IV111 2020

March 14, 2021

1.

Definition 1.1 (Sample space)**.** The (non-empty) set of the possible outcomes of a random experiment (náhodný pokus) is called the sample space (základný prostor) of the experiment and it will be denoted S. The outcomes of a random experiment (elements of the sample space) are denoted sample points(základné, elementárne body)

Sample space is **determined by purpose**. e.g. $0, 1, 2$, $S_1 = \{0, 1, 2\}$, $S_2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$

Definition 1.2 (Event)**.** Event (jav) of a random experiment with sample space **S** is any **subset of S**.

Definition 1.3 (Mutually exclusive)**.** Events $A_1, A_2, \ldots, A_n$ are **mutally exclusive** (nezlúčiteľné) or mutually disjoint if and only if:

$$\forall i,\ i \neq j . A_i \cap A_j = \emptyset$$

Event (jav) of a random experiment with sample space **S** is any **subset of S**.

Definition 1.4 (Collectively exhaustive)**.** Events $A_1, A_2, \ldots, A_n$ are **collectively exhaustive** (tvoria úplný systém) if and only if:

$$A_1 \cup A_2 \cup \cdots \cup A_n = \mathbf{S}$$

Event (jav) of a random experiment with sample space **S** is any **subset of S**.

Definition 1.5 (Partition)**.** Mutually exclusive and collectively exhaustive list is called a **partition** (rozklad) of the sample space **S**.

Definition 1.6. A **probability** function P on a (discrete) sample space **S** with a set of all events $\mathscr{F} = 2^{\mathbf{S}}$ is a function $P : \mathscr{F} \to [0, 1]$ such that:

- $P(\mathbf{S}) = 1$
- $P(A) \geq 0$
- $P(A \cup B) = P(A) + P(B)$, A and B are mutually exclusive
- for any countable sequence of mutually exclusive events $A_1, A_2, \ldots, A_n$:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Definition 1.7. A discrete **probability space** or **system** is a triple (**S**, $\mathscr{F}$, **P**)

- **S** is **sample space**, the set of all possible outcomes,
- $\mathscr{F} = 2^{\mathbf{S}}$, collection of all events,
- **P** : $\mathscr{F} \to [0, 1]$, probability function

Definition 1.8. A **probability space** or **system** is a triple (**S**, $\mathscr{F}$, **P**)

- **S** is **sample space**, the set of all possible outcomes,
- $\mathscr{F} \subseteq 2^{\mathbf{S}}$ is a **σ-field** if
  - $\emptyset \in \mathscr{F}$
  - if $A_1, A_2, \cdots \in \mathscr{F}$, then $P \bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$
  - $A \in \mathscr{F}$, then $\overline{A} \in \mathscr{F}$
  - Why? Because 2^S (in continuous world) is a too large collection for probabilities to be assigned reasonably to all its members.
- **P** : $\mathscr{F} \to [0, 1]$, probability function

Definition 1.9 (Conditional probability)**.** For $P(B) \neq 0$, the **conditional probability** of A given B:

$$\begin{aligned} P(A \mid B) &= \sum_{s \in A} P(s \mid B) \\ &= \sum_{s \in A \cap \overline{B}} P(s \mid B) + \sum_{s \in A \cap B} P(s \mid B) \\ &= \sum_{s \in A \cap B} P(s \mid B) \\ &= \sum_{s \in A \cap B} \frac{P(s)}{P(B)} \\ &= \frac{P(A \cap B)}{P(B)}, P(B) \neq 0 \end{aligned}$$

$$P(s \mid B) = \begin{cases} \frac{P(s)}{P(B)} & if s \in B, \\ 0 & if s \in \overline{B}, \end{cases}$$

A rearrangement of this definition yields the following **multiplication rule**:

$$P(A \cap B) = \begin{cases} P(B) \cdot P(A \mid B) & if P(B) \neq 0, \\ P(A) \cdot P(B \mid A) & if P(A) \neq 0, \\ 0 & otherwise \end{cases}$$

Definition 1.10 (Independence of events)**.** Events A and B are said to be independent, if:

$$P(A \cap B) = P(A) \cdot P(B) (= P(A \mid B) \cdot P(B))$$

$P(A \mid B) = P(A)$ is saying that the probability of the event A **does not change** regardless of whether event B occurred.

- Indipendence is not transitive,
- If A and B are independent, then also A and $\overline{B}$, $\overline{A}$ and B, and $\overline{A}$ and $\overline{B}$ are independent

Definition 1.11 ((mutually) independent)**.** Events $A_1, A_2, \dots A_n$ are (mutually) **independent** if and only if for any set $\{i_1, i_2, \dots i_k\} \subseteq \{1, \dots n\} (2 \leq k \leq n)$ of distinct indices it holds that:

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \dots P(A_{i_k})$$

Definition 1.12 ((pairwise) independent)**.** Events $A_1, A_2, \dots A_n$ are (pairwise) **independent** iff for all distinct indices $i, j \in \{1, \dots n\}$ it holds that:

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j)$$

Definition 1.13 (Law of Total Probability)**.** Let A be an event and $\{B_1, \dots, B_n\}$ be an event space. Then

$$\begin{aligned}
\sum_{i=1}^{n} P(A) \cdot P(A \mid B_i) &= \sum_{i=1}^{n} P(A \cap B_i) \\
&= P(\bigcup_{i=1}^{n} (A \cap B_i)) \\
&= P(A \cap \bigcup_{i=1}^{n} B_i) \\
&= P(A \cap \mathbf{S}) \\
&= P(A)
\end{aligned}$$

Definition 1.14 (Theorem (Bayes' Rule))**.** Let A be an event and $\{B_1, \dots, B_n\}$ be an event space. Then of every B_j

$$P(B_j \mid A) = \frac{P(A|B_j) \cdot P(B_j)}{P(A)} = \frac{P(A|B_j) \cdot P(B_j)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$$

2.

Definition 2.1. random variable A (discrete) **random variable** X on a sample space **S** is a function $X : S \to \mathbb{R}$ that assigns a real number $X(s)$ to each sample point $s \in S$.
We define **image** of a random variable X as the set $\mathbf{Im}(X) = \{X(s) \mid s \in \mathbf{S}\}$ - countable set; we can imagine this like die with colors instead of numbers, and random variable is assigning numbers to these colors.

Definition 2.2. inverse image A random variable partitions its sample space **S** into a *mutually exclusive* and *collectively exhaustive* set of events. Thus for a random variable **X** and a real number x, we define the event A_x - called **inverse image**:

$$A_x = [X = x] = \{s \in \mathbf{S} \mid \mathbf{X}(s) = x\}$$

It's clear that $A_x \cap A_y = \emptyset$. if $x \neq y$ and also:

$$\bigcup_{x \in \mathbb{N}} A_x = \mathbf{S}$$

Thus probability that the value of the random variable X obtained on a performance of the experiment is equal to x:

$$\begin{aligned} P(A_x) &= P([\mathbf{X} = x]) \\ &= P(\{s \mid \mathbf{X}(s) = x\}) \\ &= \sum_{X(s)=x} P(s) \end{aligned}$$

We would like to compute the probability for some subset $A \subseteq \mathbb{R}$ of the event $\{s \mid X(s) \in A\}$. We write:

$$P(\{s \mid X(s) \in A\}) = P([X \in A]) = P(X \in A) = P(\bigcup_{x_i \in A} \{s \mid X(s) = x_i\}) = \sum_{x_i \in A} p_X(x_i)$$

Definition 2.3. cumulative distribution function/distribučná funkcia of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ given by:

$$F_{X(x)} = P(X \leq x) = \sum_{t \leq x} p_X(t)$$

It follows that:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

Properties:

- $0 \leq F(x) \leq 1$ for $-\infty < x < \infty$. This follows because $F(x)$ is a probability.
- $F(x)$ is a **monotone increasing** function of x;that is,if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.
- $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to \infty} F(x) = 1$.
- $F(x)$ is a right-continuous function; $F(x) = \lim_{i \to x^+} F(i)$

Definition 2.4. The **Bernoulli** pmf (cinknutá minca)
The **Bernoulli** pmf of a discrete random variable X is the density function of a discrete random variable X having 0 and 1 as its ONLY possible values. $(p+q=1)$

$$p_X(0) = p_0 = P(X=0) = 1-p = q$$
$$p_X(1) = p_1 = P(X=1) = 1 = p$$

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ q & \text{for } 0 \le x \le 1, \\ 1 & \text{for } x \ge 1. \end{cases}$$

Definition 2.5. The **binomial** pmf
Binomial random variable X, denoted by $B((k;)n,p)$, has two $(+\ k$ successes) parameters, count n and probability p. It models the number of successes (outcomes 1) in n consecutive Bernoulli trials with probability p of success in each trial.

