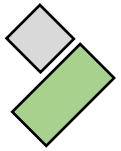
$$E[x^2] = \int_0^1 x^2 f_X(x)\,dx = \int_0^1 x^2\left(\frac{x}{2} + \frac{3}{4}\right)dx = \left(\frac{x^4}{8} + \frac{3x^3}{12}\right)\Bigg|_0^1 = \frac{9}{24}$$

$$p(x,y) = \frac{1}{2}x + \frac{3}{2}y, 0 \le x, y \le 1$$

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)} = \frac{\frac{x}{2} + \frac{3y}{2}}{\frac{1}{4} + \frac{3y}{2}} = \frac{2x+6y}{1+6y}$$

$$E[x \mid Y = y] = \int_0^1 x f_{X|Y}(x \mid y)\,dx = \int_0^1 x \frac{2x+6y}{1+6y}\,dx = \frac{1}{1+6y}\left(\frac{2x^3}{3} + 3yx^2\right)\Big|_0^1$$

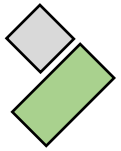
$$= \frac{\frac{2}{3} + 3y}{1+6y}$$



Linearity of Expectations

- The expected value is a *linear operator*. For example, given a real-valued function f(x) = ag(x)+bh(x) where a, b ∈ R and x ∈ R^D, we obtain

$$\begin{aligned}\mathbb{E}_X[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int [ag(\boldsymbol{x}) + bh(\boldsymbol{x})]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= a\int g(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}x + b\int h(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= a\mathbb{E}_X[g(\boldsymbol{x})] + b\mathbb{E}_X[h(\boldsymbol{x})]\,.\end{aligned}$$



Covariance

- **Definition (Covariance (Univariate))**. The *covariance* between two univariate random variables X, Y ∈ R is given by the *expected product of their deviations from their respective means*, i.e.,

$$\mathrm{Cov}_{X,Y}[x, y] := \mathbb{E}_{X,Y}\left[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])\right] .$$

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] .$$

Note. $E_X[x]$ is often written as $E[x]$)

Variance, Standard deviation

- The *covariance* of a variable with itself Cov[x, x] is called the *variance* and is denoted by $V_X[x]$

➔ $V[x] = E[x^2] - (E[x])^2$

- The *standard deviation,* σ(x), is defined by

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

Covariance, Variance – Ex 1

- Consider joint distribution of two random variables X_1, X_2

X_1 \ X_2	0	1	2
0	0.1	0.25	0.16
1	0.15	0.22	0.12

Compute $Cov[x_1, x_2]$, $V[x_1]$, $V[x_2]$, $\sigma(x_1)$, and $\sigma(x_2)$

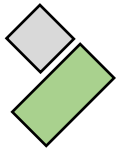
X_1 \ X_2	0	1	2	
0	0.1	0.25	0.16	0.51
1	0.15	0.22	0.12	0.49
	0.25	0.47	0.28	1

- Marginal distributions

X_1	0	1
$p(x_1)$	0.51	0.49

X_2	0	1	2
$p(x_2)$	0.25	0.47	0.28

- $E[x_1x_2] = (0\times0)(0.1) + (0\times1)(0.25) + (0\times2)(0.16) + (1\times0)(0.15) + (1\times1)(0.22) + (1\times2)(0.12) = 0.46$
- $E[x_1] = 0(0.51) + 1(0.49) = 0.49$, $E[x_2] = 1.03$
- $Cov[x_1, x_2] = E[x_1x_2] - E[x_1]E[x_2] = 0.46 - (0.49)(1.03) = -0.0477$



X_1 \ X_2	0	1	2	
0	0.1	0.25	0.16	0.51
1	0.15	0.22	0.12	0.49
	0.25	0.47	0.28	1

- $V[x_1] = E[x_1x_1] - (E[x_1])^2$

$= (0^2)0.51 + (1^2)(0.49) - 0.49^2 = 0.2499$

➔ $\sigma(x_1) = 0.4999$

- $V[x_2] = E[x_2x_2] - (E[x_2])^2$

$= (0^2)0.25 + (1^2)0.47 + (2^2)0.28 - 1.03^2$

$= 0.5291$

➔ $\sigma(x_2) = 0.7273926$

	x_1	x_2
x_1	$V[x_1]$	$Cov[x_1,x_2]$
x_2	$Cov[x_2,x_1]$	$V[x_2]$

Covariance matrix of $X = [x_1, x_2]^T$

Covariance, Variance – Ex 2

- Given *pdf* of a bivariate random variable

$$p(x, y) = \frac{1}{2}x + \frac{3}{2}y, 0 < x, y \leq 1$$

- Compute Cov[x, y], V[x], V[y]

Cov[x, y] = E[xy] – E[x]E[y]

V[x] = Cov[x, x] = $E[x^2]$ – E[x]E[x]

V[y] = Cov[y, y] = $E[y^2]$ – E[y]E[y]

$$E[\mathrm{xy}] = \int_0^1 \int_0^1 xyf\left(x,y\right)dxdy = \int_0^1 \left[\int_0^1 xy\left(\frac{x}{2}+\frac{3y}{2}\right)dy \right] dx$$

$$E[x] = \int_0^1 xf_X\left(x\right)dx$$

$$E[\mathrm{y}] = \int_0^1 yf_Y\left(y\right)dy$$

$$f_X\left(x\right) = \int_0^1 f\left(x,y\right)dy = \int_0^1 \left(\frac{x}{2}+\frac{3y}{2}\right)dy$$

$$E[x^2] = \int_0^1 x^2 f_X\left(x\right)dx$$

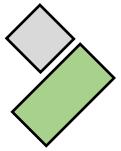
$$f_Y\left(y\right) = \int_0^1 f\left(x,y\right)dx = \int_0^1 \left(\frac{x}{2}+\frac{3y}{2}\right)dx$$

$$E[\mathrm{y}^2] = \int_0^1 y^2 f_Y\left(y\right)dy$$

Exercise – Compute Cov(X, Y)

Let X and Y have the joint density function

$$f(x,y) = \begin{cases} x+y & \text{if } 0 < x, y < 1 \\ \\ 0 & \text{elsewhere .} \end{cases}$$



Exercise

- Show that if X and Y are two r.v., then

$$Cov[ax + b, cy + d] = acCov[x, y],$$

where and a, b, c, d are constants.

$$Cov[ax + b, cy + d]$$

$$= E[(ax + b)(cy + d)] - E[ax + b]E[cy+d]$$

$$= E[acxy + adx + bcy + bd] - (aE[x] + b)(cE[y] + d)$$

$$= acE[xy] + adE[x] + bcE[y] + bd - (acE[x]E[y] + adE[x] + bcE[y] + bd)$$

$$= ac(E[xy] - E[x]E[y])$$

$$= acCov[x,y]$$

Covariance (Multivariate)

