Probability and Entropy

Dominik Tschimmel

February 7, 2021

Abstract

Probability theory is not often a core part of the mathematical education in physics, and is usually taught along statistical physics. As such, it can happen that not much detail is spent on the concepts of probability theory, which has its uses in many areas of applied sciences and applied mathematics. This report tries to give a concise and coherent introduction to the basic concepts of probability theory and the information theoretic view of entropy. It contains only standard results and is inspired by lectures on these topics (not publicly available). For the probability section, any introductory book on probability theory and stochastics will do. The second section loosely follows [CT12].

Contents

1 Probability

1.1 Fundamental concepts

At first glance, the concept of probability needs no further introduction. Intuitively, it is described as the likelihood of a specific event to happen. Yet, this is no precise definition. Although bulky at first, the axiomatic construction of probability allows to derive precise statements for random processes. That is, even randomness is not devoid of order and structure. Before we can define probability spaces, we need to introduce some technical concepts:

Definition 1.1.
Let Ω be a set and $\Sigma \subseteq \mathcal{P}(\Omega)$ be a subset of the power set of Ω. It is called a **σ-algebra** if:

i) $\Omega \in \Sigma$,

ii) $A \in \Sigma \quad \Rightarrow \quad \Omega \setminus A \in \Sigma$,

iii) $A_k \in \Omega$ for all $k \in I \subseteq \mathbb{N}$, then $\bigcup_{k\in I} A_k \in \Sigma$.

The second technical definition we need is the following:

Definition 1.2.
Let Σ be a σ-algebra of Ω. A map $p\colon \Sigma \longrightarrow [0,\infty]$ is called **measure** if

i) $p(\emptyset) = 0$,

ii) p is **additive**, i.e. for disjunct $A_k \in \Sigma$ for $k \in I \subseteq \mathbb{N}$ it holds that

$$p\left(\bigcup_{k\in I} A_k\right) = \sum_{k\in I} p(A_k)\ .$$

If $p\colon \Sigma \longrightarrow [0,1]$ and $p(\Omega) = 1$, it is called a **probability measure**.

With these technical definitions out of the way, we may define the object of interest:

Definition 1.3.
A **probability space** is a tuple (Ω, Σ, p) consisting of

- a **sample space** Ω of all possible outcomes,
- a σ-algebra Σ, called **set of events /events**,
- a **probability measure** $p\colon \Sigma \longrightarrow [0,1]$.

Remark 1.4.
Note that the result can only be an outcome. Yet, the events are more general, allowing to ask for the probability of a combination of outcomes.

Example 1.5.
Maybe the most basic example is a fair coin toss. Here, the outcomes are either heads h or tails t, so $\Omega = \{h, t\}$. The possible events are $\emptyset, \{h\}, \{t\}, \{h, t\} = \Omega$. The probability measure is uniquely defined by $p(\{h\}) = p(\{t\}) = 1/2$. Note that $\emptyset$ is impossible, as $p(\emptyset) = 0$. The event $\{h, t\} = \{h\} \cup \{t\} = \Omega$ has the meaning of either heads or tails as outcome, and is guaranteed to occur, since $p(\Omega) = 1$.

1.2 Conditional probabilities

The concept of conditional probability can be be summarized as follows: assume that one knows that the conditions for a certain event B are already met, what is then the probability of event A to occur. This seems very technical, so a real world example is a good starting point to illustrate the concepts and notions better:

Example 1.6.
We consider a medical test, designed to test for a single disease. There are four possible outcomes: The patient has the disease $\boldsymbol{D}$, the patient does not have the disease $\overline{\boldsymbol{D}}$, the test is positive $\boldsymbol{P}$ and the test is not positive $\overline{\boldsymbol{P}}$.
Note that the probability of the test being positive for an arbitrary person should be different than the probability of the test being positive if the person has the disease. That is, if the test is any good in detecting the disease, the probability for a positive result should be higher for patients with the disease. Formally one says, that given the **condition** D, what is the probability of P, and writes $p(P \mid D)$.
For medical tests, there are two important parameters. The probability $p(P \mid D)$, called the sensitivity, i.e. how well does the test detect the disease. And the probability $p(\overline{P} \mid \overline{D})$, called the specificity, i.e. how well does a negative result indicate absence of the disease.

Definition 1.7.
Let $A, B \in \Sigma$ be events, then $p(A \cap B)$ is called the **joint probability**. Let B be a condition, then the **conditional probability** (of A under the condition B) is defined as

$$p(A \mid B) \coloneqq \frac{p(A \cap B)}{p(B)} \ .$$

One may understand this definition intuitively by saying one divides out/normalizes to the probability it takes to achieve the condition.

Theorem 1.8 (Bayes' theorem).
It holds that

$$p(A \mid B) = \frac{p(B \mid A)\ p(A)}{p(B)} \ .$$

Proof 1.9.
This follows from the definition of conditional probabilities:

$$p(A \mid B)p(B) = p(A \cap B) = p(B \cap A) = p(B \mid A)p(A) \ . \qquad \square$$

Corollary 1.10.
Let $\{A_j\}$ be a decomposition of Ω, i.e. $A_i \cap A_j = \emptyset$ for all $i \neq j$ and $\bigcup_j A_j = \Omega$. Then it holds that

$$p(B) = \sum_j p(B \mid A_j)p(A_j) \ .$$

Proof 1.11.
Again, this follows from the definition of conditional probabilities and the assumptions (allowing to use the additivity of the probability measure):

$$\begin{aligned}\sum_j p(B \mid A_j)p(A_j) = \sum_j p(B \cap A_j) = p\left(\bigcup_j (B \cap A_j)\right) = p\left(B \cap \bigcup_j A_j\right) = p(B \cap \Omega) \\ = p(B) \ .\end{aligned}$$

$\square$

1.3 Probability distributions

Working out the details of the probability spaces can be cumbersome and rather abstract. To calculate properties, it is desirable to translate the probability space into a model that works with numbers, hence the definition of random variables:

Definition 1.12.
A **random variable** is an injective measureable function $X : \Omega \longrightarrow \mathbb{K}$, where $\mathbb{K}$ is usually $\mathbb{R}$ or $\mathbb{C}$. If $X(\Omega)$ is countable, it is called **discrete**, otherwise it is called **continuous**.

Loosing the flexibility for constructions, the σ-algebra of events offered, we gain the computational power of analysis, as we will see, to derive useful properties. To be a translation of a probability space, the concept of probabilities has to be defined for random variables. For discrete random variables, this is straightforward. However, the continuum harbors some technical difficulties, preventing that there is a continuous analogue of the probability for all random variables. Yet, with a modified concept, this can be fixed.

Definition 1.13.
Let $X\colon \Omega \longrightarrow \mathbb{R}$ be a real random variable. If X is discrete, the **probability mass function (PMF)** is defined by

$$p\colon \mathbb{R} \longrightarrow [0,1]\ , \quad p(x) \equiv p(X = x) \coloneqq p(\{\omega \in \Omega \mid X(\omega) = x\})\ .$$

If X is continuous, the **cumulative distribution function (CDF)** is defined as

$$F\colon \mathbb{R} \longrightarrow [0,1]\ , \quad F(x) \coloneqq p(X \leq x) \equiv p(\{\omega \in \Omega \mid X(\omega) \leq x\})\ .$$

Let X be continuous. A function $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ is called **probability density function (PDF)**, if

$$p(X \in [a,b]) = \int_a^b f(x)\ dx\ .$$

Remark 1.14.
The PDF need not be restricted to $[0,1]$ as the probability is given by the integration. In fact, using measure theory to construct the integral (Lebesgue integral), the function f may be infinite on a null set, so in particular on finitely many points.

While the CDF exists for all continuous random variable and are more well behaved, it can happen that there is no PDF. Generally speaking however, the existence of PDFs usually is desirable. Not only is the concept closer to the PMF and thus to the probability of outcomes, but it also simplifies calculations. As this is only a short introduction, we will take the liberty and show the statements only for the cases where there is a PDF (which is the case for most applications).

Corollary 1.15.
Let $X\colon \Omega \longrightarrow \mathbb{R}$ be a continuous real random variable with PDF f and CDF F. Then it holds that

$$F(x) = \int_{-\infty}^{x} f(y)\ dy\ .$$

If the PDF f is continuous, it holds that $f(x) = \frac{d}{dx}F(x)$.