$$\begin{aligned} p_k &= P(X=k) \\ &= p_X(k) \\ &= \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } 0 \le k \le n, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

$$B(k;n,p) = F_X(k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

Definition 2.6. the **Geometric** pmf
Has one paramter p and models the number of Bernoulli trials until the FIRST 'success' occurs $(q = 1-p)$. **Geometric** pmf is the only discrete distribution with **memory-less** property:
$P(X > t) = P(X > t+i \mid X > i)$, be carefull e.g. $P(X > 40 \mid X > 30) = P(X > 10)$ but $P(X > 40 \mid X > 30) \neq P(X > 40)$

$$p_X(i) = P(X=i) = pq^{i-1} = p(1-p)^{i-1}$$

$$\begin{aligned} F_X(t) &= P(X \le t) \\ &= \sum_{i=1}^{\lfloor t \rfloor} p(1-p)^{i-1} \\ &= 1 - (1-p)^t \end{aligned}$$

Definition 2.7. The discrete **uniform** pmf (Hracia kocka)
Let X be a discrete random variable with a finite image $\{x_1, x_2, \ldots, x_n\}$ and let us assign to all elements of the image the same probability $p_X(x_i) = p$.

$$p_X = \begin{cases} \frac{1}{n} & x_i \in \mathbf{Im}(X) \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x < x_0 \\ \frac{j}{n} & x_{j-1} \le x \le x_j, j \in \{1, \ldots, n\} \\ 1 & x \ge x_n \end{cases}$$

Definition 2.8. Constant random variable
Fo a real number c, the function X defined by $X(s) = c$ for all $s \in \mathbf{S}$ is a discrete random variable.

$$p_X(x) = \begin{cases} 1 & \text{if } x = c \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & \text{for } x < c, \\ 1 & \text{for } x \geq c \end{cases}$$

Definition 2.9. indicator random variable
Assume that event A partitions the sample space **S** into two mutally exclusive and collectively exhaustive subsets, A and $\overline{A}$. The **indicator** of event A is a random variable I_A defined by

$$I_A(s) = \begin{cases} 0 & \text{if } s \in A, \\ 1 & \text{if } s \in \overline{A} \end{cases}$$

$$p_{I_A}(0) = P(\overline{A}) = 1 - P(A)$$
$$p_{I_A}(1) = P(A)$$

$$F_{I_A}(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 - P(A) & \text{for } 0 \leq x < 1, \\ 1 & \text{for } x \geq 1 \end{cases}$$

Definition 2.10. Discrete random **vector**
Often we may be interested in studying relationships between TWO OR MORE random variables defined on a given sample space.
Let $X_1, X_2, \ldots, X_r$ be discrete random variables defined on a sample space S.
The random vector $X = (X_1, X_2, \ldots, X_r)$ is an r-dimensional vector-valued function $X : S \to \mathbb{R}^r$ with $X(s) = \mathbf{x} = (X_1(s) = x_1, X_2(s) = x_2, \ldots, X_r(s) = x_r)$, where
joint (or compound) **probability distribution**

$$p_X(\mathbf{x}) = P(X = \mathbf{x}) = (X_1(s) = x_1 \wedge X_2(s) = x_2 \wedge \cdots \wedge X_r(s) = x_r)$$

joint (or compound) **distribution function**

$$F_X(\mathbf{x}) = (X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \cdots \wedge X_r \leq x_r)$$

Definition 2.11. marginal pmf
In other words, to obtain marginal pmf $p_{X(x)}$. we erect a vertical column at $X = x$ and sum the probabilities of all event points touched by the column.

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= P(\bigcup_j \{X = x, Y = y_j\}) \\ &= P(\sum_j \{X = x, Y = y_j\}) \\ &= \sum_j p_{X,Y}(x, y_j) \end{aligned}$$

Definition 2.12. multinomial pmf
which is a generalization of the binomial pmf. Consider n balls falling into r bins with probability $p_1, p_2, \ldots, p_r$. We define the random vector $X = (X_1, X_2, \ldots, X_r)$ such that X_i is the number of trials that resulted in i-th outcome, i.e. the number of balls in the $i - th$ bin. (search google multinomial distribution, e.g.) Sum over i, n_i must be equal to n
Compound probability:

$$\begin{aligned} p_X(n) &= P(X_1 = n_1, X_2 = n_2, \ldots, X_r = n_r) \\ &= \binom{n}{n_1, n_2, \ldots, n_r} p_1^{n_1}, p_2^{n_2}, \ldots, p_r^{n_r} \end{aligned}$$

Marginal probability:

$$p_{X_i}(n_i) = \binom{n}{n_i} p^{n_i}(1-p)^{n-n_i}$$

Definition 2.13. Independent random variables
Two discrete random variables X and Y are defined to be independent:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \qquad \forall x, y$$

Also if X and Y are two **independent** random variables, then for any two subsets $A, B \subseteq \mathbb{R}$ the events $X \in A$ and $Y \in B$ are independent:

$$P(X \in A \wedge Y \in B) = P(X \in A)P(Y \in B)$$

Definition 2.14. mutually indipendent
Let $X_1, ..., X_r$ be discrete random variables with probability distributions $p_{X_1}, \ldots, p_{X_r}$. These random variables are mutually independent iff

$$p_{X_1,X_2,\ldots,X_r}(X_1, X_2, \ldots, X_r) = p_{X_1}(x_1)p_{X_2}(x_2) \ \ldots \ p_{X_r}(x_r)$$

Definition 2.15. discrete **convolution**
Let X and Y be (non-negative) independent random variables. Random variable $Z = X + Y$ is called convolution of X and Y and its probability distribution is

$$\begin{aligned} p_Z(t) &= p_{X+Y}(t) \\ &= \sum_{x=0}^{t} p_X(x)p_Y(t - x) \end{aligned}$$

Definition 2.16. Random variable
A random (continuous) variable X on a **probability space** (**S**, $\mathcal{F}$,**P**) is a function $X : S \to \mathbb{R}$ that assigns a real number $X(s)$ to each sample point $s \in S$, such that:

$$\forall r \in \mathbb{R} \quad \{s \mid X(s) \leq r\} \in \mathcal{F}$$

Definition 2.17. cumulative distribution function
of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(x) = P(X \leq x)$$

it's good to realize that $\forall r \in \mathbb{R} \quad P(X = x) = 0$.

Definition 2.18. probability **density** function
For a continuous random variable X, is a derivative function of cumulative distribution function; thus

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

Let's think about real (one-dimensional) things. If you think of the total amount of probability as a liquid poured over the real number line, the areas where there is more probability will have thicker levels of liquid. You can describe the position of the surface of the liquid by a function $f(x)$ - a density function, if you will, since it tells you how thick the liquid is at any location.

$(f1)$ $\forall x \quad f(x) \geq 0$

$(f2)$ $\int_{-\infty}^{\infty} f(x)dx = 1$

3.

Definition 3.1. Expectation (stredá hodnota)
The expected value of the random variable X denotates the "center" of a probability mass (or density) function in tje sense of a weighted averge, or better, in the senes of a center of gravity.

$$E[X] = \begin{cases} \sum_i x_i p(x_i) & \text{if X is discrete,} \\ \int_{-\infty}^{\infty} x f(x) dx, & \text{if X is continuous.} \end{cases}$$

Provided the relevant sum or integral is absolutely convergent.

Definition 3.2. Some examples

- Uniform r. dis. $E[X] = \sum_i x_i p(x_i) = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Bernoulli r. dis. $E[X] = \sum_i x_i p(x_i) = 0 \cdot (1-p) + 1 \cdot p = p$
- Indicator r. var. $E[X] = E(I_A) = P(A)$
- Constant r. var. $E[X] = 1 \cdot c$
- Geometric prob. dis. (1) $E(X) = \sum_i x_i p(x_i) = \sum_i x \cdot (1-p)^{x-1} p$

$$(1-p) \cdot E(X) = (1-p) \cdot \sum_i x \cdot (1-p)^{x-1} p \quad / \cdot (1-p)$$

$$(1-p) \cdot E(X) = (1-p) \cdot (1 \cdot (1-p)^0 \cdot p + 2 \cdot (1-p)^1 \cdot p + 3 \cdot (1-p)^2 \cdot p + \cdots + n \cdot (1-p)^{n-1}$$

$$(2) \quad (1-p) \cdot E(X) = 1 \cdot (1-p)^1 \cdot p + 2 \cdot (1-p)^2 \cdot p + 3 \cdot (1-p)^3 \cdot p + \cdots + n \cdot (1-p)^n$$