- **Definition.** If X and Y are two *multivariate* random variables with states $x \in R^D$ and $y \in R^E$ respectively, the *covariance* between X and Y is defined as

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

Dx1 1xE

Variance

- **Definition (Variance).** The *variance* of a random variable X with states $x \in \mathbb{R}^D$ and a *mean vector* $\mu \in \mathbb{R}^D$ is defined as

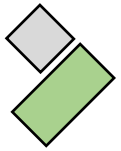
$$\begin{aligned}\mathbb{V}_X[\boldsymbol{x}] &= \mathrm{Cov}_X[\boldsymbol{x}, \boldsymbol{x}] \\ &= \mathbb{E}_X[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_X[\boldsymbol{x}]\mathbb{E}_X[\boldsymbol{x}]^\top \\ &= \begin{bmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \dots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \dots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \dots & \dots & \mathrm{Cov}[x_D, x_D] \end{bmatrix}\end{aligned}$$

variances of the marginals $V[x_i]$

cross-covariance terms $Cov[x_i, x_j]$

Covariance matrix

The *variance* describes the relation between individual dimensions of the random variable



Covariance Matrix

- The *covariance matrix* is *symmetric* and *positive semidefinite* and tells us something about the *spread* of the data
  - On *its diagonal*, the covariance matrix contains the *variances* of the marginals
  - The *off-diagonal entries* are the *cross-covariance* terms $\mathrm{Cov}[x_i, x_j]$

Correlation

- The normalized version of covariance is called the *correlation*
- **Definition.** The *correlation* between two random variables X, Y is given by

$$-1 \leq \mathrm{corr}[x, y] = \frac{\mathrm{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \leq 1$$

- The *correlation matrix* is the covariance matrix of *standardized* random variables, x/σ(x)

Correlation matrix – Ex

$X = (x_1, x_2)$

- $Cov[x_1, x_2] = -0.0477$
- $\sigma(x_1) = 0.4999$
- $\sigma(x_2) = 0.7483315$
- $corr[x_1, x_2] = Cov[x_1,x_2]/(\sigma(x_1)\sigma(x_2))$

$= -0.1223$

	x_1	x_2
x_1	0.2499	-0.0477
x_2	-0.0477	0.5291

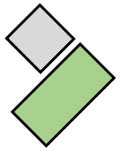
Covariance matrix

	x_1	x_2
x_1	1	-0.1223
x_2	-0.1223	1

Correlation matrix

The *correlation* indicates how two random variables are *related*

- $corr[x, y] > 0$ means that when x grows, then y is also expected to grow
- $corr[x, y] < 0$ means that as x increases, then y decreases



Statistical Independence

- **Definition (Independence).** Two random variables X, Y are *statistically independent* if and only if

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

- Two random variables X and Y are independent if the value of y (once known) does not add any additional information about x (and vice versa)
  - $p(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y})$
  - $p(\boldsymbol{x} \mid \boldsymbol{y}) = p(\boldsymbol{x})$
  - $\mathbb{V}_{X,Y}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{x}] + \mathbb{V}_Y[\boldsymbol{y}]$
  - $\mathrm{Cov}_{X,Y}[\boldsymbol{x}, \boldsymbol{y}] = \mathbf{0}$

Independence – Ex 1

- Given the joint distribution of two r. v. X, Y
- Are X and Y independent? Are X and X + Y independent?

Marginal distributions

p(x)

x	0	1
p(x)	0.7	0.3

p(y)

y	0	1
p(y)	0.7	0.3

p(x, y)		y	
		0	1
x	0	0.5	0.2
	1	0.2	0.1

$p(x = 0, y = 0) = 0.5$
$p(x = 0) = 0.7, p(y = 0) = 0.7$
➔ $p(x, y) \neq p(x)p(y)$

Independence – Ex 1

- Are X and X + Y independent?

x	0	1
p(x)	0.7	0.3

p(x, y)		y	
		0	1
x	0	0.5	0.2
	1	0.2	0.1

p(x+y = 1) = p(x = 0, y = 1) + p(x = 1, y = 0) = 0.2 + 0.2 = 0.4
p(x = 1) = 0.3
p(x = 1, x + y = 1) = p(x = 1, y = 0) = 0.2
➔ $p(x = 1, x + y = 1) \neq p(x = 1)p(x + y = 1)$
➔ X and X + Y independent are not independent

Independence – Ex 2

- Given the joint pdf of a bivariate r. v.

$$p(x,y) = \begin{cases} x+y, & \text{for } 0 < x < 1, 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

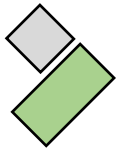
Are X and Y independent?

$$p(x) = \int_0^1 p(x,y)dy = \int_0^1 (x+y)dy = x + ½$$

$$p(y) = \int_0^1 p(x,y)dx = \int_0^1 (x+y)dx = y + ½$$

So, $p(x, y) \neq p(x)p(y)$

➔ X and Y are NOT independent



Theorem

If X and Y are independent, then Cov[x, y] = 0.

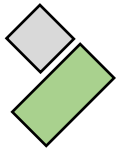
Note that Cov[x, y] = 0 does not imply X and Y are independent.

Independent – Example 3

Let X, Y be r.v with pdf

$$f(x,y)=\begin{cases}\frac{1}{4}, & \text{if } (x,y)\in\{(0,1),(0,-1),(1,0),(-1,0)\}\\ 0, & \text{otherwise}\end{cases}$$

Compute Cov[x,y]. Are X and Y independent?



Conditional Independence

- **Definition.** Two random variables X and Y are *conditionally independent* given Z if and only if

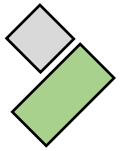
$$p(x, y \mid z) = p(x \mid z)p(y \mid z) \text{ for all } z \in Z,$$

or $p(x \mid y, z) = p(x \mid z)$

where Z is the set of states of random variable Z

Notation: $X \perp\!\!\!\perp Y \mid Z$

Read: *"X is conditionally independent of Y given Z"*



Empirical Means and Covariances

- The *empirical mean* vector is the arithmetic *average* of the observations for each variable, and it is defined as

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

- The empirical covariance matrix is a D×D matrix ($x_n \in R^D$)

$$\boldsymbol{\Sigma} := \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^\top.$$

Three Expressions for the Variance

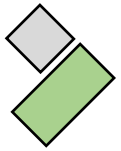
- For a single random variable X

$$\mathbb{V}_X[x] := \mathbb{E}_X[(x - \mu)^2]$$

$$\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$$

$$\frac{1}{N^2}\sum_{i,j=1}^{N}(x_i - x_j)^2 = 2\left[\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2\right]$$

➔ There is an equivalence between the pairwise distances and the distances from the center of the set of points



Sums and Transformations of Random Variables

Consider two random variables X, Y with states $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$.