Proof 1.16.
For the first equality, we use the definition of the PDF and calculate:

$$F(x) = p(X \leq x) = p(X \in [-\infty, x]) = \int_{-\infty}^{x} f(y)\ dy\ .$$

Assuming continuity of the PDF f, the second clam is the result of the fundamental theorem of calculus. □

Lemma 1.17.
Let $X\colon \Omega \longrightarrow \mathbb{R}$ be a continuous random variable with CDF $F_X(x)$ and PDF $f_X(x)$. Let $g\colon \mathbb{R} \longrightarrow \mathbb{R}$ be an invertible function. Then the random variable $Y = g(X)$ has the CDF

$$F_Y(y) = \begin{cases} F_X(g^{-1}(y)) & \text{if } g \text{ is mon. increasing} \\ 1 - F_X(g^{-1}(y)) & \text{if } g \text{ is mon. decreasing} \end{cases}$$

If g is also differentiable (i.e. it is an diffeomorphism), then Y has the PDF

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

Proof 1.18.
Let g be mon. increasing first, then

$$F_Y(y) \equiv p(Y \leq y) = p(g(X) \leq y) = p(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \ .$$

If g is mon. decreasing it holds that $g(a) \leq b \ \Leftarrow \ a \geq g^{-1}(b)$ for all $a, b \in \mathbb{R}$. Then we calculate:

$$\begin{aligned} F_Y(y) &\equiv p(Y \leq y) = p(g(X) \leq y) = p(X \geq g^{-1}(y)) = 1 - p(X < g^{-1}(y)) \\ &= 1 - p(X < g^{-1}(y)) \pm p(X = g^{-1}(y)) \\ &= 1 - p(X \leq g^{-1}(y)) + p(X = g^{-1}(y)) = 1 - p(X \leq g^{-1}(y)) = \\ &= 1 - F_X(g^{-1}(y)) \ . \end{aligned}$$

Note that we used that $p(X = g^{-1}(y)) = 0$. This is, since the random variable has a PDF, such that indeed

$$p(X = g^{-1}(y)) = p(X \in [g^{-1}(y), g^{-1}(y)]) = \int_{g^{-1}(y)}^{g^{-1}(y)} f(x)\ dx = 0 \ .$$

For the PDF, let g be mon. increasing first. We calculate:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = \frac{d}{dy(x)} F_X(x) = \frac{1}{g'(x)} \frac{d}{dx} F_X(x) \\ &= \frac{f_X(x)}{g'(x)} = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|} \ . \end{aligned}$$

For mon. decreasing g, we obtain:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy}(1 - F_X(g^{-1}(y))) = \frac{d}{dy(x)}(1 - F_X(x)) = \frac{1}{g'(x)} \frac{d}{dx}(1 - F_X(x)) \\ &= -\frac{f_X(x)}{g'(x)} = -\frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|} \ . \end{aligned}$$

Observe that for a mon. decreasing g it holds that $g' \leq 0$, such that we could trade the minus sign for the absolute value of the denominator in the last step. □

1.4 Expected values, variance and higher moments

Assume that $X\colon \Omega \longrightarrow \mathbb{R}$ is a discrete random variable with finite Ω. Assuming we repeat the random experiment N times and assume that the probability is perfectly matched in the results, i.e. $n(\omega) = N \cdot p(\omega)$ is the number of the outcome ω. Then, we obtain as average the value

$$\overline{X} = \frac{1}{N} \sum_{\omega \in \Omega} n(\omega) X(\omega) = \frac{1}{N} \sum_{\omega \in \Omega} N p(\omega) X(\omega) = \sum_{\omega \in \Omega} p(\omega) X(\omega) \ .$$

This means that half of the resulting values $X(\omega)$ of the experiment will be smaller than $\overline{X}$ and half of the values will be larger than $\overline{X}$. This motivates the following definition:

Definition 1.19.
Let $X\colon \Omega \longrightarrow \mathbb{R}$ be a random variable. The **expected value** $E(X)$ is defined as follows:

$$E(X) \coloneqq \begin{cases} \displaystyle\sum_{\omega \in \Omega} p(\omega) X(\omega) & \text{if } X \text{ is discrete} \\ \displaystyle\int_{\Omega} p(\omega) X(\omega) \, d\omega & \text{if } X \text{ is continuous} \end{cases}$$

Remark 1.20.
The notation for the expected value for continuous variables is suggestive here, to resemble the result that is used in most calculations. In fact, a more mathematical notation would be $\int_\Omega X(\omega) \, dp(\omega)$ to indicate that in fact the Lebesgue integral w.r.t. the measure $p(\omega)$ is used. For this, one needs that Ω is a Borel-σ-algebra and $\int_\Omega |X(\omega)| \, dp(\omega) < \infty$ [Geo07, Satz 4.12].

Corollary 1.21.
The expected value is linear

$$E(\alpha X + Y) = \alpha E(X) + E(Y)) \ ,$$

monotone

$$E(X) \leq E(Y) \qquad \forall X \leq Y \ ,$$

i.e. $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. Finally, for any constant value α it holds that the expected value acts trivially:

$$E(\alpha) = \alpha \ .$$

Proof 1.22.
Linearity and monotony follow from basic properties of sums (limits) and (Lebesgue) integrals, together with the fact that $p(\omega) \geq 0$.

For the last property, we calculate (using that probabilities sum/integrate to one):

$$E(\alpha) = \sum_{\omega \in \Omega} p(\omega)\alpha = \alpha \sum_{\omega \in \Omega} p(\omega) = \alpha \ .$$

$$E(\alpha) = \int_{\Omega} \alpha p(\omega) \ d\omega = \alpha \int_{\Omega} p(\omega) \ d\omega = \alpha \ . \qquad \square$$

Proposition 1.23 (See [Geo07, Korollar 4.13] for the proof)**.**
Let $X \colon \Omega \longrightarrow \mathbb{R}$ be a continuous random variable with PDF $f_X(x)$, then the expected value is given by

$$E(X) = \int_{\mathbb{R}} f_X(x) x \ dx \ .$$

Let $g \colon \mathbb{R} \longrightarrow \mathbb{R}$ be an integrable function, then it holds that

$$E(g(X)) = \int_{\mathbb{R}} f_X(x) g(x) \ dx \ .$$

For a discrete random variable, we can also "untie" the expected value from the probability space. This may be the case if the model one is interested in is constructed in terms of a random variable and a PMF, without concrete realization of a probability space.

Corollary 1.24.
Let $X \colon \Omega \longrightarrow \mathbb{R}$ be a discrete random variable, then it holds that

$$E(X) = \sum_{x \in X(\Omega)} p(x) x \ .$$

Let $g \colon \mathbb{R} \longrightarrow \mathbb{R}$ be a function, then it holds that

$$E(g(X)) = \sum_{x \in X(\Omega)} p(x) g(x) \ .$$

Proof 1.25.
Recall that $p(x)$ is defined as follows:

$$p(x) = p(X(\omega) = x) = p(\{\omega \in \Omega \mid X(\omega) = x\}) = p(\omega) \ .$$

Thus:

$$E(X) = \sum_{\omega \in \Omega} p(\omega) X(\omega) = \sum_{x \in X(\Omega)} p(x) x \ .$$

For the second claim, we start again with the basic definition:

$$E(g(X)) = \sum_{\omega \in \Omega} p(\omega) g(X(\omega)) = \sum_{x \in X(\Omega)} p(x) g(x) \ . \qquad \square$$

Remark 1.26.
Form a physical point of view, one can attach an intuitive meaning to the expected value. Consider the PMF/PDF as distribution of masses (mass point/continuous body). Then the formulas (see proposition 1.23 and corollary 1.24) for the expected values are precisely the formulas for the center of mass. Furthermore, the transformation behavior makes perfect sense. Changing the position of the masses (i.e. $x \curvearrowright g(x)$), but keeping the masses the same (i.e. $p(x) \curvearrowright p(x)$) results in the center of mass for the masses shifted to the position $g(x)$ (i.e. asking for $E(g(X))$). So the expected value may be regarded as the "**center of probability**".

As first approach to understand a probability distribution, the expected value is a good tool. However, following the analogy to physics, knowing the center of mass does not tell, how far the masses extend, i.e. how much the probability distribution spreads out. See for example figure 1, where both PDFs have the same expected value, but have different widths.

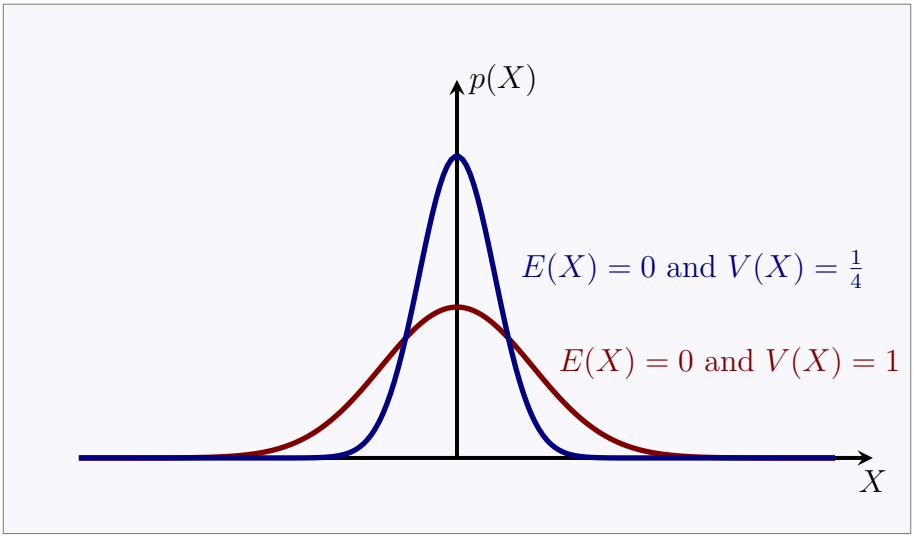

Figure 1: Two normal distributions with the same expected value but with different variances.