$$((1)-(2)) \quad p \cdot E(X) = p \cdot \sum_{i=0}^{\infty} (1-p)^i$$

$$p \cdot E(X) = p \cdot \frac{1}{1-(1-p)}$$

$$p \cdot E(X) = p \cdot \frac{1}{p)}$$

$$p \cdot E(X) = \frac{p}{p}$$

$$p \cdot E(X) = 1 \quad / \div p$$

$$E(X) = \frac{1}{p}$$

- Binomial ran. var. $E[X] = \sum_i x_i p(x_i)$

$$
\begin{aligned}
E(X) &= \sum_{x=1}^{n} x \cdot \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \cdot \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} \cdot p^x \cdot (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n \cdot (n-1)!}{(x-1)!((n-1)-(x-1))!} \cdot p^x \cdot (1-p)^{n-x} \\
&= \sum_{x=1}^{n} n \cdot \binom{n-1}{x-1} \cdot p^x \cdot (1-p)^{n-x} \\
&= n \cdot \sum_{x=1}^{n} \binom{n-1}{x-1} \cdot p^x \cdot (1-p)^{n-x} \quad \text{subst.} y = x-1 \\
&= n \cdot \sum_{y=0}^{n-1} \binom{n-1}{y} \cdot p^{y+1} \cdot (1-p)^{n-(y+1)} \\
&= n \cdot p \sum_{y=0}^{n-1} \binom{n-1}{y} \cdot p^y \cdot (1-p)^{n-1-y} \quad \text{from binomial theorem} \\
&= n \cdot p \cdot (p + (1-p))^{n-1} \\
&= n \cdot p \cdot 1^{n-1} \\
&= n \cdot p
\end{aligned}
$$

Definition 3.3. negative binomial distribution
Number of of failures before the first success is a $Geo(p)$ random variable, and number of failures before the $r-th$ success (where $r \in \mathbb{N}$) is a $Neg-Bin$ (fun fact: $NB(1,p) = Geo(p)$): **neg Binomial** random variable X, denoted by $B(r,p)$, has two parameters - count r and probability p. It models the number of Bernoulli trials (with probability p of success) to the r-th success. Hence, it is a convolution of r geometric distributions.

$$
\begin{aligned}
p_X(x) &= P(X = x) \\
&= \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & \text{for } x \geq r, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

$$
E(X) = \sum_i x_i p_X(x_i) = \cdots = r \cdot \frac{1}{p}
$$

Definition 3.4. Linearity of expectation Let X and Y be random variables (**even DEPENDENT!**). Then

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

Definition 3.5. Expectation of convolution
Let X and Y be random variables (**even DEPENDENT!**). Then

$$\begin{aligned} E(X+Y) &= \sum_i \sum_j (i+j) \cdot P(X = i \wedge Y = j) \\ &= \sum_i i \sum_j P(X = i \wedge Y = j) + \sum_j j \sum_i P(X = i \wedge Y = j) \\ &= \sum_i i \cdot P(X = i) + \sum_j j \cdot P(Y = j) \\ &= E(X) + E(Y) \end{aligned}$$

Definition 3.6. Multiplication by a constant
Let X be random variable and c be a real number. Then

$$\begin{aligned} E(c \cdot X) &= c \cdot \sum_i x_i p_X(x_i) \\ &= c \cdot E(X) \end{aligned}$$

Corollary (linearity fo expectation) Let $X_1, X_2, \ldots X_n$ be random variables and $c_1, c_2, \ldots, c_n \in \mathbb{R}$ constants. Then:

$$E\left(\sum_{i=1}^{n} c_i \cdot X_i\right) = \sum_{i=1}^{n} c_i \cdot E(X_i)$$

Definition 3.7. Multiplication
Let X and Y be random variables (**only INDEPENDENT!**). Then

$$\begin{aligned} E(X \cdot Y) &= \sum_i \sum_j (i \cdot j) \cdot P(X = i \wedge Y = j) \\ &= \sum_i \sum_j (i \cdot j) \cdot P(X = i) \cdot P(Y = j) \\ &= \sum_i i \cdot P(X = i) \cdot \sum_j j \cdot P(Y = j) \\ &= E(X) \cdot E(Y) \end{aligned}$$

Definition 3.8. Theorem
If $\mathbf{Im}(X) \subseteq \mathbb{N}$ then $E(X) = \sum_{i=1}^{\infty} P(X \geq i)$

Definition 3.9. Markov inequality
It is important to derive as much information as possible even from a partial description of random variable... Let X be a **nonnegative random variable** with finite mean value $E(X)$. Then for all $t > 0$ (higher t means lower bound, t smaller then exp. val. doesn't make sense) it holds that

$$P(X \geq t) \leq \frac{E(X)}{t} \Leftrightarrow P(X \geq k \cdot E(X)) \leq \frac{1}{k}$$

Definition 3.10. Moments
The k^{th} **moment of a random variable** X is defined as $E(X^k)$.

If X and Y are random variables with matching coresponding moments of all orders, i.e. $\forall E(X^k) = E(Y^k)$ then X and Y have the same distributions.

Definition 3.11. k-th **central** moment
Centralization - $[X - E(X)]$ we shift the values such that the zero (0) of x and y axis is the expected value.

$$\mu_k = E\left([X - E(X)]^k\right)$$

Definition 3.12. variance (rozptyl)
The second (Squaring makes each term positive so that values above the mean do not cancel values below the mean) central moment is known as the **variance** of X ($Var(X)$ or σ_X^2) and defines as

$$\mu_2 = E\left([X - E(X)]^2\right) = E(X^2) - [E(X)]^2$$

If variance is small, then X takes values close to $E(X)$ with high probability. If the variance is large, then the distribution is more 'diffused'.

Definition 3.13. standard deviation (smerodatná odchylka)
The difference or advantage upon variance is in units, the standard deviation use same units as random variable (but variance is using squared units)

$$\sqrt{\sigma_X^2} = \sigma_X$$

4.

Definition 4.1. Variance - properties
Let X be a random variable and a and b be real numbers. Then

$$\begin{aligned} Var(aX+b) &= E((aX+b) - E(aX+b))^2 \\ &= E(aX+b-aE(X)-b)^2 \\ &= E(aX-aE(X))^2 \\ &= E(a(X-E(X)))^2 \\ &= E(a^2(X-E(X))^2) \\ &= a^2E([X-E(X)]^2) \\ &= a^2 \cdot Var(X) \end{aligned}$$

Definition 4.2. **Variance** of some distributions

Distribution	Probability	$E(X)$	$Var(X)$
Uniform	$\frac{1}{n}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{n^2-1}{12}$
Constant	$P(c)=1$	c	0
Bernoulli	$\{0,1\}$, $P(1)=p$	p	$(1-p)$
Binomial	n-trials with p-prob.	$n \cdot p$	$n \cdot p \cdot (1-p)$
Geometric	p-trials (1) first succ.	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Neg. binomial	p-trials (r) succ.s	$\frac{r}{p}$	$\frac{r\cdot(1-p)}{p^2}$

Definition 4.3. Generalized additivity of variance
Let $X_1, X_2, \ldots, X_n$ be **pairwise independent** (or not independent) variables and constants $a_1, a_2, \ldots, a_n, b \in \mathbb{R}$ as

$$Var\left(\sum_{i=1}^{n} a_iX_i + b\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2 \cdot \sum_{1\le i<j\le n} a_i \cdot a_j \cdot Cov(X_i, X_j)$$

Becareful between $Var(X_1+X_1)$ and $Var(X_1+X_2)$ (slide 17/30, 4 lecture).

Definition 4.4. motivation for **covariance**
Is true that $Var(X+Y) = Var(X) + Var(Y)$?

$$\begin{aligned} Var(X+Y) &= E[(X+Y)^2 - (E[X+Y])^2] \\ &= E[(X^2+2XY+Y^2) - (E[X]+E[Y])^2] \\ &= E[X^2] + 2E[XY] + E[Y^2] - ((E[X])^2 + 2E[X]E[Y] + (E[Y])^2) \\ &= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y])^2) + 2(E[XY] - E[X]E[Y]) \\ &= Var(X) + Var(Y) + 2(E[XY] - E[X]E[Y]) \quad \text{no it is } \textbf{not} \text{ TRUE} \\ &= Var(X) + Var(Y) + 2Cov(X,Y) \quad \text{but for indipendent?} \\ &= Var(X) + Var(Y) + 2(E[XY] - E[X]E[Y]) \\ &= Var(X) + Var(Y) + 2(E[X]E[Y] - E[X]E[Y]) \\ &= Var(X) + Var(Y) \end{aligned}$$

Definition 4.5. Covariance
is a measure of how much two random variables change together. In other words, covariance is a measure of the strength of the correlation between two random variables. Also, it can be considered as a generalization of the concept of the variance of two random variables.
Covariance of X and Y is denoted $Cov(X, Y)$; Centralize (shift values) of both random variables, multiply them and get expected value:

$$E[(X - E[X])\cdot(Y - E[Y])] = \sum_{i,j} p_{x_i,y_j}(x_i = E[X]) \cdot (y_i - E[Y])$$

Another definition:

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

Corollary for **independent** random variables X and Y, it holds that $Cov(X, Y) = 0$! BUT:

$$\text{X and Y are independent} \Rightarrow Cov(X, Y) = 0$$
$$\text{X and Y are independent} \not\Leftarrow Cov(X, Y) = 0$$

Covariance measures **linear dependence** between two random variables. It is positive if the variables are "correlated", and negative when "anti-correlated".
E.g. $Y = aX$, remember $(Cov(X, X) = Var(X))$

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X]) \cdot (Y - E[Y])] \\ &= E[(X - E[X]) \cdot (aX - E[aX])] \\ &= E[(X - E[X]) \cdot (aX - aE[X])] \\ &= E[(X - E[X]) \cdot a(X - E[X])] \\ &= a \cdot E[(X - E[X]) \cdot (X - E[X])] \\ &= a \cdot Var(X) \end{aligned}$$

Definition 4.6. correlation coefficient
We define as the NORMALIZED covariance, i.e.