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{x}+\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{y}] \\
\mathbb{E}[\boldsymbol{x}-\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[\boldsymbol{y}] \\
\mathbb{V}[\boldsymbol{x}+\boldsymbol{y}] &= \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] + \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}] \\
\mathbb{V}[\boldsymbol{x}-\boldsymbol{y}] &= \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] - \mathrm{Cov}[\boldsymbol{y},\boldsymbol{x}] \,.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_X[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] &= \boldsymbol{A}\mathbb{E}_X[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}\,, \\
\mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}] &= \mathbb{V}_X[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_X[\boldsymbol{x}]\boldsymbol{A}^\top
\end{aligned}
$$

Transformations – Ex

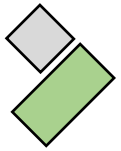
Suppose X = (X_1, X_2) is r. v. with mean zero and covariance matrix $\Sigma=\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}$. Let Y = X_1 + X_2, compute Var[Y].

- Var[Y] = Var[X_1 + X_2] = Var[X_1] + Var[X_2] + 2Cov[X_1, X_2]

= 1 + 2 + 0 = 3.

- Note that if we write Y = $\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}X_1\\ X_2\end{bmatrix} = AX$

Then Var[Y] = ACov[X]A^T = $\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 2\end{bmatrix}\begin{bmatrix}1\\ 1\end{bmatrix}$ = 3



Inner Products of Random Variables

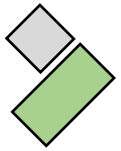
- General Inner Products.

Let V be a vector space and $\langle \cdot , \cdot \rangle : V \times V \to \mathbb{R}$ is a bilinear mapping

$\langle \cdot , \cdot \rangle$ is called an *inner product* on V if

- $\langle \cdot , \cdot \rangle$ symmetric $\langle x , y \rangle = \langle y , x \rangle$
- $\langle \cdot , \cdot \rangle$ positive definite $\langle x , x \rangle > 0, \langle 0 , 0 \rangle = 0$

- **Ex1.** Dot product as inner product
- **Ex2.** Consider $V = \mathbb{R}^2$, $\langle x , y \rangle := x_1y_1 - (x_1y_2 + x_2y_1) + 2x_2y_2$ is an *inner product*.



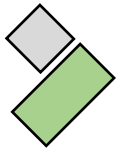
- If we have two uncorrelated random variables X, Y, then

$$V[x + y] = V[x] + V[y]$$

- Define <X, Y> := Cov[x, y],

for zero mean random variables X and Y, we obtain an inner product.

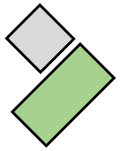
- We see that the covariance is symmetric, positive definite, and linear in either argument.
- The length of a random variable is $||X|| = \sigma(x)$
- The “longer” the random variable, the more uncertain it is; and a random variable with length 0 is deterministic.



- If we look at the angle θ between two random variables X, Y

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

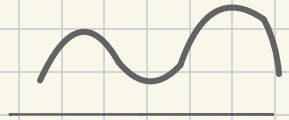
- This means that we can think of correlation as the cosine of the angle between two random variables when we consider them geometrically
- $X \perp Y \Leftrightarrow \langle X, Y\rangle = 0$



Gaussian Distribution – Introduction

- Gaussian distribution has many computationally convenient properties
- There are many other areas of ML that also benefit from using a Gaussian distribution, for example Gaussian processes, reinforcement learning, etc.
- It is also widely used in other application areas (signal processing, control, and statistics, etc.)
- With Gaussian random variables, variable transformations are often not need

Mixture of $p_1(x)$, $p_2(x)$, $0 < \alpha < 1$



$p(x) = \alpha\, p_1(x) + (1-\alpha)\, p_2(x)$ is pdf of X

- $\mu = E(X) = \alpha \mu_1 + (1-\alpha)\mu_2$

$$= \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_{-\infty}^{\infty} x \left(\alpha p_1(x) + (1-\alpha) p_2(x)\right) dx$$

$$= \alpha \underbrace{\int x p_1(x)\, dx}_{\mu_1} + (1-\alpha) \underbrace{\int x p_2(x)\, dx}_{\mu_2}$$

$$= \alpha \int_{-\infty}^{\infty} x^2 p_1(x)\, dx + (1-\alpha) \underbrace{\int_{-\infty}^{\infty} x^2 p_2(x)\, dx}_{-\mu_2}$$

$$V(X) = \sigma^2 = \int_{-\infty}^{\infty} x^2 p(x)\, dx - \mu^2$$

$$= \alpha(\sigma_1^2 + \mu^2) + (1-\alpha)(\sigma_2^2 + \mu_2^2) - [\alpha\mu_1 + (1-\alpha)\mu_2]^2$$

$p_1(x) = N(\mu_1, \sigma_1^2)$ $\quad$ $p_2(x) = N(\mu_2, \sigma_2^2)$

$= N(\mu_1 = 1, \sigma_1^2 = 3)$ $\quad$ $= N(\mu = 1, \sigma_2^2 = 2)$

$\alpha = \frac{1}{3}$

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

$\Rightarrow E(x) = \frac{1}{3} \times 2 + \frac{2}{3} \times 1 = \frac{4}{3}$

$V(x) = \frac{1}{3}(4+3) + \frac{2}{3}(1+2) - \left(\frac{4}{3}\right)^2 = \frac{23}{9}$