In physics, there is a natural quantity to measure the spreading of a mass configuration, the moment of inertia. It measures how much torque (the pendent to force for angular movement) is needed to create angular movement. The further the masses extend outward form the center of rotation and the larger the mass is, the larger is the moment of inertia. A mass point contributes to the moment of inertia with it mass m and its distance to the center of rotation r by mr^2. Again, we interpret the PMF as mass points and the PDF as mass density. As center of rotation (as is the case in a physical system without external forces) we take the center of probability, i.e. the expected value $E(X)$. The distance to this center is given by $X - E(X)$. For the "**moment of probability-inertia**" we obtain (formulas for momenta of inertia):

$$I_X = \sum_{x \in X(\Omega)} p(x)(x - E(X))^2 = E((X - E(X))^2) \ .$$

And for a continuous random variable:

$$I_X = \int_{\mathbb{R}} f_X(x)(x - E(X))^2 \, dx = E((X - E(X))^2) \; .$$

This is the motivation for the following definition:

Definition 1.27.
Let $X : \Omega \longrightarrow \mathbb{R}$ be a random variable. The **variance** is defined as

$$V(X) \coloneqq E((X - E(X))^2) \; .$$

The **standard deviation** of X is defined as $\sigma_X = \sqrt{V(X)}$, thus σ_X^2 is an other common notation for the variance.

Both the variance and the standard deviation measure how far the probability distribution spreads outward from the expected value. Yet, there is a subtle difference between the two, justifying the additional name for the square root of the variance. The variance is motivated by the moment of inertia. However, this introduces the square of the random variable, so the variance has and probability distribution have different dimensions. For better comparability, it is useful to "undo" the square by taking the square root of the variance.

Proposition 1.28.
The variance has the following properties:

i) $V(X) = E(X^2) - E(X)^2$

ii) $V(X) = E((X - a)^2) - (E(X) - a)^2 \qquad \forall \; a \in \mathbb{R}$

iii) $V(aX + b) = a^2 V(X)$

iv) $V(X) \geq 0$ *and* $V(X) = 0 \quad \Rightarrow \quad X = E(X) = const. \in \mathbb{R}.$

Proof 1.29.
We use the properties of the expected value from corollary 1.21 in the following calculations. For the first property, we calculate:

$$\begin{aligned} E(X^2) - E(X)^2 &= E(X^2) - 2E(X)E(X) + E(X)^2 = E(X^2 - 2XE(X) + E(X)^2) \\ &= E((X - E(X))^2) = V(X) \; . \end{aligned}$$

For the second property, we use the first property:

$$\begin{aligned} E((X - a)^2) - (E(X) - a)^2 &= E(X^2 - 2aX + a^2) - E(X)^2 - a^2 + 2aE(X) \\ &= E(X^2) - 2aE(X) + a^2 - E(X)^2 - a^2 + 2aE(X) \\ &= E(X^2) - E(X)^2 = V(X) \; . \end{aligned}$$

The third property follows from:

$$V(aX + b) = E((aX + b)^2) - E(aX - b)^2$$

$$
\begin{aligned}
&= E(a^2X^2 + b^2 - 2abX) - (aE(X) - b)^2 \\
&= a^2E(X^2) + b^2 - 2abE(X) - a^2E(X)^2 - b^2 + 2abE(X) \\
&= a^2E(X^2) - a^2E(X)^2 = a^2(E(X^2) - E(X)^2) \\
&= a^2V(X)
\end{aligned}
$$

Using as random variable $Y = 0$, it holds that $(X - E(X))^2 \geq Y$ and thus it follows from the monotony of the expected value that

$$V(X) = E((X - E(X))^2) \geq E(Y) = 0 \ .$$

Now, let $V(X) = 0$. It holds that $(X - E(X))^2 \geq 0$ and $(X - E(X))^2 = 0$ only if $X = E(X)$. In the discrete case, we calculate:

$$0 = V(X) = E((X-E(X))^2) = \sum_{x \in X(\Omega)} p(x)(x-E(X))^2 \quad \Rightarrow \quad p(x) = 0 \quad \forall x \neq E(X) \ .$$

Since $\sum_{x \in X(\Omega)} p(x) = 1$ it follows that $p(E(X)) = 1$ and thus $X(\omega) = E(X)$ for all $\omega \in \Omega$, i.e. $X = E(X)$. In the continuous case, we calculate:

$$0 = V(X) = E((X - E(X))^2) = \int_\Omega (X(\omega) - E(X))^2 \ dp(\omega) \ .$$

Again, since the probability measure is positive, $X(\omega) = E(X)$ for (almost) all $\omega \in \Omega$ (note that values of null sets are not specified and do not enter probability). □

As a measure for the spreading of the probability distribution, the variance gives so far a qualitative idea of how unlikely outliers are, i.e. events that are far away from the expected value. Yet, one can quantify this notion and construct an upper bound for outliers. Before we can proof this upper bound, we need the following lemma:

Lemma 1.30.
Let $\chi_A \colon \mathbb{R} \longrightarrow \{0, 1\}$ be the indicator function of $A \subseteq \mathbb{R}$, i.e. $\chi_A(x) = 1$ if $x \in A$ else if $\chi_A(x) = 0$. For a random variable $X \colon \Omega \longrightarrow \mathbb{R}$ with PMF/PDF it holds that:

$$p(A) \equiv p(x \in A) = E(\chi_A(X)) \ .$$

Proof 1.31.
For a discrete random variable, we calculate:

$$E(\chi_A(X)) = \sum_{x \in X(\Omega)} p(x)\chi_A(x) = \sum_{x \in X(\Omega) \cap A} p(x) = p(x \in A) \ .$$

For a continuous random variable with PDF f_X we calculate:

$$E(\chi_A(X)) = \int_{\mathbb{R}} f_X(x)\chi_A(x) \ dx = \int_A f_X(x) \ dx = p(x \in A) \ .$$ □

Proposition 1.32.
Let $X\colon \Omega \longrightarrow \mathbb{R}$ be a random variable. Then, the probability of X to deviate from $E(X)$ by r is bounded form above by

$$p(|x - E(X)| \geq r) \leq \frac{V(X)}{r^2} \ .$$

Proof 1.33.
Define function $g, h\colon \mathbb{R} \longrightarrow \mathbb{R}$ by

$$g(x) \coloneqq \begin{cases} 1 & |x - E(X)| \geq r \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad h(x) \coloneqq \frac{1}{r^2}(x - E(X))^2 \ .$$

Then for all $x \in \mathbb{R}$ it holds that $g(x) \leq h(x)$, so $g(X) \leq h(X)$. By the monotony of the expected value, this means that $E(g(X)) \leq E(h(X))$. Observe that g is the indicator function for the set $\{x \in \mathbb{R} \mid |x - E(X)| \geq r\}$. Thus, with lemma 1.30 we obtain:

$$\begin{aligned} \frac{V(X)}{r^2} &= \frac{1}{r^2} E((X - E(X))^2) = E(\tfrac{1}{r^2}(X - E(X))^2) = E(h(X)) \\ &\geq E(g(x)) = p(|x - E(X)| \geq r) \ . \end{aligned}$$

□

Recall that the variance is defined as the expected value of the square of the difference between the random variable and the expected value. Such terms are used frequently, hence they get a term for quick reference:

Definition 1.34.
Let $X\colon \Omega \longrightarrow \mathbb{R}$ be a random variable. The ***n*-th moment** of X is $E(X^n)$. The ***n*-th central moment** is $E((X - E(X))^n)$.

In this terminology, the expected value is the first moment and the variance is the second central moment.

1.5 Stochastic independence and collections of random variables

Recall the definition of conditional probabilities (definition 1.7), which reads

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)} \ .$$

Now, following the terminology, one might wonder what happens if B is no condition for A at all. That is, assume that the conditional probability does not change the probability $p(A \mid B) = p(A)$. In this case, we are tempted to call A and B independent:

$$p(A \mid B) = p(A) \quad \Leftrightarrow \quad p(A) = p(A \mid B) = \frac{p(A \cap B)}{p(B)} \quad \Leftrightarrow \quad p(A \cap B) = p(A) \cdot p(B) \ .$$

Definition 1.35.
Two events $A, B \in \Sigma$ are called **stochastically independent** if

$$p(A \cap B) = p(A) \cdot p(B) \ .$$

This formula is familiar. Consider repeating an identical random process twice, which implies that there is no influence between the results. Then the probability that A and then B occurs is given by $p(A) \cdot p(B)$. However, stochastic independence does not mean that there is no causality between two events. Consider the following example:

Example 1.36.
Consider rolling a die twice and define the following events:

A: The sum of both results is odd.

B: The first roll yields an even number.

Here, the question if the sum is odd (A) depends on a causal level on the first roll (B). Yet, evaluating all possible results yields

$$p(A) = p(B) = \frac{1}{2} \quad \Rightarrow \quad p(A \cap B) = \frac{1}{4} = p(A) \cdot p(B) \ ,$$

so A and B are stochastically independent. This is no contradiction, as causality (here temporal causality of allowed outcomes) and the interdependence of probability are not the same concepts, though they may coincide in some cases. So stochastic independence may best be understood as follows. The occurrence of one event, does not change the likelihood of the second event, but may change the allowed outcomes such that the conditions for the second event are met.

Definition 1.37.
Let $\{A_i\}_{i \in I}$ be a family of events $A_i \in \Sigma$. The family is called **stochastically independent** if

$$p\left(\bigcap_{i \in I'} A_i\right) = \prod_{i \in I'} p(A_i) \qquad \forall \ I' \subseteq I \ .$$

At first, this definition seems redundant. However, stochastic independence between all events of a family does not imply stochastic independence of the whole family, and vice versa.

