Introduction to Deep Learning for Speech and Language Processing
Exercise Sheet 3: Math for Machine Learning

Thang Vu

4th November 2025

Probabilities

Exercise 1.
What is the probability of throwing a fair dice and getting a number greater than 3?
Exercise 2.
What is the probability of throwing two fair dices and getting a sum that is greater than 8?
Exercise 3.
Given two random variables X and Y and the following table that summarizes how many times the two events occur:

Y/X	x_1	x_2	x_3	x_4
y_1	4	3	5	10
y_2	1	8	3	2

Your tasks are:

- Compute the joint probability $P(X = x_2, Y = y_1)$ and $P(Y = y_2, X = x_1)$
- Compute the marginal probability $P(X = x_2)$
- Compute the marginal probability $P(Y = y_1)$
- Compute the conditional probability $P(X = x_3 \mid Y = y_2)$
- Compute the conditional probability $P(Y = y_2 \mid X = x_3)$

Exercise 4.
We happen to have a dice where its faces do not show up equally. Let X be the random variable representing the value of the dice after rolling and we know that:

- $\mathrm{P(X = 1) = 0.1}$
- $\mathrm{P(X = 2) = 0.1}$
- $\mathrm{P(X = 3) = 0.1}$
- $\mathrm{P(X = 4) = 0.2}$
- $\mathrm{P(X = 5) = 0.2}$
- $\mathrm{P(X = 6) = 0.3}$

You tasks are:

- Compute the expected value $E[X]$.
- Calculate the variance $Var[X]$.

Exercise 5.

Let's consider two events that are dependent from each other. The events will either take place ('1') or be canceled ('0'). Let X_1 be the random variable representing the outcome of the first event and X_2 be the random variable representing the outcome of the second event. The following table presents the joint probability of these two events:

x_1	x_2	$\mathrm{P}(\mathrm{X}_1 = x_1, X_2 = x_2)$
0	0	0.2
0	1	0.1
1	0	0.1
1	1	0.6

You tasks are:

- Compute the expected values $E[X_1]$ and $E[X_2]$.
- Calculate the covariance $Cov(X_1, X_2)$ between the two events.

Optimization

Exercise 6.

(1) Given $f(x) = (x+2)^2 + 3$ with $x \in \mathbb{R}$ perform one step of gradient descent starting from $x^0 = 1$ with a learning $\eta = 0.1$.

(2) Given $f(x) = x_1^3 + 2x_2^2 - 1$ with $x \in \mathbb{R}^2$ perform one step of gradient descent starting from $x^0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ with a learning $\eta = 0.1$.