Gaussian distribution

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

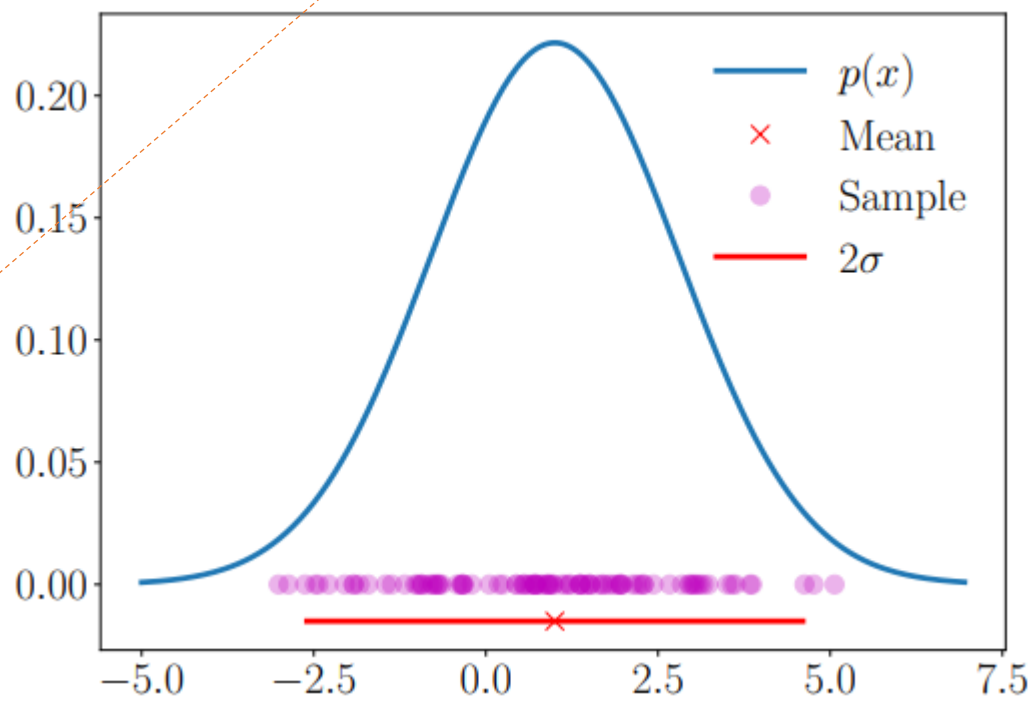
Normal distribution

The *univariate* Gaussian distribution is *fully characterized* by a *mean* μ and a *variance* σ^2

Notation.

$p(x) = N(x \mid \mu, \sigma^2)$

$X \sim N(\mu, \sigma^2)$

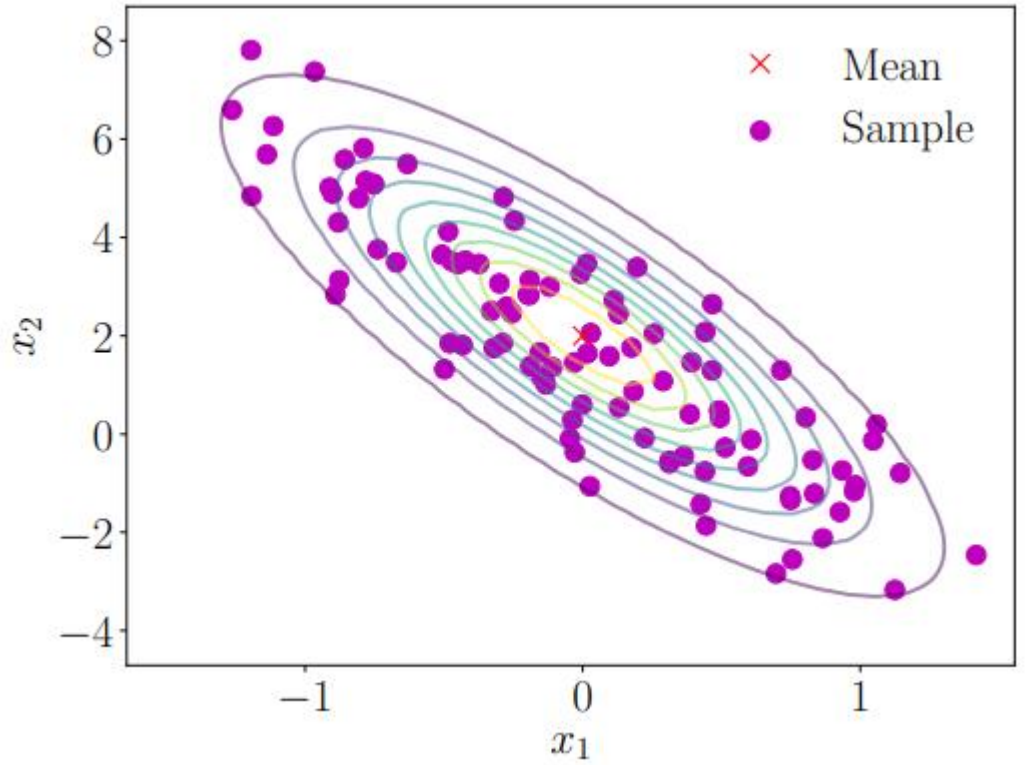


Univariate (one-dimensional) Gaussian

The *multivariate* Gaussian distribution is fully characterized by a *mean vector* μ and a *covariance matrix* Σ

Notation.

$$p(x) = N(x \mid \mu, \Sigma)$$
$$X \sim N(\mu, \Sigma)$$



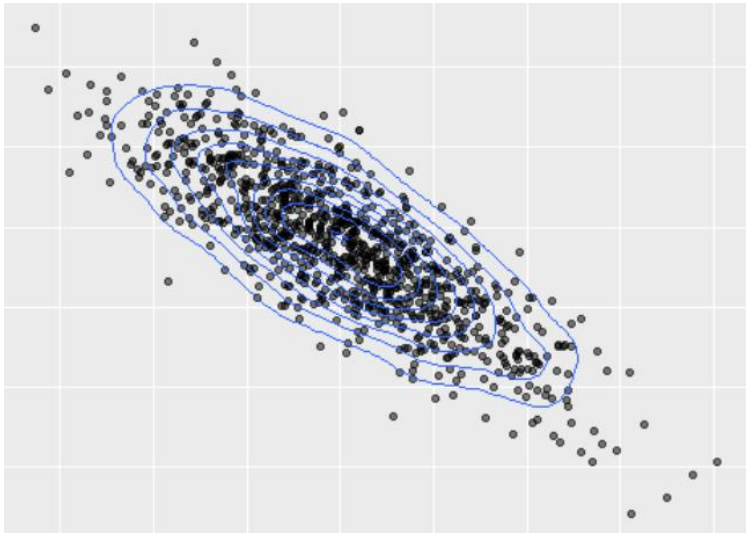
Multivariate (two-dimensional) Gaussian $x \in \mathbb{R}^2$