Example 1.38.
Consider the random generation of a binary tuple $\vec{a} = (a_1, a_2, a_3) \in \{0, 1\}^3$. Let s denote the number of ones, i.e. $s = a_1 + a_2 + a_3$, and define the following probabilities for these

cases:

$$p(\vec{a}) \coloneqq \begin{cases} 5/16 & s = 0 \\ 0 & s = 1 \\ 3/16 & s = 2 \\ 1/8 & s = 3 \end{cases}$$

Let A_j be the event, that the j-th component is one, i.e.

$$A_j \coloneqq \{\vec{a} \in \{0,1\}^3 \mid a_j = 1\} \ .$$

Direct calculation shows that

$$p(A_1) = p(A_2) = p(A_3) = \frac{1}{2} \ .$$

Observe that

$$p(A_1 \cap A_2 \cap A_3) = p(\{(1,1,1)\}) = \frac{1}{8} = p(A_1) \cdot p(A_2) \cdot p(A_3) \ .$$

However,

$$p(A_1 \cap A_2) = p(\{(1,1,0)\} \cup \{(1,1,1)\}) = p(\{(1,1,0)\}) + p(\{(1,1,1)\}) = \frac{3}{16} + \frac{1}{8} = \frac{5}{16}$$
$$\neq \frac{1}{4} = p(A_1) \cdot p(A_2) \ .$$

We have introduced the concept of stochastic independence on the fundamental level of probability spaces. However, as the previous sections have illustrated, we are particularly interested in random variables for calculations.

Definition 1.39 (collection of discrete random variables).
Let $X_1, \ldots, X_n$ be a collection of discrete random variables.
The **joint PMF** is defined as

$$p(x_1, \ldots, x_n) \coloneqq p(X_1 = x_1, \ldots, X_n = x_n) \ .$$

The **marginal PMF** for X_k is defined as

$$p_{X_k}(x_k) \coloneqq \sum_{x_1, \ldots, \widehat{x_k}, \ldots, x_n} p(x_1, \ldots, x_n) \ ,$$

where the hat denotes omission of that variable.

The definition of the joint PMF is straightforward. It is just the probability, that X_i yields x_i for all $i = 1, \ldots n$. The idea of the marginal PMF is, to obtain the probability for the single random variable left, by summing out the other random variables. For continuous random variables, the definition of the joint CDF follows the same reasoning. Also, the marginal PDF is defined by integrating out the remaining random variables from the joint PDF. However, the definition of the joint PDF is again not analogue to the PMF.

Definition 1.40 (**collection of continuous random variables**).
Let $X_1, \dots, X_n$ be a collection of continuous random variables.
The **joint CDF** is defined as

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) \coloneqq P(X_1 \leq x_1, \dots, X_n \leq x_n) \ .$$

The **joint PDF** (if it exists) is defined as

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) \coloneqq \frac{\partial^n}{\partial x_1 \dots \partial x_n} F_{X_1,\dots,X_n}(x_1, \dots, x_n) \ .$$

The **marginal CDF** for X_k is defined as

$$F_{X_k}(x_k) = F_{X_1,\dots,X_n}(\infty, \dots, x_k, \dots, \infty) \ .$$

The **marginal PDF** for X_k is defined as

$$f_{X_k}(x_k) \coloneqq \int_{\mathbb{R}^{n-1}} f_{X_1,\dots,X_n}(x_1, \dots, x_n) \ dx_1 \dots d\widehat{x_k} \dots dx_n \ ,$$

where the hat denotes omission of that variable.

The definition of the joint PDF is a direct generalization of the relation between PDF and CDF, which is $\frac{d}{dx}F(x) = f(x)$ (see corollary 1.15). Note, that again $\frac{d}{dx}F_{X_k}(x_k) = f_{X_k}(x_k)$. In fact, if the joint PDF exists (and is well defined, i.e. does not depend on the order of differentiation), it holds that

$$\begin{aligned} F_{X_1,\dots,X_n}(x_1, \dots, x_n) &= \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} \frac{\partial^n}{\partial y_1 \dots \partial y_n} F_{X_1,\dots,X_n}(y_1, \dots, y_n) \ dy_1 \dots dy_n \\ &= \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f_{X_1,\dots,X_n}(y_1, \dots, y_n) \ dy_1 \dots dy_n \ . \end{aligned}$$

Definition 1.41 (**stochastic independence of discrete random variables**).
Let $X_1, \dots, X_n$ be a collection of discrete random variables. They are called **stochastically independent** if

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i) \ .$$

Remark 1.42.
Unlike the stochastic independence of events, it is enough for the stochastic independence of random variables, to demand the factoring of the whole collection, since then any sub-collection factors, too.

For the definition of stochastic independence, the marginal PMFs are used. However, for stochastic independence, it is enough to find any probability functions that factor out the individual random variables, as the following lemma shows.

Lemma 1.43.
Let $X_1, \ldots, X_n$ be a collection of random variables and $p_i \colon X_i(\Omega_i) \longrightarrow [0,1]$ be functions that satisfy the probability axioms, i.e. $\sum_{x_i} p_i(x_i) = 1$. If

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i) \ ,$$

then $X_1, \ldots, X_n$ are stochastically independent and the p_i are the marginal probabilities $p_i \equiv p_{X_i}$.

Proof 1.44.
We just need to apply the definition of the marginal PMF to find

$$\begin{aligned} p_{X_i}(x_i) &= \sum_{x_1, \ldots, \widehat{x_k}, \ldots, x_n} p(x_1, \ldots, x_n) = \sum_{x_1, \ldots, \widehat{x_k}, \ldots, x_n} \prod_{i=1}^{n} p_i(x_i) \\ &= p_k(x_k) \prod_{i=1, \neq k}^{n} \sum_{x_i} p_1(x_i) = p_k(x_k) \ . \end{aligned}$$

But then, by the assumption of the lemma

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i) = \prod_{i=1}^{n} p_{X_i}(x_i) \ ,$$

which is the definition of stochastic independence. □

Definition 1.45.
Let $X_1, \ldots, X_n$ be a collection of continuous random variables. They are called **stochastically independent** if

$$F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i) \ .$$

Again, the condition that the individual factors have to be the marginal CDFs can be lifted.

Lemma 1.46.
Let $X_1, \ldots, X_n$ be a collection of continuous random variables and let $F_i \colon \mathbb{R} \longrightarrow \mathbb{R}$ be integrable functions such that $F_i(\infty) = 1$. If

$$F_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i) \ ,$$

then $F_i(x_i) = F_{X_i}(x_i)$ and thus $X_1, \ldots, X_n$ are stochastically independent.

Proof 1.47.

$$F_{X_k}(x_k) = F_{X_1,\dots,X_n}(\infty,\dots,x_k,\dots,X_k) = F_k(x_k)\prod_{i=1\neq k}^{n} F_i(\infty) = F_k(X_k)\ . \qquad \square$$

Furthermore, the condition for stochastic independence carries over to PDFs.

Lemma 1.48.
Let $X_1,\dots,X_n$ be a collection of continuous random variables with joint PDF $f_{X_1,\dots,X_n}(x_1,\dots,x_n)$. If and only if there are functions $f_i\colon \mathbb{R}\longrightarrow\mathbb{R}$ that are integrable with $\int_{\mathbb{R}} f_i(x)\ dx = 1$, such that

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} f_i(x_i)\ .$$

In this case $f_{X_k}(x_k) = f_k(x_k)$.

Proof 1.49.
First, we show that if the PDF factors through the random variables, the factors are the marginal PDFs:

$$\begin{aligned} f_{X_k}(x_k) &= \int_{\mathbb{R}^{n-1}} f_{X_1,\dots,X_n}(x_1,\dots,x_n)\ dx_1\dots d\widehat{x_k}\dots dx_n \\ &= \int_{\mathbb{R}^{n-1}} \prod_{i=1}^{n} f_i(x_i)\ dx_1\dots d\widehat{x_k}\dots dx_n \\ &= f_k(x_k)\prod_{i=1,\neq k}^{n}\int_{\mathbb{R}} f_i(x)\ dx = f_k(x_k) \end{aligned}$$

Now, it is enough to show that $f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} f_i(x_i)$ is equivalent to $X_1,\dots,X_n$ being stochastically independent. For this, recall that $f_{X_k}(x_k) = \frac{d}{dx_k}F_{X_k}(x_k)$ and thus vice versa $F_{X_k}(x_k) = \int_{-\infty}^{x_k} f_{X_k}(x)\,dx$. Assume now that $X_1,\dots,X_n$ are stochastically independent, then:

$$\begin{aligned} f_{X_1,\dots,X_n}(x_1,\dots,x_n) &= \frac{\partial^n}{\partial x_1\dots\partial x_n}F_{X_1,\dots,X_n}(x_1,\dots,x_n) \\ &= \frac{\partial^n}{\partial x_1\dots\partial x_n}\prod_{i=1}^{n} F_{X_i}(x_i) = \prod_{i=1}^{n}\partial_{x_i}F_{X_i}(x_i) \\ &= \prod_{i=1}^{n} f_{X_i}(x_i)\ . \end{aligned}$$