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

Definition 4.7. The **conditional** probability distribution
of random variable Y given random variable X (their joint distribution is $p_{Y,X}$) is

$$\begin{aligned} p_{Y,X}(y \mid x) &= P(Y = y \mid X = x) \\ &= \frac{P(Y = y, X = x)}{P(X = x)} \\ &= \frac{p_{X,Y}(x, y)}{p_X(x)} \end{aligned}$$

Definition 4.8. conditional expectation We may consider $Y \mid X = x$ to be a neww random variable that is given by conditional probability distribution $p_{Y|X}$.

expectation:

$$\begin{aligned} E[Y \mid X = x] &= \sum_y y \cdot P(Y = y \mid X = x) \\ &= \sum_y y \cdot p_{y|x}(y \mid x) \\ &= \sum_y y \cdot \frac{p_{y,x}(y, x)}{p_x(x)} \end{aligned}$$

variance:

$$Var(Y \mid X = x) = E[Y^2 \mid X = x] - (E[Y \mid X = x])^2$$

Definition 4.9. Theorem of **total expectation**
Let X, Y be random variable, then

$$E(Y) = \sum_x E[Y \mid X = x] \cdot p_{X(x)}$$

Definition 4.10. Theorem of **total moments**
Let X, Y be random variable, then

$$E(Y^k) = \sum_x E[Y^k \mid X = x] \cdot p_{X(x)}$$

Doesn't hold for CENTRAL moments, e.g. Variance!

Definition 4.11. Chebyshev inequality
In case we know both MEAN VALUE ($E(X)$) and VARIANCE ($Var(X)$). Let X be a random variable (we don't need non negative) with finite variance. Then

$$P[\mid X - E(X) \mid \geq t] \leq \frac{Var(X)}{t^2}, t > 0$$

How we obtain equation above? It is MARKOV INEQUALITY to the non negative variable $[X - E(X)]^2$ and we replace t by t^2. We use the fact that $[(X - E(X))^2 \geq t^2] = [\mid X - E(X) \mid \geq t]$

5.

Definition 5.1. idea

1. Markov inequality uses $E(X)$
2. Chebyshev inequality uses $Var(X)$
3. Chernoff bound will use **moment generating function**

Definition 5.2. the **Moment generating function** of a random variable X is

$$M_X(t) = E(e^{t \cdot X})$$

So let $M_X(t) = E(e^{tX})$ be a moment generating function of X. Assuming that exchanging the expectation and differentation operands is legitimate, $\forall n, n > 1$ we have

$$E(X^n) = M_X^{(n)}$$

Proof.

$$\begin{aligned} M_X^{(n)}(t) &= (E[e^{tX}])^{(n)} \\ &= E[(e^{tX})^{(n)}] \\ &= E[X^{(n)} e^{tX}] \end{aligned}$$

□

Definition 5.3. Let X and Y be two random varaibles and ... slide 9/29 5 lecture.

Definition 5.4. If X a Y are **independent** random variables, then

$$\begin{aligned} M_{X+Y}(t) &= E(e^{t(X+Y)}) \\ &= E(e^{tX} \cdot e^{tY}) \\ &= E(e^{tX}) \cdot E(e^{tY}) \\ &= M_X(t) \cdot M_Y(t) \end{aligned}$$

Definition 5.5. Chernoff bounds (not inequality) for random variable X is obtained by applying the MARKOV INEQUALITY to e^{tX} for some suitably chosen t.

$$P(X \geq a) = P(tX \geq ta) = P(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}}$$

for any $t > 0$ or $t < 0$.

Definition 5.6. Corollary Let $X_1, \ldots, X_n$ be independent Poisson trial and $X = \sum_{i=1}^{n} X_i$. For $0 < \delta < 1$,

$$P(\mid X - E(X) \geq \delta E(X)) \leq 2e^{\frac{-E(X)\delta^2}{3}}$$

Definition 5.7. idea (law of large numbers) proves formally that the mathematically defined statement of probability corresponds to our motivation based on frequencies in repeated experiments, i.e. the expected value is really "expected" as the average of the repeated experiments.

If $X_1, X_2, \ldots$ is a sequence of **independent** identically distributed random variables with expectation $\mu = E(X_k)$, then for every $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) = 0$$

The probability that average of X_i differs from the expectation by more than arbitrarily small ϵ goes to 0

Converges in probability - the probability of being close enough converges to 1. If it is long enough, the p

Definition 5.8. (strong) law of large numbers. If $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables with expectation $\mu = E(X)$, then

$$P\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1$$

Converges **with** probability 1.

6.

Definition 6.1. Stochastic process is a collection of random variables $X = \{X_t | t \in T\}$ (parametrized by time). The index t often represents time; $X_t = y$ is called state of X at time t.
Classification:
time:

- **discrete**-time
- **continuous**-time (customers in a queue in the shop)

state:

- **finite**-state (printer)
- **discrete**-state (whatever)
- **continuous**-state (cpu usage)

How does the sample space look like? Every *sample point* is a completle different "univese" e.g. $EIPPIIIP$, $IIIPPPPEEE$...(printer, state $\in P_{rinting}, S_{ending}, I_{dleing}$)

Definition 6.2. Markov chain is a *discrete-time* stochastic process $\{X_0, X_1, X_2, \dots\}$ if:

$$P(X_t = a \mid X_{t-1} = b, X_{t-2} = a_{t-2}, \dots, X_0 = a_0) = P(X_t = a \mid X_{t-1} = b) = p_{a,b}$$

Markov or **memoryless** property is, the value of X_t should depend on the value of X_{t-1}, but does **not depend on history** of how we arrived at X_{t-1}.
Markov chain can be drawn as an **automaton**. Vertices are states (values) and edge from a to b has label $p_{a,b}$.
If the Markov chains whose "state spaces" (values that X_t can take) are finite, mark them from $1, \dots, n$. We can define probability P with **transition matrix**:

$$P = [p_{i,j} (= P(X_t = j) \mid X_{t-1} = i) \mid 1 \leq i, j \leq n] = \begin{bmatrix} p_{1,1} & \dots & p_{1,n} \\ \vdots & \ddots & \vdots \\ p_{n,1} & \dots & p_{n,n} \end{bmatrix}$$

$$(\textit{every row equals to }) \sum_{j=1}^{n} p_{i,j} = P(X_t = ? \mid X_{t-1} = i) = 1$$

Question: $P(A, B, C, D)$?
Answer (TRICK is to know $P(A \cap B \mid C) = P(B \mid C) \cdot P(A | C \cap B)$):

$$\begin{aligned} P(A, B, C, D) &= P(A \cap B \cap C \cap D) \\ &= P(A) \cdot P(B \cap C \cap D \mid A) \\ &= P(A) \cdot P(B|A) \cdot P(C \cap D \mid A \cap B) \\ &= P(A) \cdot P(B|A) \cdot P(C \cap D \mid B) \\ &= \quad \vdots \\ &= P(A) \cdot P(B|A) \cdot P(C|B) \cdot P(D|C) \end{aligned}$$

Whats the probability that in time t we are in state i - $\lambda_i(t) = P(X_t = i)$. $\vec{\lambda}(t) = (\lambda_1(t), \lambda_2(t), \ldots, \lambda_n(t))$, and $\lambda(\vec{0})$ is the initial distribution.

Definition 6.3. alternative
Every (discrete-time finite-state) Markov chain can be alternatively (uniquely) define by an initial vector $\vec{\lambda}(0)$ and a a transition matrix P.