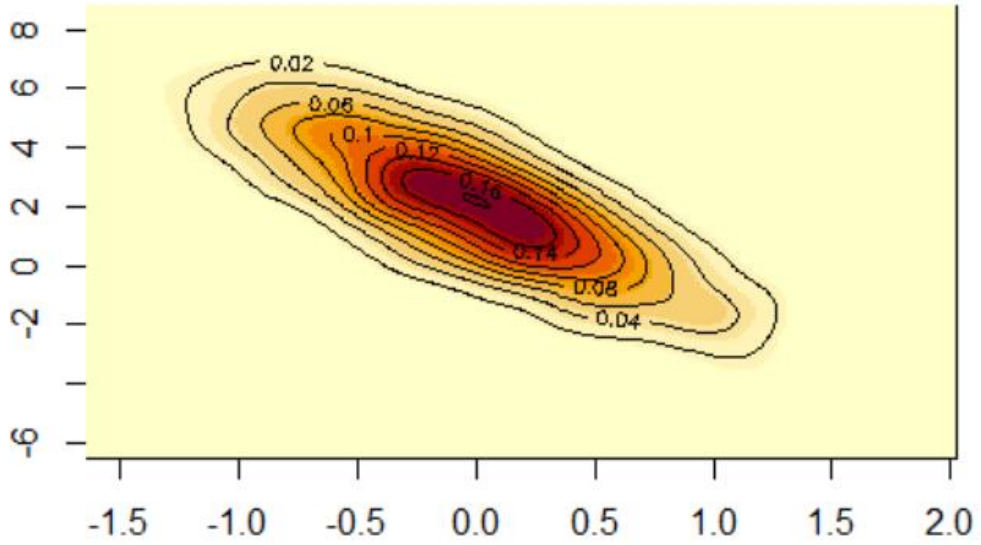
Example

- A sample from the bivariate Gaussian distribution

$$p(x,y) = N\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right)$$

Mean vector Covariance matrix



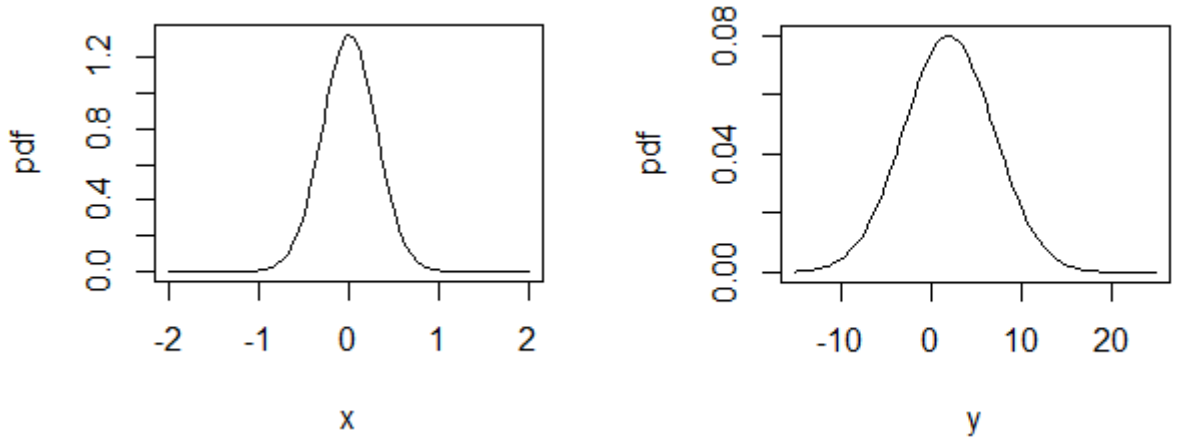


Example (cont.)

$$p(x_1, x_2) = N\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right) = N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

- *Marginal distributions*

are also Gaussians: p(x) = N(0, 0.3) and p(y) = N(2, 5)



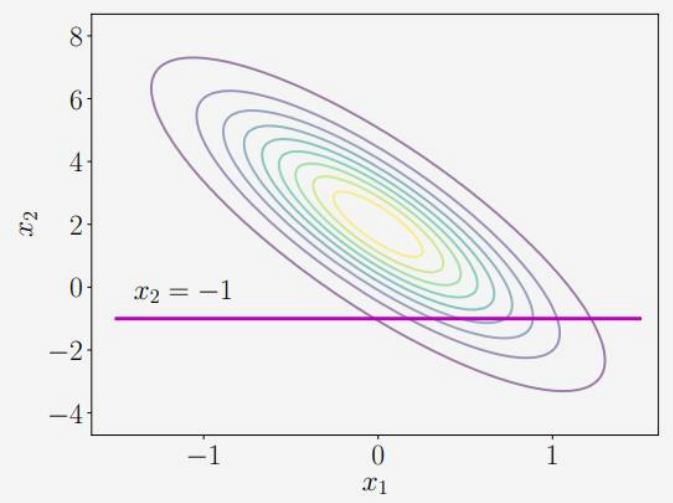
Conditional distributions are also Gaussians

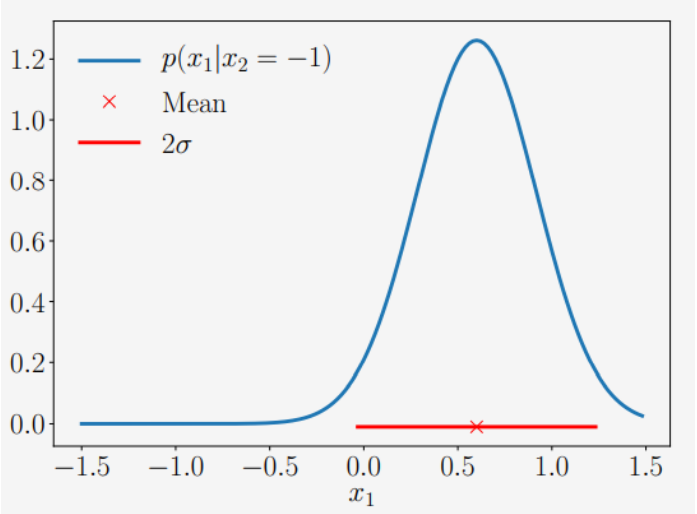
- The conditional distribution p(x | y) is also Gaussian and given by

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

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

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





Example

$$p(x, y) = N\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.3 & -1 \\ -1 & 5 \end{bmatrix}\right) = N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

- Compute the parameters of the univariate Gaussian, conditioned on y = −1

- $\mu_{x \mid y = -1}$ = 0 + (-1)(0.2)(−1 − 2) = 0.6

- $\sigma_{x}^{2}{}_{\mid y = -1}$ = 0.3 − (-1)(0.2)(-1) = 0.1

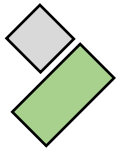
➔ p(x | y = -1) = N(0.6, 0.1)

$$p(x, y) = N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_{x \mid y}, \boldsymbol{\Sigma}_{x \mid y})$$

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

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

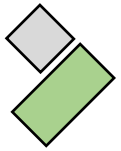


Standard normal distribution

Standard normal distribution is Gaussian with *zero mean vector* and covariance $\Sigma = I$ (identity matrix)

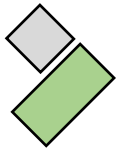
- $p(x) = N(x \mid 0, I)$ or $X \sim N(0, I)$

- Univariate $X \sim N(0, 1)$



Product of Gaussian Densities

- The product of two Gaussians $N(x \mid \mathbf{a}, A)$, $N(x \mid \mathbf{b}, B)$ is a Gaussian distribution scaled by a $c \in \mathbb{R}$, given by $cN(x \mid \mathbf{c}, C)$
  - $C = (A^{-1} + B^{-1})^{-1}$
  - $\mathbf{c} = C(A^{-1}a + B^{-1}b)$
  - $c = N(a \mid b, A + B) = N(b \mid a, A + B)$