For the opposite direction, assume that $f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$, then

$$\begin{aligned} F_{X_1,\dots,X_n}(x_1,\dots,x_n) &= \int_{-\infty}^{x_1}\dots\int_{-\infty}^{x_n} f_{X_1,\dots,X_n}(y_1,\dots,y_n)\ dy_1\dots dy_n \\ &= \int_{-\infty}^{x_1}\dots\int_{-\infty}^{x_n}\prod_{i=1}^{n} f_{X_i}(y_i)\ dy_1\dots dy_n \end{aligned}$$

$$= \prod_{i=1}^{n} \int_{-\infty}^{x_i} f_{X_i}(y_i)\ dy_i = \prod_{i=1}^{n} F_{X_i}(x_i)\ ,$$

which is the definition for $X_1, \ldots, X_n$ to be stochastically independent. □

Lemma 1.50.
Let $X, Y \colon \Omega \longrightarrow \mathbb{R}$ be two stochastical independent random variables that are both discrete/continuous and have PMFs/PDFs. Then

$$E(XY) = E(X)E(Y)\ .$$

Proof 1.51.
For the discrete case, we calculate

$$\begin{aligned} E(XY) &= \sum_{x,y} p(x,y)xy = \sum_{x,y} (p_X(x)x)(p_Y(y)y) = \left(\sum_x p_X(x)x\right)\left(\sum_y p_Y(y)y\right) \\ &= E(X)E(Y)\ . \end{aligned}$$

For the continuous case, we use the existence of the PDF to obtain:

$$\begin{aligned} E(XY) &= \int_{\mathbb{R}^2} f_{X,Y}(x,y)xy\ dxdy = \int_{\mathbb{R}^2} f_X(x)xf_Y(y)y\ dxdy \\ &= \int_{\mathbb{R}} f_X(x)x\ dx \int_{\mathbb{R}} f_Y(y)y\ dy = E(X)E(Y)\ . \end{aligned}$$

Note that we used that both X and Y are discrete/continuous to avoid mixing integration and (possibly infinite) sums for this proof. □

1.6 Covariance and correlation

Another tool to describe the relation between random variables is the correlation. For this, we generalize the notion of variance:

Definition 1.52.
Let X, Y be real random variables. The **covariance** is defined as

$$\mathrm{Cov}(X,Y) \coloneqq E((X - E(X))(Y - E(Y)))\ .$$

Note that the covariance reduces to the variance for a single random variable:

$$\mathrm{Cov}(X,X) = E((X - E(X))(X - E(X))) = E((X - E(X))^2) = V(X)\ .$$

Lemma 1.53.
Let X, Y be random variables, both discrete/continuous, with PMFs/PDFs. If X and Y are stochastically independent, then $\mathrm{Cov}(X, Y) = 0$.

Proof 1.54.
The assumptions for X and Y are the same as in lemma 1.50, such that it applies here. Then:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E((X - E(X))(Y - E(X))) = E(X - E(X))E(Y - E(Y)) \\ &= (E(X) - E(X))(E(Y) - E(Y)) = 0\end{aligned}$$

□

Remark 1.55.
The opposite direction does not hold. Zero covariance does not imply stochastic independence. However, as is the case for all negations of implications, a non-zero covariance implies stochastic dependence.

The covariance is not normalized, such that the particular values do not tell on their own, how much the random variables depend upon each other.

Definition 1.56.
Let X, Y be random variables. The **Pearson correlation** is defined by

$$\rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} .$$

For the proof that the Pearson correlation is properly normalized, we need the following lemma:

Lemma 1.57.
It holds that $V(X + Y) = V(X) + V(Y) + 2\mathrm{Cov}(X, Y)$.

Proof 1.58.

$$\begin{aligned}V(X+Y) &= E((X + Y - E(X - Y))^2) = E(([X - E(X)] + [Y - E(Y)])^2) \\ &= E((X - E(X))^2) + E((Y - E(Y))^2) + 2E((X - E(X))(Y - E(Y))) \\ &= V(X) + V(Y) + 2\mathrm{Cov}(X,Y) .\end{aligned}$$

□

Proposition 1.59.
It holds that $\rho_{X,Y} \in [-1,1]$. Furthermore, the following extreme value properties hold:

(i) It holds that $\rho_{X,Y} = \pm 1$, if and only if $Y = \pm aX + c$ for any $a \geq 0$ and $c \in \mathbb{R}$.

(ii) If X and Y are stochastically independent, it follows that $\rho_{X,Y} = 0$.

Proof 1.60 (see [Cox97]).
First, using the properties of the variance (proposition 1.28), we have:

$$0 \leq V(tX+Y) = t^2 V(X) + 2t\mathrm{Cov}(X,Y) + V(Y) \ .$$

Note that this is a second order polynomial in t with $V(tX+Y) \geq 0$. This implies that the discriminant has to less than or equal to zero, as $V(tX+Y) = 0$ has at most one real solution:

$$0 \geq 4\mathrm{Cov}(X,Y)^2 - 4V(X)V(Y) \quad \Leftrightarrow \quad \mathrm{Cov}(X,Y)^2 \leq V(X)V(Y)$$

$$\Leftrightarrow \quad \frac{\mathrm{Cov}(X,Y)^2}{V(X)V(Y)} \in [0,1] \quad \Leftrightarrow \quad \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} \equiv \rho_{X,Y} \in [-1,1] \ .$$

***(i)*:**
For property (i), let $\rho_{X,Y} = \pm 1$, then $\mathrm{Cov}(X,Y)^2 = V(X)V(Y)$, and thus $0 = 4\mathrm{Cov}(X,Y)^2 - 4V(X)V(Y)$. But this means that the discriminant is zero, so there is exactly one $\alpha \in \mathbb{R}$ such that $V(\alpha X + Y) = 0$. By proposition 1.28 (iv) this means that $\alpha X + Y = c$ for a constant $c \in \mathbb{R}$. Put differently, $Y = -\alpha X + c$. It remains to check the sign of $a = -\alpha$.

$$\begin{aligned}
\mathrm{Cov}(X, aX+c) &= E((X - E(X))(aX + c - E(aX+c))) \\
&= E(aX^2 + cX - aXE(X) - cX - aXE(X) - cE(X) + aE(X)^2 + cE(X)) \\
&= aE(X^2) + cE(X) - aE(X)^2 - cE(X) - aE(X)^2 - cE(X) + aE(X)^2 + cE(X) \\
&= aE(X^2) - aE(X)^2 = a(E(X^2) - E(X^2)) = aV(X) \ .
\end{aligned}$$

Thus $\mathrm{Cov}(X, aX+c) > 0$ if $a > 0$ and $\mathrm{Cov}(X, aX+c) < 0$ if $a < 0$. From this, the statement (and the opposite) direction follow.
***(ii)*:**
This is a consequence of lemma 1.53. □

Remark 1.61 (Warnings).

- Stochastic independence implies that the random variables are uncorrelated. However, uncorrelated random variables need not be stochastically independent.
- Correlation is no causality

– **Inverse causality:**
When X and Y correlate, this does not tell, if Y depends on X or vice versa.

– **Common causality:**
There may be an effect, that causes both X and Y, hence the correlation, although neither X causes Y nor vice versa.

– **Coincidental correlation:**
The correlation can be purely accidental.

- The Pearson correlation is only linear. For example, perfect quadratic correlation $Y = X^2$ yields zero linear correlation.

1.7 The normal distribution

Definition 1.62.
A continuous random variable $X : \Omega \longrightarrow \mathbb{R}$ is **normally distributed** with **mean** μ and variance σ^2 if its PDF is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

In this case, one writes $X \sim \mathcal{N}(\mu, \sigma)$ for short.

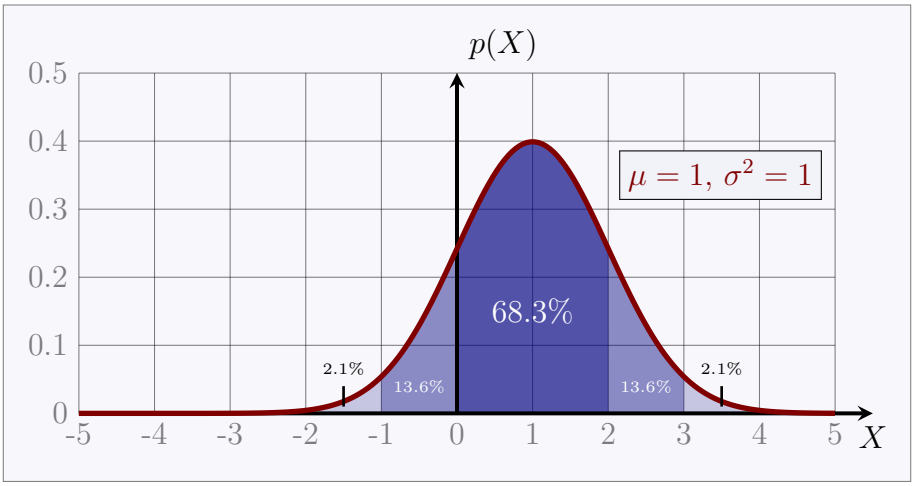

Figure 2: Example of a normal distribution with $\mu = 1$ and $\sigma^2 = 1$, together with the 3σ rule.

A normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2 has the following properties:

1. The expected value and variance are

$$E(X) = \mu \qquad \text{and} \qquad V(X) = \sigma^2 .$$

2. The higher central moments are given by

$$E((X-\mu)^n) = \begin{cases} 0 & n \text{ is odd} \\ \sigma^n \cdot (n-1)!! & n \text{ is even} \end{cases} .$$

3. The **3σ rule**, which states that

n	$p(\lvert x-\mu\rvert \leq n\sigma)$
1	$68,3\%$
2	$95,4\%$
3	99.7%

1.8 Law of large numbers and central limit theorem

Without further technicalities, the law of large numbers can be stated. In fact, the proof is not hard, and up to a single detail the following lemma covers, can be done with the tools we have established so far.

Lemma 1.63.
Let $X, Y : \Omega \longrightarrow \mathbb{R}$ be two random variables. Then

$$V(X+Y) = V(X) + V(Y) \ .$$

Proof 1.64.

$$\begin{aligned} V(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\ &= E(X^2 + X^2 + 2XY) - E(X)^2 - E(Y)^2 - 2E(XY) \\ &= E(X^2) + E(Y^2) + 2E(XY) - E(X)^2 - E(Y)^2 - 2E(XY) \\ &= E(X^2) - E(X^2) + E(Y^2) - E(Y)^2 \\ &= V(X) + V(Y) \ . \end{aligned}$$

□

Theorem 1.65 (weak law of large numbers (LLN)).
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of stochastically independent, identically distributed real random variables with $E(X_n) = \mu$ and $V(X_n) = \sigma^2$ for all $n \in \mathbb{N}$. Then it holds that

$$\lim_{n\to\infty} p\left(\left|\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right| \geq \varepsilon\right) = 0 \qquad \forall\ \varepsilon > 0 \ .$$

Proof 1.66.
Using corollary 1.21 we obtain:

$$E\left(\frac{1}{n}\sum_{i=1}^{n} X_n\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \ ,$$

With the lemma 1.63 and proposition 1.28 we obtain

$$V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n} \ .$$

Now, proposition 1.32 delivers the desired result:

$$p\left(\left|\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right| \geq \varepsilon\right) \leq \frac{V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\longrightarrow\infty} 0 \ . \qquad \square$$

As the name suggests, there is also a strong law of large numbers. Yet, the difference is only the convergence behavior.

Definition 1.67.
A sequence $(X_n)_{n\in\mathbb{N}}$ is said to **converge in distribution** to a random variable X, i.e. $X_n \xrightarrow{D} X$ if the CDFs satisfy

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(z) \qquad \forall\ x \in \mathbb{R} \text{ where } F_X \text{ is continuous .}$$

Convergence in distribution is essentially point wise convergence of the CDFs, yet excluding discontinuities of the limit-CDF. So in particular, whenever the CDFs converge point wise, the random variables converge in distribution.

Theorem 1.68 (Central limit theorem (Lindeberg-Lévi)).
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables with $E(X_n) = \mu$ and $V(X_n) = \sigma^2 < \infty$ for all $n \in \mathbb{N}$. Let $A_n = \frac{1}{n}\sum_{i=1}^{n} X_n$ denote the sequence of averages, then

$$\sqrt{n}(A_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma) \ .$$

The Lindeberg-Lévi central limit theorem (**CLT**) is considered as the classical CLT. There are different versions, where some of the restrictions can be lifted to weaker conditions, and there are generalization to higher dimensions. Yet, as with the weak law of large numbers, this version is enough to convey the message.

Remark 1.69.
Loosely speaking, the LLN states that the average of the results from repeated identical random experiments approaches the expected value of the distribution. The CLT states that $\sqrt{n}(A_n - \mu)$, where A_n is again the average of the results, approaches a normal distribution.
Strictly speaking, the results of the CLT and the LLN only apply to the limit $n \to \infty$. However, in practice the behavior that the average tends to stabilize at the expected value, and that the distribution becomes a normal distribution, can be observed for finite n.

1.9 Propagation of uncertainties

Taking into account, that there are always small fluctuations in systems, even in classical systems considered in the thermodynamic limit, the accuracy of a measurement is limited. Hence, there is reason to believe that the measured value x is accompanied with an uncertainty Δx, i.e. the real value lies in $[x - \Delta x, x + \Delta x]$. This is denoted as $x \pm \Delta x$.

However, then the question arises, how these uncertainties behave under a function. That is, let $x_1, \dots, x_n$ be measurements with uncertainties $\Delta x_1, \dots, \Delta x_n$ and $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ a differentiable function, what is the uncertainty Δf? In fact, there exist many manuals that contain a handy formula for Δf:

$$\boxed{\Delta f = \sqrt{\sum_{i=1}^{n} (\partial_i f(\vec{x}) \cdot \Delta x_i)^2} \ .} \tag{1}$$

In this subsection, we want to give a derivation of this formula, that is floating around in many manuals.

From a statistical point of view, we may regard the result of a measurement as random variable, and the uncertainty as standard deviation. From this point of view, the expected value can be regarded as the true result, and the random variable arises because of the limited precision of measurements. While the probability distribution is not important, knowledge of the expected value is required. In practice, one conducts the experiment many times. Justified by the LLN and CLT, one considers the average of the results as expected value, and obtains the standard derivation by calculating the variance of the results (usually assuming equal distribution).

So, let $X_1, \dots, X_n$ be random variables with expected values $\mu_1, \dots, \mu_n$. Let $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function. For short, we define the vectors

$$\vec{X} = (X_1, \dots, X_n) \quad \text{and} \quad \vec{\mu} = (\mu_1, \dots, \mu_n) \ .$$

Then the first order Taylor approximation around $\vec{\mu}$ reads:

$$f(\vec{X}) \approx f(\vec{\mu}) + \nabla f(\vec{\mu}) \cdot (\vec{X} - \vec{\mu}) = f(\vec{\mu}) + \sum_{i=1}^{n} \partial_i f(\vec{\mu})(X_i - \mu_i) \ .$$

Now, we can approximate the expected value for f:

$$\begin{aligned} E(f(\vec{X})) &\approx E\left(f(\vec{\mu}) + \sum_{i=1}^{n} \partial_i f(\vec{\mu})(X_i - \mu_i) \right) = f(\vec{\mu}) + \sum_{i=1}^{n} \partial_i f(\vec{\mu}) \underbrace{E(X_i - \mu_i)}_{=0} \\ &= f(\vec{\mu}) \ . \end{aligned}$$

Thus, the expected value of the function is approximately the function of the expected value. For the uncertainty Δf, we calculate the variance/standard deviation:

$$\begin{aligned} V(f(\vec{X})) &\approx V\left(f(\vec{\mu}) + \sum_{i=1}^{n} \partial_i f(\vec{\mu})(X_i - \mu_i) \right) \\ &= E\left(\left(f(\vec{\mu}) + \sum_{i=1}^{n} \partial_i f(\vec{\mu})(X_i - \mu_i) - E(f(\vec{X})) \right)^2 \right) \\ &\approx E\left(\left(f(\vec{\mu}) + \sum_{i=1}^{n} \partial_i f(\vec{\mu})(X_i - \mu_i) - f(\vec{\mu}) \right)^2 \right) \\ &= E\left(\left(\sum_{i=1}^{n} \partial_i f(\vec{\mu})(X_i - \mu_i) \right)^2 \right) \end{aligned}$$

$$
\begin{aligned}
&= E\left(\sum_{i,j=1}^{n}(\partial_i f(\vec{\mu})(X_i-\mu_i))(\partial_j f(\vec{\mu})(X_j-\mu_j))\right)\\
&= \sum_{i,j=1}^{n}\partial_i f(\vec{\mu})\cdot\partial_j f(\vec{\mu})\cdot E\left((X_i-\mu_i)(X_j-\mu_j)\right)\\
&= \sum_{i,j=1}^{n}\mathrm{Cov}(X_i,X_j)\cdot\partial_i f(\vec{\mu})\cdot\partial_j f(\vec{\mu})\ .
\end{aligned}
$$

The next assumption is, that the random variables X_i are all uncorrelated, i.e. the covariance $\mathrm{Cov}(X_i, X_j)$ vanishes for $i \neq j$:

$$
V(f(\vec{X})) \approx \sum_{i=1}^{n}(\partial_i f(\vec{\mu}))^2 V(X_i) = \sum_{i=1}^{n}(\partial_i f(\vec{\mu}))^2\sigma_{X_i}^2
$$

$$
\Rightarrow \quad \sigma_{f(\vec{X})} = \sqrt{V(f(\vec{X}))} \approx \sqrt{\sum_{i=1}^{n}(\partial_i f(\vec{\mu})\cdot\sigma_{X_i})^2}\ .
$$

Now, applying the interpretation, where $\mu_i \equiv x_i$ is the result of the measurement and $\sigma_{X_i} \equiv \Delta x_i$ is the uncertainty, (1) follows.

Remark 1.70 (Warnings).
Along the derivation of (1), we made some assumptions, that limit the applicability of the formula (which often is not mentioned):

- We used the first order Taylor approximation, which is only a good approximation, if $X_i - \mu_i$ is small. This means that the variance/square of uncertainty $V(X_i) \equiv (\Delta x_i)^2$ has to be small for the formula to be a good approximation of the propagated uncertainty. The more so, the more non-linear the function is.
- The random variables/measurements $X_i \equiv x_i$ have to be uncorrelated. Otherwise, the expression with non-zero covariance has to be used.