Definition 6.4. m-step
For any m we define the m-step transition matrix $P^{(m)}$ such that:

$$p_{i,j}^{(m)} = P(X_{t+m} = j \mid X_t = i)$$

Thus, for any $t \geq 0$ and $m \geq 1$ we have

$$\vec{\lambda}(t+m) = \vec{\lambda}(t)P^m$$

Definition 6.5. The **hitting time** (is a random variable) of a subset A of states of a Markov chain is a random variable $H^A : \mathbf{S} \to \{0, 1, 2, \ldots\} \cup \{\infty\}$ given by:

$$H^A = inf\{n \geq 0 \mid X_n \in A\}$$

where $\mathbf{S}$ is the sample space of the Markov chain and ∞ is infimum of $\emptyset$.
We are looking for first index of X_i, when we are, where we want to be! e.g. in state 1, after how many steps, we reach fourth state?

Definition 6.6. Starting from a state i, the **probability of hitting** A is:

$$h_i^A = P(H^A < \infty \mid X_0 = i)$$

and the **mean time**(steps) taken to reach a set of states A is

$$k_i^A = E(H^A \mid X_0 = i) = \sum_{n<\infty} n \cdot P(H^A = n) + \infty \cdot P(H^A = \infty)$$

Definition 6.7. The vector of **hitting probabilities** $h^A = h_1^A, h_2^A, \ldots$ is the minimal non-negative solution to the system of linear equations.

$$h_i^A = \begin{cases} 1 & \text{for } i \in A \\ \sum_j p_{i,j} \cdot h_j^A & \text{for } i \notin A \end{cases}$$

probability of reaching every potential next state j, from state i $\cdot$ probability of reaching some of the A states from potential next state j

Definition 6.8. The vector of **mean hitting times** $k^A = k_1^A, k_2^A, \ldots$ is the minimal non-negative solution to the system of linear equations.

$$k_i^A = \begin{cases} 0 & \text{for } i \in A \\ 1 + \sum_{j \notin A} p_{i,j} \cdot k_j^A & \text{for } i \notin A \end{cases}$$

We need to proceed ONE step and count all the possibilities where we can go.

Definition 6.9. Transient analysis

- ∗ distribution after k-steps
  - ♢ the k-th power of the transition matrix P
- ∗ reaching/hitting probability
  - ♢ equations for hitting probabilities h_i
- • (**mean**) hitting time k_i [#steps]
  - ♢ equations for hitting probabilities k_i

7.

Definition 7.1. absorbing
A state of a Markov chain is said to be **absorbing** iff it cannot be left, once it is entered; *i.e.* : $p_{i,j} = 1$ (self loop). Special case of recurrent MC.

Definition 7.2. recurrent
A state i of a Markov chain is said to be **recurrent** iff, starting from state i, the process eventually returns to state i with probability 1.

$$P(i \overset{+}{\Longrightarrow} i) = 1.$$

In a **finite-state** Markov chain, each recurrent state is almost either **not visited** or **visited infinitely many times**; because if it is visited, then it is revisited with probability one. Hence in an infinite run, it is visited infinitely many times with probability one.

Definition 7.3. Theorem
In a **finite-state** Markov chain, a state is **recurrent** if and only if it is in a **bottom strongly connected component** of the Markov chain graph representation. All other states are **transient**.

Definition 7.4. transient
A state of a Markov chain is said to be **transient** (or non-recurrent) iff there is a positive probability that the process will not return to this state.

$$P(i \overset{+}{\Longrightarrow} i) < 1.$$

Every transient state is visited **finitely many times** almost surely (with probability 1).

$$\begin{aligned} P(\text{finit. many visits} \mid \text{start. in } i) &= \text{visit once} + (\text{visit twice}) + \text{visit thrice} + \dots \\ &= \text{visit last time} + (\text{return back} \cdot \text{visit last t.}) + ((\text{ret. back})^2 \cdot \text{visit last t.}) + \dots \\ &= p + ((1-p) \cdot p) + ((1-p)^2 \cdot p) + \dots \\ &= p \cdot \left(\frac{1}{1-(1-p)} \right) \\ &= p \cdot \frac{1}{p} \\ &= 1 \end{aligned}$$

The pitcher goes so often to the well, that is is broken at last!

Definition 7.5. A Markov chain is said to be **irreducible** if every state can be reached from every other state in a finite number of steps, i.e.:

$$\forall i, j \; P(i \overset{+}{\Longrightarrow} j) > 0$$

A Markov chain is **irreducible** if and only if its graph representation is a *single strongly connected component*.

Corollary:
All states of a **finite-state** irreducible Markov chain are recurrent!

Definition 7.6. equilibirum
Let P be the transition matrix of a Markov chain and $\vec{\lambda}$ be a probability distribution on it's states. If

$$\vec{\lambda} \cdot P = \vec{\lambda}$$

then $\vec{\lambda}$ is a **stationary** (steady-state, invariant, **equilibrium**) distribution of the Markov chain.

Definition 7.7. If a finite-state Markov chain is irreducible then there is a unique stationary distribution.

Definition 7.8. For each finite-state Markov chain, there is stationary distribution.

Definition 7.9. Let P be a transition matrix of a finite-state Markov chain and $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$ be a stationary distribution corresponding to P. For any state i of the Markov chain, we have (want probability of coming to state i should be equal to probability of leaving state i, we omit self loops - $i = j$):

$$\sum_{j \neq i} \pi_j \cdot P_{i,j} = \sum_{j \neq i} \pi_i \cdot P_{j,i}$$

Definition 7.10. Stationary distributions

$$\mathrm{P} = \begin{bmatrix} (1-p) & p \\ q & (1-q) \end{bmatrix}$$

$$\pi_1(1-p) + \pi_2 q = \pi_1$$
$$\pi_2(1-q) + \pi_1 p = \pi_2$$
$$\pi_1 = \pi_2$$

or with cut-sets:

$$\pi_1 \cdot p = \pi_2 \cdot q$$

Definition 7.11. Expected long-run frequency
let us have a finite-state irreducible Markov chain and the unique stationary distribution $\vec{\pi}$. It holds that:

$$\pi_i = \lim_{n \to \infty} \frac{E(\text{number of visits of the state } i \text{ during the first } n \text{ steps})}{n}$$

It means, that in the limit, the real flow, or weight of each state, shows. There are many paths in Stochastic process, but all not the true ones are with probability 0.

Definition 7.12. Mean inter visit time
Let us have a finite-state irreducible Markov chain and the unique stationary distribution $\vec{\pi}$. It holds that

$$\pi_i = \frac{1}{m_i}$$

where m_i is the **mean inter visit time** (or expected return time) of state i.

Definition 7.13. A state j in a Markov chain is **periodic** if there exists an integer $\Delta \geq 2$ such that $P(X_{t+s} = j \mid X_t = j) = 0$ unless s is divisable by Δ. Periodic means $gcd(path_1, path_2, \ldots) < 2$. A Markov chain is periodic if any state in the chain is periodic. A state or chain that is not periodic is **aperiodic**.

Definition 7.14. Limit of transient distributions
Let us have a finite-state aperiodic irreducible Markov chain and the unique stationary distribution $\vec{\pi}$. It holds that

$$\vec{\pi} = \lim_{n \to \infty} \vec{\lambda} \cdot P^n$$

where $\vec{\lambda}$ is an arbitrary distribution on state (it doesn't matter where we start)

Definition 7.15. Infinite-state Markov chain

- It is **no longer true** that each Markov chain has a stationary distribution.
- It is **no longer true** that all states are recurrent in irreducible Markov chain.
- A state can be recurrent and the mean inter visit time can be infinite.

- In discrete time
  - distribution on **where** we go in the next step
- In continuous time
  - distribution on **WHEN** we do next step
  - distribution on **where** we go in the next step

Definition 8.1. Continuous-time variant of the geometric distribution - we want **Markovian property**, which is: subsequent behaviour depends **on the current state only** so neither on **where** we were going nor **when** we did the (last) step(s). It also does not depend on the **time we have been waiting** for the next step!

Definition 8.2. Exponential distribution

- one parameter λ (called **rate**)
- expected value $\frac{1}{\lambda}$
- probability density function:
$$f(t) = \lambda \cdot \frac{1}{e^{\lambda \cdot t}}$$
- cumulative distribution function:
$$F(t) = \int_0^t \lambda \cdot \frac{1}{e^{\lambda \cdot t}} dx = 1 - \frac{1}{e^{\lambda \cdot t}}$$

Theorem:
For an exponentially distributed random variable X and every $t, t_0 \geq 0$, it holds that

$$P(X > t_0 + t \mid X > t_0) = P(X > t)$$

$P(X > t) = 1 - F(t) = \frac{1}{e^{\lambda \cdot t}}$

Definition 8.3. Continuous-Time Markov chain CTMC
is an event-driven system with exponentially distributed events.