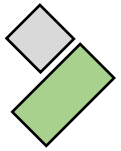


Sums and Linear Transformations

- If X, Y are independent Gaussian random variables, then
- $p(x + y) = N(\mu_x + \mu_y, \Sigma_x + \Sigma_y)$
- $p(ax + by) = N(a\mu_x + b\mu_y, a^2\Sigma_x + b^2\Sigma_y)$
- Given $X \sim N(\mu, \Sigma)$ and y = Ax, then

$$p(y) = N(y \mid A\mu, A\Sigma A^T)$$

- Given $p(y) = N(y \mid Ax, \Sigma)$ and y = Ax, then
  - $p(x) = N(x \mid \mu_x, \Sigma_x)$, where
    - $\mu_x = (A^TA)^{-1}A^Ty$
    - $\Sigma_x = (A^TA)^{-1}A^T\Sigma A(A^TA)^{-1}$



Mixture of two univariate Gaussian densities

- **Theorem.** Consider a *mixture* of two (different) univariate Gaussian densities $p_1(x)$, $p_2(x)$ with weight $0 < \alpha < 1$

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

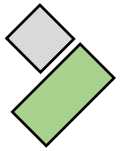
Then,

- The *mean* of the mixture density p(x) is given by

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

- The *variance* of the mixture density p(x) is given by

$$V[x] = \alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2 + \alpha\mu_1^2 + (1 - \alpha)\mu_2^2 - [\alpha\mu_1 + (1 - \alpha)\mu_2]^2$$



Introduction

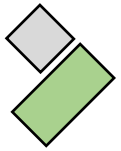
Recall the desiderata for manipulating probability distributions in the machine learning context:

1. There is some “closure property” when applying the rules of probability, e.g., Bayes’ theorem. By closure, we mean that applying a particular operation returns an object of the same type

2. As we collect more data, we do not need more parameters to describe the distribution

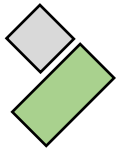
3. Since we are interested in learning from data, we want parameter estimation to behave nicely

➔ *exponential family* meets these properties



Introduction

- The main motivation for *exponential families* is that they have finite-dimensional sufficient statistics. Additionally, *conjugate distributions* are easy to write down, and the *conjugate distributions* also come from an exponential family. From an inference perspective, maximum likelihood estimation behaves nicely because empirical estimates of *sufficient statistics* are optimal estimates of the population values of *sufficient statistics* (the mean and covariance of a Gaussian). From an optimization perspective, the *log-likelihood function* is concave, allowing for *efficient optimization* approaches to be applied.



Bernoulli distribution

- The *Bernoulli distribution* is a distribution for a single *binary* random Bernoulli distribution variable X with state $x \in \{0, 1\}$

  with $p(X = 1) = \mu$, single continuous parameter

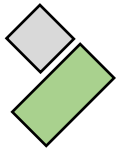
- The *Bernoulli distribution* $\text{Ber}(\mu)$ is defined as

$$p(x \mid \mu) = \mu^x(1 - \mu)^{1-x}, x \in \{0, 1\},$$

$$E[x] = \mu,$$

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

Ex. Bernoulli distribution can be used to model the probability of "heads" when flipping a coin (e.g., $\mu = 0.5$ for a fair coin)



Binomial Distribution

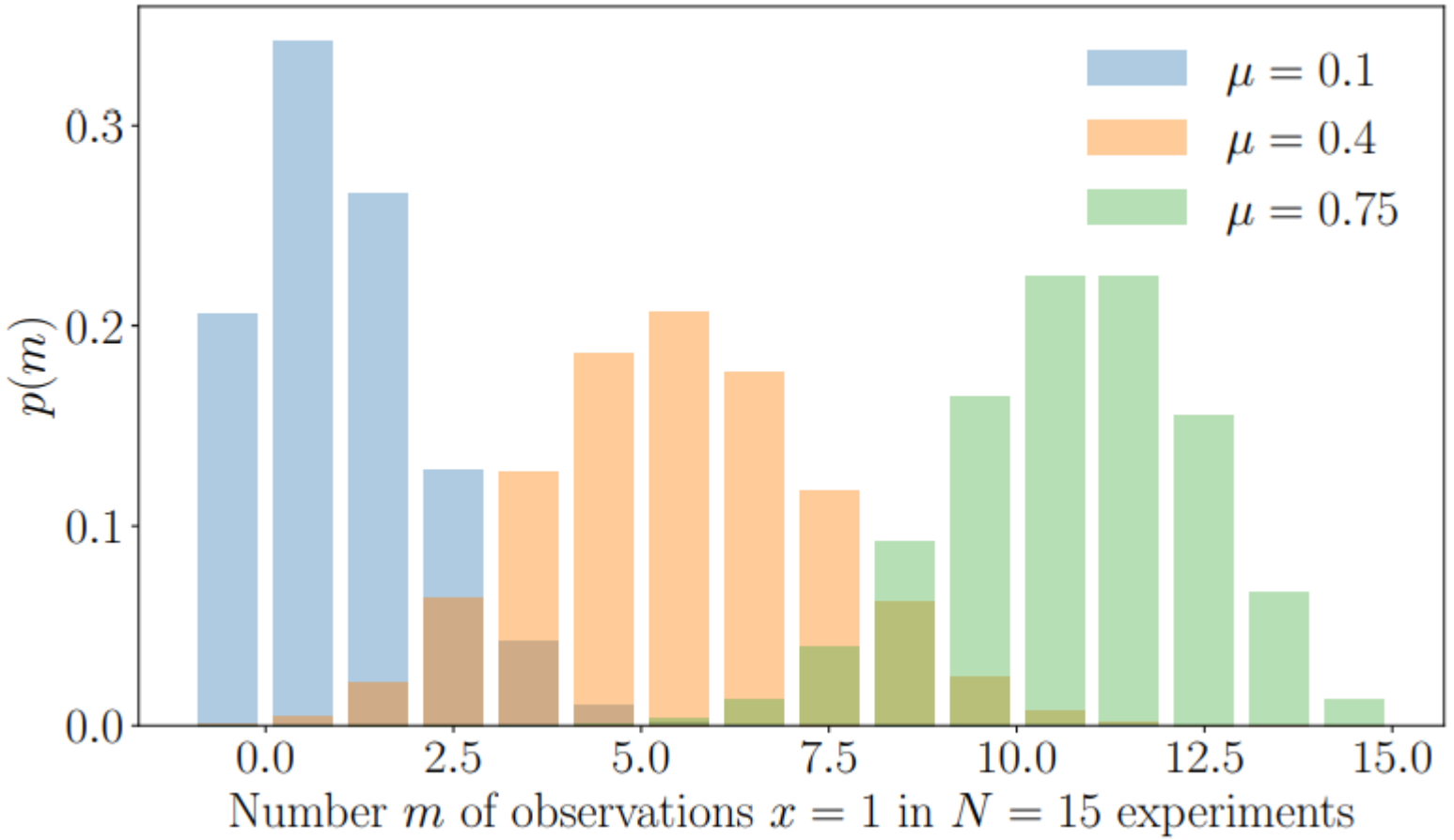
- The *Binomial distribution* is used to describe the probability of *observing m occurrences of X = 1* in a set of N samples from a Bernoulli distribution where $p(X = 1) = \mu \in [0, 1]$.
- The Binomial distribution $\mathrm{Bin}(N, \mu)$ is defined as