2 Information entropy

2.1 Axiomatic derivation of the information entropy

Consider a finite discrete random variable $X\colon \Omega \longrightarrow \mathbb{R}$, i.e. $X(\Omega) = \{x_1, \ldots, x_n\}$ with probabilities as $p_i = p_X(x_i)$, where p_X is the PMF of X. Here, we consider the choice of X to be also the choice of the PMF p_X. A natural question is, how well a result of this random variable (with its PMF) can be predicted. That is, we want to assign random variables/PMFs a numerical value (i.e. its information entropy) that indicates how hard a result can be predicted. Let Δ_n denote the set of all discrete random variables $X \equiv (X, p_X)$ with n possible outcomes. Then, we want to construct an information entropy

$$
H_n\colon \Delta_n \longrightarrow \mathbb{R}\ .
$$

In fact, this shall be a function that depends on the probabilities:

$$H_n(X) \equiv H_n(p_X) = H_n(p_1, \ldots, p_n) \ .$$

Yet, one can construct many function of random variables/PMFs that have nothing to do with their predictability. Thus, we make the following requirements/axioms (and explanations):

I : $H_n(p_1, \ldots, p_n)$ is continuous $p_1, \ldots, p_n$.
Small changes in the probability configuration changes the predictability only slightly, thus the information entropy should change only slightly.

II : $H_n(\frac{1}{n}, \ldots, \frac{1}{n})$ is a monotonously increasing sequence in n (note that the number of arguments also has to be increased).
If the system[1] in maximally undetermined, i.e. all results are equally probable, increasing the size n of the system makes every single result less probable, since $p_i = \frac{1}{n}$. Hence, the information entropy should increase, as the predictability decreases for increasing n.

III : The position of the respective probability does not matter, i.e.

$$H_n(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_n) = H_n(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_n) \ .$$

Changing the positions of the p_i is nothing but relabeling the results x_i. Yet, this should not change the predictability of the whole system.

IV : Let $q = \sum_{i=1}^{m} q_i$, then it shall hold that

$$H_{m+n}(q_1, \ldots, q_m, p_1, \ldots, p_n) = H_{1+n}\left(q, p_1, \ldots, p_n\right) + q \cdot H_m\left(\frac{q_1}{q}, \ldots, \frac{q_m}{q}\right) \ .$$

We may reduce a system artificially, by combining several results y_i with probabilities q_i to a single result y with probability q. Artificial means, that in the background the actual system is not changed, but that we count any result y_i as the result y, i.e. ignore which y_i it actually was. Then, the information entropy should account for this. Here, we demand that the internal entropy of the y_i is added with the probability q that any of the y_i has occurred as weight. For the internal probability, we take the probabilities q_i, normalized to the probability q that any results y_i occurred.

These axioms seem reasonable and innocent enough. One might think, that they allow for many information entropies. However, they determine a unique information entropy, at least up to a scaling factor.

[1] System refers to the system the random variable describes. We choose this terminology, as it is more intuitive.

Theorem 2.1.
If H_n satisfies the above axioms, it holds that (for an arbitrary $c \in \mathbb{R}^+$)

$$H_n(p_1, \ldots, p_n) = -c \sum_{i=1}^{n} p_i \ln(p_i) \ . \tag{2}$$

Proof 2.2 (using ideas from [Ale05, proof of Theorem 2.1])**.**
First, we approximate the probabilities p_i by fractions $p_i = \frac{m_i}{m}$, where $m = \sum_{i=1}^{n} m_i$. Because of the continuity, we may assume that

$$H_n(\tfrac{m_1}{m}, \ldots, \tfrac{m_n}{m}) \underset{m\to\infty}{\simeq} H_n(p_1, \ldots, p_n) \ .$$

This approximation can be understood as follows. We consider m uniformly distributed events $y_{i,j}$ with probability $\frac{1}{m}$. Then we bunch together the events $y_{i,1}, \ldots, y_{i,m_i} = x_i$ which thus have the probability $\sum_{j=1}^{m_i} \frac{1}{m} = \frac{m_i}{m} \approx p_i$. Now, using (2), axiom **IV** and then $\frac{m_i}{m} \approx p_i$ yields:

$$\begin{aligned} H_n(p_1, \ldots, p_n) \approx H_n(\tfrac{m_1}{m}, \ldots, \tfrac{m_n}{m}) &= H_m(\tfrac{1}{m}, \ldots, \tfrac{1}{m}) + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{1}{m} H_{m_i}(\tfrac{1}{m_i}, \ldots, \tfrac{1}{m_i}) \\ &= H_m(\tfrac{1}{m}, \ldots, \tfrac{1}{m}) + \sum_{i=1}^{n} \frac{m_i}{m} H_{m_i}(\tfrac{1}{m_i}, \ldots, \tfrac{1}{m_i}) \\ &\approx H_m(\tfrac{1}{m}, \ldots, \tfrac{1}{m}) + \sum_{i=1}^{n} p_i H_{m_i}(\tfrac{1}{m_i}, \ldots, \tfrac{1}{m_i}) \ . \end{aligned}$$

Note that the terms of the form $H_m(\frac{1}{m}, \ldots, \frac{1}{m})$ contain m arguments $\frac{1}{m}$. Since the arguments are entirely determined by the integer m in this case, we write H_m for short:

$$\Rightarrow \quad H_n(\tfrac{m_1}{m}, \ldots, \tfrac{m_n}{m}) = H_m + \sum_{i=1}^{n} \frac{m_i}{m} H_{m_i} \approx H_m + \sum_{i=1}^{n} p_i H_{m_i} \ . \tag{3}$$

From the axioms of the information entropy, we can already deduce the form of such terms. Let for a moment $m_i = \alpha$, then $m = \sum_{i=1}^{n} m_i = \alpha \cdot n$. Hence $\frac{m_i}{m} = \frac{\alpha}{\alpha \cdot n} = \frac{1}{n}$. The above equation then becomes

$$H_n + \sum_{i=1}^{n} \frac{1}{n} H_{m_i} = H_n + H_\alpha = H_m = H_{\alpha \cdot n} \ .$$

As function of the integer, this reads $H_n + H_\alpha = H_{\alpha \cdot n}$. To find a functional expression for this, we want to solve the following functional equation:

$$f(x \cdot y) = f(x) + f(y) \ .$$

Considering $g(x) = f(e^x)$ this becomes:

$$g(x) + g(y) = f(e^x) + f(e^y) = f(e^x \cdot e^y) = f(e^{x+y}) = g(x + y) \ .$$

This is **Cauchy's functional equation**, for which it can be shown that it only has the solution $g(x) = c \cdot x$ with $c \in \mathbb{R}$, if one requires f to be continuous. But then

$$f(e^x) = c \cdot x \qquad \Rightarrow \qquad f(x) = c \cdot \ln(x) \ .$$

Hence, terms of the form H_m are $H_m = c \cdot \ln(m)$. Since we require H_m to be monotonously increasing for increasing m, c has to be positive, i.e. $c \in \mathbb{R}^+$. Returning to (3), we find:

$$\begin{aligned} H_n(p_1, \ldots, p_n) \approx H_n(\tfrac{m_1}{m}, \ldots, \tfrac{m_n}{m}) \approx H_m + \sum_{i=1}^{n} p_i H_{m_i} &= c\ln(m) + \sum_{i=1}^{n} p_i \cdot c \cdot \ln(m_i) \\ - c\left(\sum_{i=1}^{n} p_i \ln(m_i) - \ln(m)\right) &= -c\left(\sum_{i=1}^{n} p_i \ln(m_i) - \ln(m)\sum_{i=1}^{m} p_i\right) \\ = -c\left(\sum_{i=1}^{n} p_i \ln(m_i) - \sum_{i=1}^{m} p_i \ln(m)\right) &= -c\sum_{i=1}^{n} p_i(\ln(m_i) - \ln(m)) \\ = -c\sum_{i=1}^{n} \ln(\tfrac{m_i}{m}) \approx -c\sum_{i=1}^{n} p_i \ln(p_i) \ . & \end{aligned}$$

□

Based on this theorem, one makes the following definition:

Definition 2.3.
Let $X : \Omega \longrightarrow \mathbb{R}$ be a discrete random variable with PMF p. The **information entropy**n is defined as

$$H(X) \coloneqq -c \sum_{x \in X(\Omega)} p(x) \ln(p(x)) \ ,$$

for any $c \in \mathbb{R}^+$.