- What is the mean waiting time in a state?
  - waiting time for exponentially distributed E_1, E_2 with rates λ_1, λ_2; waiting for the first from these two distributions $= min(E_1, E_2)$ is equal to $P(min(E_1, E_2) > t) = \frac{1}{e^{\lambda_1+\lambda_2 \cdot t}}$. So the waiting time is *exponentionaly distributed* with rate $\lambda_1 + \lambda_2$
  - mean waiting time is $\frac{1}{\lambda}$, where $\lambda = \lambda_1 + \lambda_2 + \ldots$.
- What is probability that a particular event wins in a state?
  - probability that E_1 wins, means $P(E_1 < E_2) = \frac{\lambda_1}{(\lambda_1+\lambda_2)}$
- How the waiting time depends on the wining event?
  - probability that E_1 wins for a given winning time: $\forall t \,.\, P(E_1 < E_2 \mid min(E_1, E_2) = t) = \frac{\lambda_1}{(\lambda_1+\lambda_2)}$. So waiting time is independent on event that comes!
- How the winning event depends on the waiting time?

Imagine you have a number of lily pads in a pond. Each lily pad is a "state of the system". A single frog starts on one of the lily pads (maybe one chosen at random).

At any moment the frog may jump from one lily pad to another. The rate of jumping from any pad to any other is known. No matter how long the frog has been on a lily pad, its rate of jumping to the other possible pads in a short interval of time is constant. Hence you need just a single "rate of jumping" from every pad to every other pad. Some of the rates may be zero indicating the frog cannot jump that far.

It turns out that the amount of time the frog spends on any one pad before it jumps away is a random variable with exponential distribution (just like the time for a radioactive atom to decay spontaneously).

Under reasonable conditions we can work out the long-run probability of the frog being on any lily pad at a randomly chosen instant of time.

Definition 8.4. CTMC is defines by (an initial distribution $\lambda(0)$) and a rate matrix Q where $Q[i, j]$ is the rate of an event leading from state i to state j.

Definition 8.5. A distribution π on states of a CTMC is called **stationary** iff $\pi = \lambda(t)$ for every time $t \geq 0$, when starting the CTMC in $\pi = \lambda(0)$

Definition 8.6. The rate $\lambda = \lambda_1 + \lambda_2 + \lambda_3$ is called **exit rate** of the state s and the probabilities p_1, p_2, p_3 are called **exit probabilities** ($p_i = \frac{\lambda_i}{\lambda}$).

Definition 8.7. important questions for a given CTMC with $\lambda(0)$ and Q are:

- What is the distribution on states at a given time?
  - For **uniformed** CTMC, we use **DTMC** matrix and the **Poisson distribution** to **approximate** it.

- Does the distribution converge to a stationary distribution?
  - YES, for **finite-state CTMC** there is always a stationary distribution π and the distributions $\lambda(t)$ converges to π almost surely. There is also no problem with periodic Markov chains.
- What is continuous (time) frequency, i.e. fraction of time spent?
  - Invariant distribution for DTMC rebalanced using waiting times
  - For **uniformed** CTMC, it is the invariant distribution of DTMC.
- What is the queue-sever **utilization** (e.g. what portion of time is serve running? complement of [not running] = q_0) and the mean queue length?
  - Using stationary distribution π of the CTMC, the utilization is $1 - \pi_0$ and the mean queue length is defined by $\sum_s s \cdot \pi_s$.
    The **mean queue length** is equal to $\lim_{t\to\infty} E(X_t)$ X_t = length of given queue

$*$ Every CTMC can be transformed to the one with **uniform** exit rates. (Find a state with the highest exit rate and added appropriate **self-loops** to all others.)

- Uniform exit rates allow for easier discretization
- enables analysis of distribution on states at a given time

9.

Definition 9.1. Uncertainty

- Let us fix the number of outcomes of an experiment and compare the uncertainty of different probability distributions. Natural requirement is that **the most uncertain is the experiment with the uniform probability distribution**, i.e. $H(p_1, \ldots, p_n)$ is maximal for $p_1 = \cdots = p_n = \frac{1}{n}$.
- **Permutation** of probability distribution **does not change** the uncertainty
- Uncertainty should be **nonnegative** and equals to zero if and only if we are sure about the outcome of the experiment.
- If we include into an experiment an **outcome with zero probability**, this does not change our uncertainty
- As justified before, having the uniform probability distribution on n outcomes **cannot be more uncertain** than having the uniform probability distribution on $n+1$ outcomes.
- $H(p_1, \ldots, p_n)$ is a **continuous** function of its parameters.
- Uncertainty of an experiment consisting of a **simultaneous** throw of m and n sided die is as uncertain as an **independent** throw of m and n sided die implying

$$H(\overbrace{1/(mn), \ldots, 1/(mn)}^{mn\times}) = H(\overbrace{1/m, \ldots, 1/m}^{m\times}) + H(\overbrace{1/n, \ldots, 1/n}^{n\times})$$

- Let us consider a random choice of one of $n+m$ balls, m being red and n being blue. Let $p = \sum_{i=1}^{m} p_i$ be the probability that a red ball is chosen and $q = \sum_{i=m+1}^{m+n} p_i$ be the probability that a blue one is chosen. Then the uncertainty which ball is chosen is the uncertainty whether red or blue ball is chosen **plus weighted uncertainty** that a particular ball is chosen provided blue/red ball was chosen. Formally,

$$\begin{aligned} &H(p_1, \ldots, p_m, p_{m+1}, \ldots, p_{m+n}) = \\ =&H(p, q) + pH\left(\frac{p_1}{p}, \ldots, \frac{p_m}{p}\right) + qH\left(\frac{p_{m+1}}{q}, \ldots, \frac{p_{m+n}}{q}\right) \end{aligned}$$

- It can be shown that any function satisfying Axioms $1-8$ is of the form

$$H(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log_a p_i = -(\log_a 2) \sum_{i=1}^{m} p_i \log_2 p_i$$

the function is defined uniquely up to multiplication by a constant, which effectively changes only the **base of the logarithm**. The base specifies units in which we measure the entropy (for bits, we use $\log_2$).

Definition 9.2. Entropy
Let X be a random variable with a probability distribution $p(x) = P(X = x)$. Then the (Shannon) entropy of the random variable X is defined as

$$\begin{aligned} H(X) &= -\sum_{x \in \mathrm{lm}(X)} p(x) \cdot \log p(x) \\ &= \sum_{x \in \mathrm{lm}(X)} p(x) \cdot \log \frac{1}{p(x)} \\ &= E_p\left[\log \frac{1}{p(X)}\right] \\ &= -E_p[\log p(X)] \end{aligned}$$

Definition 9.3. Let X be a random variable with a probability distribution $p(x) = P(X = x)$. Then

$$H(X) \geq 0$$

Proof:
As $p(x)$ is a probability distribution, we know that $\forall x \; p(x) \leq 1$. Then $\log p(X) \leq 0$, and so $E_p[\log p(X)] \leq 0$. Hence,

$$H(X) = -E_p[\log p(X)] \geq 0$$

Definition 9.4. Let X and Y be random variables with a joint probability distribution $p(x, y) = P(X = x, Y = y)$ We define the **joint (Shannon) entropy** of X and V as

$$H(X, Y) = -\sum_{x \in \mathrm{lm}(X) y \in \mathrm{lm}(Y)} p(x, y) \log p(x, y)$$

or, alternatively,

$$H(X, Y) = -E_p[\log p(X, Y)] = E_p\left[\log \tfrac{1}{p(X,Y)}\right]$$

Definition 9.5. conditional entropy given $Y = y$
How *uncertain* we are about an outcome of a random variable X given a particular outcome y of a random variable Y. Let X and Y be random variables and $y \in \mathrm{Im}(Y)$. The conditional entropy of X given $Y = y$ is

$$H(X \mid Y = y) = -\textstyle\sum_{x \in \mathrm{lm}(X)} P(X = x \mid Y = y) \log P(X = x \mid Y = y)$$

Definition 9.6. conditional entropy given Y
Let X and Y be random variables with a orangejoint probability distribution $p(x, y) = P(X = x, Y = y)$. Let us denote $p(x \mid y) = P(X = x \mid Y = y)$ The conditional entropy of X given Y is

$$\begin{aligned} H(X \mid Y) &= \sum_{y \in \mathrm{lm}(Y)} p(y) H(X \mid Y = y) \\ &= -\sum_{y \in \mathrm{lm}(Y)} p(y) \sum_{x \in \mathrm{lm}(X)} p(x \mid y) \log p(x \mid y) \\ &= -\sum_{x \in \mathrm{lm}(X) y \in \mathrm{lm}(Y)} p(x, y) \log p(x \mid y) \\ &= -E_p[\log p(X \mid Y)] \end{aligned}$$

Definition 9.7. Chain rule of conditional entropy
Let X and Y be random variables. Then

$$H(X,Y) = H(Y) + H(X|Y)$$

size of information in Y + size of information remains in X whenever Y is revealed
COROLLARY:
Let X, Y, and Z be random variables.

$$H(X,Y|Z) = H(Y|Z) + H(X|Y,Z)$$

Definition 9.8. cross-entropy, which measures "entropy" of q in the probability space defined by p. The cross-entropy of the distribution q relative to a distribution p (on a common set of sample points $\mathrm{Im}(X)$) is defined as

$$H_p(q) = -\sum_{x \in \mathrm{lm}(X)} p(x) \log q(x) = -E_p[\log q(X)]$$

Definition 9.9. relative entropy
The relative entropy or **Kullback-Leibler divergence** between two probability distributions $p(x)$ and $q(x)$ (on a common set of sample points $\mathrm{Im}(X)$) is defined as

$$D(p\|q) = \sum_{x \in \mathrm{lm}(X)} p(x) \log \frac{p(x)}{q(x)} = E_p\left[\log \frac{p(X)}{q(X)}\right] = H_p(q) - H_q(p)$$

and describes how different B is from A from the perspective of A.