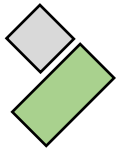
$$p(m \mid N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$E[m] = N\mu,$$

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

Binomial Distribution – example

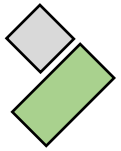




Example

- Consider variable X with state $x \in \{0, 1\}$ where x = 1 means “you *correctly* select a choice in a multiple choice question” and x = 0 means “you answer a question *incorrectly*”
- Given p(X = 1) = 0.3, N = 10, then

$$p(m = 4 \mid N = 10, \mu = 0.3) = \binom{10}{4}(0.3)^4(1 - 0.3)^{10-4} = 0.200$$



Beta Distribution

- The *Beta distribution* is a distribution over a continuous random variable $\mu \in [0, 1]$, which is often used to represent the probability for some *binary event*
- The Beta distribution Beta(α, β) itself is governed by two parameters $\alpha > 0$, $\beta > 0$ and is defined as

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

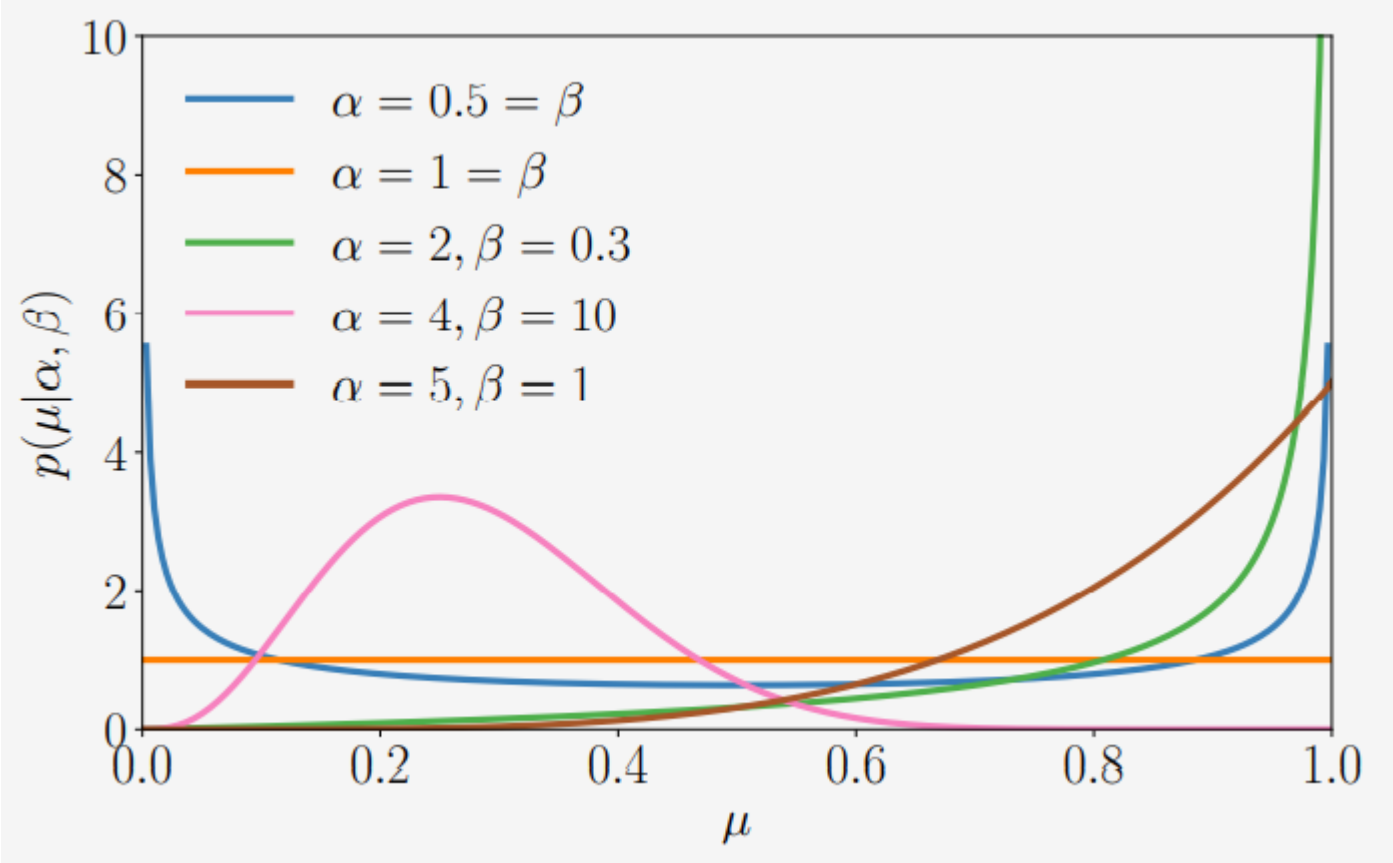
$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta}, \qquad \mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

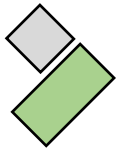
$\Gamma(\cdot)$ is the Gamma function defined as

$$\Gamma(t) := \int_0^\infty x^{t-1} \exp(-x) dx,$$

$$\Gamma(t+1) = t\Gamma(t).$$

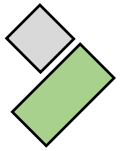
Beta Distribution





Conjugacy

- **Definition (Conjugate Prior).** A *prior* is *conjugate* for the likelihood function if the *posterior* is of the *same form/type* as the prior
- Conjugacy is particularly convenient because we can *algebraically calculate* our *posterior distribution* by updating the parameters of the *prior distribution*



Example (Beta-Binomial Conjugacy)

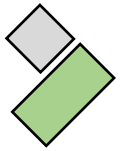
- Consider a Binomial random variable x ~ Bin(N, μ) where

$$p(x \mid N, \mu) = \binom{N}{x} \mu^x (1-\mu)^{N-x}, \quad x = 0, 1, \ldots, N$$

is the probability of finding x times the outcome “heads” in N coin flips, where μ is the probability of a “head”

- Place a *Beta prior* on the parameter μ, μ ~ Beta(α, β), where

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

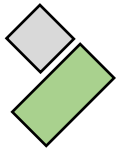


Example (Beta-Binomial Conjugacy)