2.2 Elementary properties of the information entropy

First, we may address the unspecified constant c. It is nothing but a scaling factor for the information entropy, and can be regarded as choice of units for the information entropy. For example, in thermodynamics one chooses the Boltzmann constant $c = k_B$, that relates energy to temperature (a statistical property), since $[k_B] = \frac{\text{Energy}}{\text{Temperature}}$. Note that logarithms with different bases are related to each other as follows:

$$\log_a(\bullet) = \log_a(b) \log_b(\bullet) \ .$$

Thus, choosing c appropriately, one can change the basis of the logarithm. A common choice in information theory is

$$H(X) = -\sum_{x \in X(\Omega)} p(x) \log_2(p(x)) \ ,$$

which is an entropy in the units of bits. Other direct observations are:

- $H(X) = E(-\ln(p(X)))$, by definition of the expected value.
- $H(X) \geq 0$, since $p(x) \leq 1$ for all $x \in X(\Omega)$

In the explanation of axiom **II**, we called a system maximally undetermined if the results are uniformly distributed for a finite discrete random variable. In fact, this is a result that follows from the information entropy:

Theorem 2.4.
The information entropy $H(p_1, \ldots, p_n)$ has its global maximum at $(p_1, \ldots, p_n) = (\frac{1}{n}, \ldots, \frac{1}{n})$.

Proof 2.5 (Here we use the argument from [Exc20][2])**.**
The probabilities are restricted to the set $P := \{\vec{p} \in \mathbb{R}^n \mid \sum_{i=1}^n p_i = 1\}$, which is compact. Thus, the function $H(p_1, \ldots, p_n)$ has a global maximum on P. Now, assume without loss of generality that $p_i < p_j$ for $i \neq j$. Then there is an $\varepsilon > 0$ such that $p_i + \varepsilon < p_j - \varepsilon$. We calculate (to keep the notation short, we choose $c = 1$. This does not change the argument, since $c > 0$ anyway):

$$\begin{aligned}
&H(p_1, \ldots, p_i + \varepsilon, \ldots, p_j - \varepsilon, \ldots, p_n) - H(p_1, \ldots, p_n) \\
&\quad = -(p_i + \varepsilon)\ln(p_i + \varepsilon) - (p_j - \varepsilon)\ln(p_j - \varepsilon) + p_i \ln(p_i) + p_j \ln(p_j) \\
&\quad = -p_i \ln\left(\tfrac{p_i+\varepsilon}{p_i}\right) - \varepsilon \ln(p_i + \varepsilon) - p_j \ln\left(\tfrac{p_j-\varepsilon}{p_j}\right) + \varepsilon \ln(p_j - \varepsilon) \\
&\quad = -p_i \ln\left(1 + \tfrac{\varepsilon}{p_i}\right) - p_j \ln\left(1 - \tfrac{\varepsilon}{p_j}\right) - \varepsilon\left(\ln\left(p_i\left(1 + \tfrac{\varepsilon}{p_i}\right)\right) - \ln\left(p_j\left(1 - \tfrac{\varepsilon}{p_j}\right)\right)\right) \\
&\quad = -p_i \ln\left(1 + \tfrac{\varepsilon}{p_i}\right) - p_j \ln\left(1 - \tfrac{\varepsilon}{p_j}\right) - \varepsilon\left(\ln(p_i) + \ln\left(1 + \tfrac{\varepsilon}{p_i}\right) - \ln(p_j) - \ln\left(1 - \tfrac{\varepsilon}{p_j}\right)\right)
\end{aligned}$$

Next, we recall that $\ln(1 + x) = \sum_{k=0}^n \frac{(-1)^{k-1}}{k} x^k = x + \mathcal{O}(x^2)$. Thus,

$$\begin{aligned}
&H(p_1, \ldots, p_i + \varepsilon, \ldots, p_j - \varepsilon, \ldots, p_n) - H(p_1, \ldots, p_n) \\
&\quad = -p_i \tfrac{\varepsilon}{p_i} + p_j \tfrac{\varepsilon}{p_j} - \varepsilon\left(\ln(p_i) + \tfrac{\varepsilon}{p_i} - \ln(p_j) + \tfrac{\varepsilon}{p_j}\right) + \mathcal{O}(\varepsilon^2) \\
&\quad = -\varepsilon - \varepsilon \ln(p_i) + \varepsilon + \varepsilon \ln(p_j) + \mathcal{O}(\varepsilon^2) = \varepsilon \ln\left(\tfrac{p_j}{p_i}\right) + \mathcal{O}(\varepsilon^2)\ .
\end{aligned}$$

Since $p_i < p_j$, this difference is positive for sufficiently small ε. However, then the entropy is larger for $p_i + \varepsilon$ and $p_j - \varepsilon$, as long as $p_i < p_j$. Fixing all but p_i and p_j, we obtain the maximum if $p_i = p_j$. Since we can repeat this argument for all p_i, it follows that the entropy is maximal if $p_1 = \ldots = p_n$, which implies that $p_i = \frac{1}{n}$ for all i. □

[2] There are also proofs, using Lagrange multipliers to find a possible candidate for a local maximum. However, to find a sufficient condition becomes rather convoluted.

Lemma 2.6 (Boltzmann entropy)**.**
In case of a uniform distributed random variable X, i.e. $p_i = \frac{1}{n(X)}$, where $n(X)$ is the number of results, the information entropy reduces to

$$H_n(X) = c \ln(n(X)) \ .$$

Proof 2.7.

$$H(X) = -c \sum_{i=1}^{n(X)} \frac{1}{n(X)} \ln\left(\tfrac{1}{n(X)}\right) = -c \ln\left(\tfrac{1}{n(X)}\right) = c \ln(n(X)) \ . \qquad \square$$

The name Boltzmann entropy may be a stretch of terminology. Yet, choosing $c = k_B$ and denoting $n(X) = \Omega$, we obtain the Boltzmann entropy $S = k_B \ln(\Omega)$.

2.3 Joint and conditional entropies

In the section on probabilities, we have met joint and conditional probabilities for random variables. These concepts can be applied to the information entropy.

Definition 2.8.
Let $X \colon \Omega_X \longrightarrow \mathbb{R}$ and $Y \colon \Omega_Y \longrightarrow \mathbb{R}$ be discrete random variables, and let $p(x, y)$ be the joint PMF. The **joint information entropy** is defined as

$$H(X, Y) \coloneqq -c \sum_{\substack{x \in X(\Omega_X) \\ y \in Y(\Omega_Y)}} p(x, y) \ln(p(x, y)) \ .$$

As **conditional information entropy** with single condition, we define:

$$H(Y \mid X = x) \coloneqq -c \sum_{y \in Y(\Omega_Y)} p(y \mid x) \ln(p(y \mid x)) \ .$$

As **conditional information entropy** $H(Y \mid X)$, where X can take any of its values, we define:

$$H(Y \mid X) \coloneqq \sum_{x \in X(\Omega_X)} p(x) H(Y \mid X = x) \ .$$

Note that the conditional information entropy $H(Y \mid X)$ is the superposition of the conditional information entropy $H(Y \mid X)$, weighted by the probability $p(x)$ that this condition $X = x$ is met.

Corollary 2.9.
It holds that

$$H(Y \mid X) = -c \sum_{\substack{x \in X(\Omega_X) \\ y \in Y(\Omega_Y)}} p(x, y) \ln(p(y \mid x)) \ .$$

Proof 2.10.
This follows from the definition of conditional probabilities (definition 1.7)

$$\begin{aligned} H(Y \mid X) &= \sum_{x \in X(\Omega_X)} p(x) H(Y \mid X = x) \\ &= -c \sum_{x \in X(\Omega_X)} p(x) \sum_{y \in Y(\Omega_Y)} p(y \mid x) \ln(p(y \mid x)) \\ &= -c \sum_{\substack{x \in X(\Omega_X) \\ y \in Y(\Omega_Y)}} p(x) p(y \mid x) \ln(p(y \mid x)) \\ &= -c \sum_{\substack{x \in X(\Omega_X) \\ y \in Y(\Omega_Y)}} p(x, y) \ln(p(y \mid x)) \ . \end{aligned}$$

□

Lemma 2.11.
It holds that

$$H(X, Y) = H(X) + H(Y \mid X) \ .$$

Proof 2.12.
This is a straightforward calculation, using corollary 2.9, the definition of conditional probabilities and the definition of marginal PMFs:

$$\begin{aligned} H(X, Y) &= -c \sum_{x,y} p(x, y) \ln(p(x, y)) = -c \sum_{x,y} p(x, y) \ln(p(y \mid x) p(x)) \\ &= -c \sum_{x,y} p(x, y) \left(\ln(p(y \mid x)) + \ln(p(x)) \right) \\ &= -c \sum_{x,y} p(x, y) \ln(p(x)) - c \sum_{x,y} p(x, y) \ln(p(y \mid x)) \\ &= -c \sum_{x} \left(\sum_{y} p(x, y) \right) \ln(p(x)) + H(Y \mid X) \\ &= -c \sum_{x} p(x) \ln(x) + H(Y \mid X) = H(x) + H(Y \mid X) \ . \end{aligned}$$

□

References

[Ale05] Semyon Alesker. "Theory of valuations on manifolds, IV. New properties of the multiplicative structure". In: (2005). eprint: arXiv:math/0511171.

[Cox97] Dennis Cox. *Covariance and Correlation*. 1997. URL: http://www.stat.rice.edu/~dcox/Stat421/Supp2/node3.html.

[CT12] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012. ISBN: 9781118585771.

[Exc20] Stack Exchange. *Why is Entropy maximised when the probability distribution is uniform?* 2020. URL: https://stats.stackexchange.com/questions/66108/why-is-entropy-maximised-when-the-probability-distribution-is-uniform.

[Geo07] H.O. Georgii. *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik*. De-Gruyter-Lehrbuch. de Gruyter, 2007. ISBN: 9783110193497.