Definition 9.10. Mutual information
measures the *shared information* between two random variables. It is the decrease of the uncertainty about an outcome of a random variable given an outcome of another random variable. If their are independent mutual information is zero. More depedence more shared information...
Let X and Y be random variables with a joint distribution p. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product of marginal distributions p_X and p_Y

$$I(X;Y) = D\left(p\|(p_X \cdot p_Y)\right) = E_p\left[\log \frac{p(X,Y)}{p_X(X)p_Y(Y)}\right]$$

Definition 9.11. Let X and Y be random variables. Then

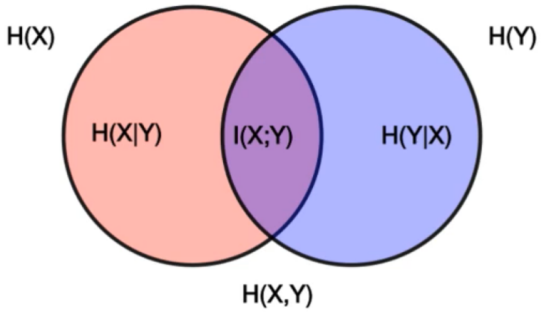
$$I(X;Y) = H(X) - H(X|Y)$$

information in X - size of information remains in X whenever Y is revealed = *shared information* between two random variables Proof:

$$\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y) \log \frac{p(x \mid y)}{p(x)} = \\
&= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x \mid y) = \\
&= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x \mid y)\right) = \\
&= H(X) - H(X \mid Y)
\end{aligned}$$

Corollary (inclusion-exclusion):

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$



Definition 9.12. Properties of entropy and mutual information

1. (general chain rule for entropy)
   - Let $X_1, X_2, \ldots, X_n$ be random variables. Then
   $$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_{n-1}, \ldots, X_1)$$
   - (PROOF) We use repeated application of the chain rule for a pair of random variables
   $$\begin{aligned} H(X_1, X_2) &= H(X_1) + H(X_2 \mid X_1) \\ H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3 \mid X_1) = \\ &= H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_2, X_1) \\ &\vdots \\ H(X_1, X_2, \ldots, X_n) &= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_{n-1}, \ldots, X_1) = \\ &= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1) \end{aligned}$$
2. (condition mutual information)
   - Let X, Y, and Z be random variables. The conditional mutual information between X and Y given $\mathbf{Z} = \mathbf{z}$ is
   $$I((X;Y) \mid Z = z) = H(X \mid Z = z) - H(X \mid Y, Z = z)$$
   The conditional mutual information between X and Y given Z is
   $$I((X;Y) \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = E\left[\log \frac{p(X, Y \mid Z)}{p(X \mid Z)p(Y \mid Z)}\right]$$
   where the expectation is taken over the joint distribution $p(x, y, z)$.

3. (chain rule for mutual information)

$$I\left((X_1, X_2, \ldots, X_n); Y\right) = \sum_{i=1}^{n} I\left(X_i; Y \mid X_{i-1}, \ldots, X_1\right)$$

4. (the conditional relative entropy) is the average of the relative entropies between the conditional probability distributions $p_{Y|X}$ and $q_{Y|X}$ averaged over the probability distribution p_X. Formally,

$$D\left(p_{Y|X} \| q_{Y|X}\right) = \sum_X p_X(x) \sum_y p_{Y|X}(y \mid x) \log \frac{p_{Y|X}(y \mid x)}{q_{Y|X}(y \mid x)} = E_p\left[\log \frac{p_{Y|X}}{q_{Y|X}}\right]$$

   where p in the base of the expectation is the joint distribution of X and Y.

5. (Chain rule for relative entropy)

$$D(p(x,y) \| q(x,y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x))$$

Definition 9.13. Let $p(x)$ and $q(x)$, be two probability distributions. Then

$$D(p \| q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Definition 9.14. Information inequality Let $p(x)$ and $q(x)$, be two probability distributions. Then

$$D(p \| q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Corollary:

- For any two random variables X, Y

$$I(X;Y) \geq 0$$

  with equality if and only if X and Y are independent.

- $$D(p(y|x) \| q(y|x)) \geq 0$$

  with equality if and only if $p(y|x) = q(y|x)$ for all y and x with $p(x) > 0$.

- $$I(X;Y|Z) \geq 0$$

  with equality if and only if X and Y are conditionally independent given Z.

Definition 9.15. Let X be a random variable, then

$$H(X) \leq log|Im(X)|$$

with equality if and only if X has a uniform distribution over $Im(X)$.

Definition 9.16. Conditioning **reduces** entropy
Let X and Y be a random variables, then

$$H(X \mid Y) \leq H(X)$$

with equality if and only if X and Y are independent.
Proof. $0 \leq I(X;Y) = H(X) - H(X \mid Y)$ with equality iff X and Y are independent.

Definition 9.17. (Independence bound on entropy) Let $X_1, X_2, \ldots, X_n$ be random variables. Then

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if X_i 's are mutually independent.
Proof. We use the chain rule for entropy

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i)$$

where the inequality follows directly from the previous theorem. We have equality if and only if X_i is independent of all $X_{i-1}, \ldots, X_1$.

10.

Definition 10.1. A code C for a random variable (memoryless source) X is a mapping $C : \mathrm{lm}(X) \to \mathbf{D}^*$, where $\mathbf{D}^*$ is the set of all finite length strings over the alphabet $\mathbf{D}$, with $|\mathbf{D}| = d$. $C(x)$ denotes the codeword assigned to x and $I_C(x)$ denotes the length of $C(x)$

Definition 10.2. The **expected length** $L_C(X)$ of a code C for a random variable X is given by

$$L_C(X) = \sum_{x \in \mathrm{lm}(X)} P(X = x) I_C(x) = E\left[I_C(X)\right]$$

In what follows, we will assume (WLOG) that the alphabet is $\mathbf{D} = \{0, 1, \ldots, d-1\}$

Definition 10.3. A code
C is said to be non-singular if it maps every element in the range of X to different string in $\mathbf{D}^*$, i.e.

$$\forall x, y \in \mathrm{lm}(X) x \neq y \Rightarrow C(x) \neq C(y)$$

Non-singularity allows **unique decoding of any single codeword**, however, in practice we send a sequence of codewords and require the complete sequence to be uniquely decodable. We can use e.g. any non-singular code and use an extra symbol $\# \notin \mathbf{D}$ as a **codeword separator**. However, this is very inefficient and we can improve efficiency by designing **uniquely decodable or prefix code**.

Definition 10.4. An **extension** C^* of a code C is the mapping from $\mathrm{Im}(X)^+$ (set of all nonempty strings over the alphabet $\mathrm{lm}(X)$) to $\mathbf{D}^*$ defined by

$$C^*\left(x_1 x_2 \ldots x_n\right) = C\left(x_1\right) C\left(x_2\right) \ldots C\left(x_n\right)$$

where $C\left(x_1\right) C\left(x_2\right) \ldots C\left(x_n\right)$ is concatenation of corresponding codewords.

Definition 10.5. A code is uniquely decodable iff its extension is non-singular.A code is uniquely decodable if any encoded string has only one possible source string.

Definition 10.6. A code is called **prefix code** (or instantaneous code) if **no codeword is a prefix of any other codeword**. The advantage of prefix codes is not only their unique decodability, but also the fact that a codeword can be decoded as soon as we read its last symbol.

Definition 10.7. KRAFT inequality
For any prefix code over an alphabet of size d, the codeword lengths (including multiplicities) $I_1, l_2, \ldots I_m$ satisfy the inequality

$$\sum_{i=1}^{m} d^{-l_i} \leq 1$$

Conversely, given a sequence of codeword lengths that satisfy this inequality, there exists a prefix code with these codeword lengths. Proof: use d-tree!, when building start from the shortest codewords.