- We see h heads in N coin flips, we can compute the *posterior distribution on* μ as

$$p(\mu \mid x = h, N, \alpha, \beta) \propto p(x \mid N, \mu)p(\mu \mid \alpha, \beta)$$
$$= \mu^h(1-\mu)^{N-h}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$
$$= \mu^{h+\alpha-1}(1-\mu)^{N-h+\beta-1}$$
$$\propto \text{Beta}(h+\alpha, N-h+\beta),$$

i.e., the *posterior* distribution is a Beta distribution as the *prior*



Beta-Bernoulli Conjugacy

- Let $x \in \{0, 1\}$ be distributed according to the Bernoulli distribution with parameter $\theta \in [0, 1]$, and

$$p(x = 1 \mid \theta) = \theta \text{ or } p(x \mid \theta) = \theta^x(1 - \theta)^{1-x}$$

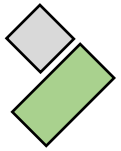
- Let θ be distributed according to $\text{Beta}(\alpha, \beta)$, that is,

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

- We get, $p(\theta \mid x, \alpha, \beta)$

$$\begin{aligned} &= p(x \mid \theta)p(\theta \mid \alpha, \beta) \\ &\propto \theta^x(1 - \theta)^{1-x}\theta^{\alpha-1}(1 - \theta)^{\beta-1} \\ &= \theta^{\alpha+x-1}(1 - \theta)^{\beta+(1-x)-1} \\ &\propto p(\theta \mid \alpha + x, \beta + (1 - x)) \end{aligned}$$

➔ $\text{Beta}(\alpha + x, \beta + (1 - x))$



Sufficient Statistics (ignore)

- Let X be a random variable with distribution $p(x \mid \theta_0)$ given an unknown θ_0. A vector $\varphi(x)$ of statistics is called *sufficient statistics* for θ_0 if they contain all possible information about θ_0
- **Theorem (Fisher - Neyman).** Let X have probability density function $p(x \mid \theta)$. Then the statistics $\varphi(x)$ are *sufficient* for θ if and only if $p(x \mid \theta)$ can be written in the form

$$p(x \mid \theta) = h(x)g_\theta(\varphi(x)),$$

where h(x): a distribution independent of θ,

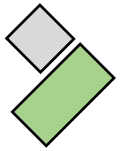
g_θ captures all the dependence on θ via sufficient statistics $\varphi(x)$.

Theorem (Fisher – Neyman)

$$p(x \mid \theta) = h(x)g_{\theta}(\varphi(x))$$

Note.

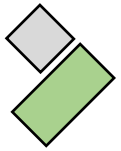
- If p(**x** | **θ**) does not depend on **θ**, then for any function φ, φ(**x**) is trivially a *sufficient statistic*
- If p(**x** | **θ**) is dependent only on φ(**x**) and not **x** itself, φ(**x**) is a *sufficient statistic* for **θ**



Exponential Family – Intro

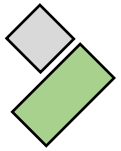
In ML, given a set of data (from an *unknown distribution*), which *distribution* gives the best fit? And as we observe more data, do we need more parameters $\boldsymbol{\theta}$? ➔ The answer is yes in general

Which class of distributions have *finite-dimensional sufficient statistics*, that is, the number of parameters needed to describe them does not increase arbitrarily? ➔ The answer is *exponential family* distributions



Exponential Family – Intro

- There are three possible levels of abstraction when considering distributions
  - Level one: we have a particular named distribution with fixed parameters, e.g., univariate Gaussian N(0, 1)
  - Level two (often in ML): we fix the parametric form and *infer the parameters from data*. For example, we assume a univariate Gaussian $N(\mu, \sigma^2)$ with unknown mean μ and unknown variance σ^2, and use a maximum likelihood fit to determine the best parameters (μ, σ^2)
  - A third level of abstraction is to consider families of distributions, and one of them is the *exponential family*



Exponential Family

- **Definition.** An *exponential family* is a family of probability distributions, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^D$, of the form

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = h(\mathbf{x})\exp(\langle \boldsymbol{\theta}, \varphi(\mathbf{x}) \rangle - A(\boldsymbol{\theta})),$$

where $\varphi(\mathbf{x})$ is the vector of sufficient statistics

- **Note.** Considered as distributions of the form

$$p(\mathbf{x} \mid \boldsymbol{\theta}) \propto \exp(\boldsymbol{\theta}^T \varphi(\mathbf{x}))$$

(use standard dot product: $\langle \boldsymbol{\theta}, \varphi(\mathbf{x}) \rangle = \boldsymbol{\theta}^T \varphi(\mathbf{x})$)

Example (Gaussian as Exponential Family)

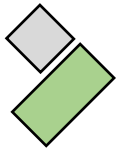
- Consider the univariate Gaussian distribution $N(\mu, \sigma^2)$
- Let $\varphi(x) = [x, x^2]^T$, $\boldsymbol{\theta} = [\theta_1, \theta_2]^T$
- By using the definition of the *exponential family*,

$$p(x \mid \boldsymbol{\theta}) \propto \exp(\theta_1 x + \theta_2 x^2)$$

- Setting $\boldsymbol{\theta} = [\mu/\sigma^2, -1/2\sigma^2]^T$
- We have $p(x \mid \boldsymbol{\theta}) \propto \exp(\mu x/\sigma^2 - x^2/2\sigma^2)$

$$\propto \exp(-(x - \mu)^2 /2\sigma^2)$$

➔ $N(\mu, \sigma^2)$ is a member of the *exponential family* with sufficient statistics $\varphi(x) = [x, x^2]^T$, and natural parameters θ_1, θ_2



Bernoulli as Exponential Family

- Recall the Bernoulli distribution

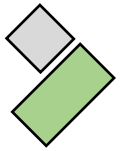
$$p(x \mid \mu) = \mu^x(1 - \mu)^{1-x},\ x \in \{0, 1\}$$

- This can be written in *exponential family* form

$$p(x \mid \mu) = \exp[\log(\mu^x(1 - \mu)^{1-x})]$$
$$= \exp[x\log\mu + (1 - x)\log(1 - \mu)]$$
$$= \exp[x\log\mu - x\log(1 - \mu) + \log(1 - \mu)]$$
$$= \exp[x\log\frac{\mu}{1-\mu} + \log(1 - \mu)]$$

The last line can be identified as being in *exponential family*

$h(x) = 1$, $\theta = \log(\mu/(1-\mu))$, $\varphi(x) = x$, $A(\theta) = -\log(1 - \mu) = \log(1 + \exp(\theta))$.



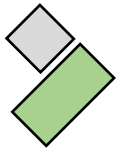
sigmoid

- The relationship between parameter μ and the natural parameter θ is known as the *sigmoid* or *logistic function*

$$\mu = \frac{1}{1 + \exp(-\theta)}$$

- Observe that $\mu \in (0, 1)$ but $\theta \in \mathbb{R}$, and therefore the *sigmoid function* squeezes a real value into the range (0, 1)

➔ This property is useful in ML, for example it is used in *logistic regression* as well as a nonlinear activation functions in *neural networks*



THANKS