Definition 10.8. McMillan Inequality *Kraft inequality* holds also for codes with countably infinite number of codewords.There exist **uniquely decodable codes** that are not prefix codes, but, as established by the following theorem, the Kraft inequality applies to general uniquely decodable codes as well and, therefore, when searching for an **optimal code** it suffices to concentrate on prefix codes. General uniquely decodable codes offer no extra codeword lengths in contrast to prefix codes. Theorem **(McMillan inequality)**: The codeword lengths of any **uniquely decodable** code must satisfy the Kraft inequality, i.e.

$$\sum_i d^{-l_i} \leq 1$$

Conversely, given a set of codeword lengths that satisfy the inequality it is possible to construct a uniquely decodable code with these codeword lengths.

Definition 10.9. The **expected length of any prefix** d **-ary code** C for a random variable X is greater than or equal to the entropy $H_d(X)$(d is the base of the logarithm), i.e.

$$L_C(X) \geq H_d(X)$$

with equality iff for all $x_i P\left(X = x_i\right) = p_i = d^{-l_i}$ for some integer l_i - i.e. $-\log_d p_i$ is an integer.

Definition 10.10. PROOF of definition 10.9
We write the difference between the expected length and the entropy as

$$\begin{aligned} L_C(X) - H_d(X) &= \sum_i p_i l_i + \sum_i p_i \log_d p_i = \\ &= \sum_i p_i \log_d \frac{1}{d^{-l_i}} + \sum_i p_i \log_d p_i = \\ &= \sum_i p_i \log_d \frac{p_i}{d^{-l_i}} \end{aligned}$$

Note that the final expression "looks like" a relative entropy. It holds that $0 \leq d^{-l_j} \leq 1$ for all j, but $\sum_j d^{-l_j}$ is not necessarily 1 . Hence, let us normalize it to a distribution r_i defined as

$$r_i = d^{-l_i}/c \quad \text{where} \quad c = \sum_j d^{-l_j}$$

We are doing this in conditional probability too.
Having $r_i = d^{-l_i}/c$ and $c = \sum_j d^{-1_j}$ we can continue by

$$\begin{aligned} L_C(X) - H_d(X) = \sum_i p_i \log_d \frac{p_i}{d^{-l_i}} &= \sum_i p_i \log_d \frac{p_i}{r_i \cdot c} \\ &= \sum_i p_i \log_d \frac{p_i}{r_i} - \log_d c \\ &= D(p\|r) + \log_d \frac{1}{c} \geq 0 \end{aligned}$$

by the nonnegativity of the relative entropy and the fact that $c \leq 1$ (Kraft inequality).

Definition 10.11. d**-adic** code
A probability distribution is called d-adic if each of the probabilities is equal to $\frac{1}{d^n}$ for some integer n.
Due to the previous theorem, the expected length is equal to the entropy if and only if the probability distribution of X is d -adic. The proof also suggests a method to find a code with optimal length in case the probability distribution is not d -adic.

11.

Definition 11.1. (Optimal-code properties)
For any distribution X with $P(X = x_i) = p_i$, there exists an optimal prefix code that satisfies the following properties:

1. If $p_j > p_k$ then $l_j \leq I_k$.
2. Two longest codewords have the same length.
3. Each longest codeword has a sibling (Siblings are codewords that differ only in their last letter).
4. The longest codewords correspond to the least likely symbols.

Definition 11.2. (**Generating Discrete Distribution Using Fair-Coin Tosses**)
The general formulation of the problem is that we have a sequence of fair coin tosses $Z_1, Z_2, \ldots$ and we want to generate a discrete random variable X with the probability distribution $\vec{p} = (p_1, p_2, \ldots, p_m)$. Let the random variable T denotes the *number of coin flips* used by the algorithm.
We can describe the algorithm mapping outcomes of $Z_1, Z_2, \ldots$ to outcomes of X by a **binary tree**. Leaves of the tree are marked by outcomes of X and the path from the root to a particular leaf represents the sequence of coin toss outcomes.

1. **full** - every node is either leaf, or it has two childs
2. There can be **more leaves labeled by the same outcome** of X (sum the probabilities).
3. the expected number of fair bits $E[T]$ requeired to generate X is **expected depth** of this tree.

Definition 11.3. Let Y denote the set of leaves of a full binary tree and Y a random variable with distribution on Y, where the probability of a leaf of the depth k is 2^{-k}. The **expected depth** of this tree is **equal to the entropy** of Y.
(Proof:)Let $k(y)$ denote the depth of y (leaf).

$$
\begin{aligned}
E(T) &= \sum_{y \in \mathbf{Y}} k(y) \cdot \frac{1}{2^{k(y)}} \\
H(Y) &= -\sum_{y \in \mathbf{Y}} \frac{1}{2^{k(y)}} \log \frac{1}{2^{k(y)}} \\
&= -\sum_{y \in \mathbf{Y}} \frac{1}{2^{k(y)}} \cdot (\log 1 - \log 2^{k(y)}) \\
&= -\sum_{y \in \mathbf{Y}} \frac{1}{2^{k(y)}} \cdot (0 - k(y)) \\
&= --\sum_{y \in \mathbf{Y}} \frac{1}{2^{k(y)}} \cdot k(y) \\
&= \sum_{y \in \mathbf{Y}} \frac{1}{2^{k(y)}} \cdot k(y) = E(T)
\end{aligned}
$$

Definition 11.4. For any algorithm generating X, the expected number of fair bits used is at least the entropy $H(X)$, i.e.

$$E(T) \geq H(X)$$

Proof. Any algorithm generating X from fair bits can be represented by a binary tree. Label all leaves by distinct symbols Y. The tree may be infinite. Consider the random variable Y defined on the leaves of the tree such that for any leaf of depth k the probability is $P(Y = y) = 2^{-k}$. By the previous lemma, we get $E(T) = H(Y)$. **The random variable X is a function of Y** and hence we have $H(X) \leq H(Y)$, i.e. some leaves may equal on outputs of X.

Definition 11.5. The expected number of fair bits $E(T)$ required by the optimal algorithm to generate a random variable X is bounded as

$$H(X) \leq E(T) < H(X) + 2$$

12.

Definition 12.1. Communication system
A **discrete channel** is a system $(\mathscr{X}, p(y \mid x), \mathscr{Y})$ consisting of an input alphabet $\mathscr{X}$, output alphabet $\mathscr{Y}$, and a probability transition matrix $p(y \mid x)$ specifying the probability we observe $y \in \mathscr{Y}$ when $x \in \mathscr{X}$ was sent.

Definition 12.2. properties A channel is said to be **without feedback** if the output distribution do not depend on past output symbols, i.e. $p\left(y_k \mid x^k, y^{k-1}\right) = p\left(y_k \mid x^k\right)$. A channel is said to be **memoryless** if the output distribution depends only on the current input and is conditionally independent of previous channel inputs and outputs.

Definition 12.3. An (M, n) code for the channel $(\mathscr{X}, p(y \mid x), \mathscr{Y})$ consists of the following:

1. An index set $\{1, 2, \ldots, M\}$. (number of outputs)
2. An encoding function $f : \{1, 2, \ldots, M\} \to \mathscr{X}^n$, yielding codewords $f(1), f(2), \ldots, f(M)$
3. A decoding function $g : \mathscr{Y}^n \to \{1, 2, \ldots, M\}$, which is a deterministic rule assigning a guess to each possible receiver vector.

(how many times we use the chanel)

Definition 12.4. **Probability of an error** for the code (M, n) and the channel $(\mathscr{X}, p(y \mid x), \mathscr{Y})$ provided the i-th index was sent is

$$\lambda_i = P(g(Y^n) \neq i \mid X^n = f(i)) = \sum_{y^n : g(y^n) \neq i} p(y^n \mid f(i))$$

Whenever we are sending i, encoding it by function f, to the sequence of X^n. Then decoding Y^n with function g and it is not i

Definition 12.5. The **maximal probability of an error** for an (M, n) code is defined as

$$\lambda_{\max} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i$$

Definition 12.6. The (arithmetic) **average probability of error** for an (M, n) code is defined as

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

Definition 12.7. The **rate** R of an (M, n) code is

$$R = \frac{\log_2 M}{n}$$

bits per transmission. Intuitively, the **rate** expresses the ratio between the number of message bits and the number of channel uses, i.e. it *expresses the number of non-redundant bits per transmission.*

Definition 12.8. The **channel capacity** (Intuitively, the noiseless throughput of the channel) of a *discrete, memoryless* channel is

$$C = \max_{p_X} I(X; Y) = \max_{p_X} [H(Y) - H(Y \mid X)]$$

where X is the input random variable, Y describes the output distribution and the maximum is taken over all possible input distributions p_X.