Lecture 8: Probability theory

Probability theory is the science of chance. It starts with **combinatorics** and leads to a theory of **stochastic processes**. Historically, probability theory initiated from gaming problems like put together in **Girolamo Cardano's** gamblers manual in the 16th century. A great moment of mathematics occurred, when **Blaise Pascal** and **Pierre Fermat** jointly laid a first foundation of mathematical probability theory.

It took mathematicians much longer to formalize "randomness" precisely. Here is the setup as which it had been put forward by **Andrey Kolmogorov**: all possible experiments of a situation are modeled by a set Ω, the "laboratory". A measurable subset of experiments is called an "event". Measurements are done by real-valued functions X. These functions are called **random variables** and are used to **observe the laboratory**.

As an example, lets model the process of throwing a coin 5 times. An experiment is a word like $htht$, where h stands for "head" and t represents "tail". The laboratory consists of all such 32 words. We could look for example at the event A that the first two coin tosses are tail. It is the set $A = \{ttttt, tttth, tttht, ttthh, tthtt, tthth, tthht, tthhh\}$. We could look at the random variable which assigns to a word the number of heads. For every experiment, we get a value, like for example, $X[tthht] = 2$.

In order to make statements about randomness, the concept of a **probability measure** is needed. This is a function P from the set of all events to the interval $[0, 1]$. It should have the property that $P[\Omega] = 1$ and $P[A_1 \cup A_2 \cup ...] = P[A_1] + P[A_2] + ...$, if A_i are disjoint events.

The most natural probability measure on a finite set Ω is $P[A] = \|A\|/\|\Omega\|$, where $\|A\|$ stands for the number of elements in A. It is the "number of good cases" divided by the "number of all cases". For example, to count the probability of the event A that we throw 3 heads during the 5 coin tosses,, we have $|A| = 10$ possibilities. Since the entire laboratory has $|\Omega| = 32$ possibilities, the probability of the event is 10/32. In order to study these probabilities, one needs **combinatorics**:

How many ways are there to:	The answer is:
rearrange or permute n elements	$n! = n(n-1)...2 \cdot 1$
choose k from n with repetitions	n^k
pick k from n if order matters	$\frac{n!}{(n-k)!}$
pick k from n with order irrelevant	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

The **expectation** of a random variable $E[X]$ is defined as the sum $m = \sum_{\omega \in \Omega} X(\omega) P[\{\omega\}]$. In our coin toss experiment, this is 5/2. The **variance** of X is the expectation of $(X - m)^2$. In our coin experiments, it is 5/4. Its square root is called the **standard deviation**. This is the expected deviation from the mean. An event happens **almost surely** if the event has probability 1.

An important case of a random variable is $X(\omega) = \omega$ on $\Omega = R$ equipped with probability $P[A] = \int_A \frac{1}{\sqrt{\pi}} e^{-x^2} \, dx$, the **standard normal distribution**. Analyzed first by **Abraham de Moivre** in 1733, it was studied by **Carl Friedrich Gauss** in 1807 and therefore also called **Gaussian distribution**.

Two random variables X, Y are called **decorrelated**, if $E[XY] = E[X] \cdot E[Y]$. If for any functions f, g also $f(X)$ and $g(Y)$ are decorrelated, then X, Y are called **independent**. Two random variables are said to have the same distribution, if for any $a < b$, the events $\{a \leq X \leq b\}$ and $\{a \leq Y \leq b\}$ are independent. If X, Y are decorrelated, then the relation $\text{Var}[X] + \text{Var}[Y] = \text{Var}[X + Y]$ holds which is just **Pythagoras theorem**, because decorrelated can be understood geometrically: $X - E[X]$ and $Y - E[Y]$ are orthogonal. A common problem is to study the sum of independent random variables X_n with identical distribution. One abbreviates this IID. Here are the three most important theorems which we formulate in the case, whee all random variables are assumed to have expectatation 0 and standard deviation 1. Let $S_n = X_1 + ... + X_n$ be the n'th sum of the IID random variables. It is also called a **random walk**.

LLN **Law of Large Numbers** assures that S_n/n converges to 0.

CLT **Central Limit Theorem**:$S_n/\sqrt{n}$ approaches the Gaussian distribution

LIL **Law of Iterated Logarithm:** $S_n/\sqrt{2n \log\log(n)}$ accumulates in $[-1, 1]$

The LLN shows that one can find out about the expectation by averaging experiments. The CLT explains why one sees the standard normal distribution so often. The LIL finally gives us a precise estimate how fast S_n grows. Things become interesting if the random variables are no more independent. Generalizing LLN,CLT,LIL to such situations is part of ongoing research.

Here are two open questions in probability theory:

Are $\pi, e, \sqrt{2}$... normal: do all digits appear with the same frequency?
What growth rates Λ_n can occur in S_n/Λ_n having limsup 1 and liminf -1?

For the second question, there are examples for $\Lambda_n = 1, \lambda_n = \log(n)$ and of course $\lambda_n = \sqrt{n \log\log(n)}$ from LIL if the random variables are independent. Examples of random variables which are not independent are $X_n = \cos(n\sqrt{2})$.

Statistics is the science of modeling random events in a probabilistic setup. Given data points, we want to find a **model** which fits the data best. This allows to **understand the past**, **predict the future** or **discover laws of nature**. The most common task is to find the **mean** and the **standard deviation** of some data. The mean is also called the **average** and given by $m = \frac{1}{n} \sum_{k=1}^n x_k$. The variance is $\sigma^2 = \frac{1}{n} \sum_{k=1}^n (x_k - m)^2$ with standard deviation σ.

A sequence of random variables X_n define a so called **stochastic process**. Continuous versions of such processes are where X_t is a curve of random random variables. An important example is **Brownian motion**, which is a model of a random particles.

Besides gambling and analyzing data, also **physics** was an important motor to develop probability theory. An example is statistical mechanics where laws of nature are studied with probabilistic methods. A famous physical law is **Ludwig Boltzmann's** relation $S = k \log(W)$ for entropy, a formula which decorates Boltzmann's tombstone. The **entropy** of a probability measure $P[\{k\}] = p_k$ on a finite set $\{1, ..., n\}$ is defined as $S = -\sum_{i=1}^n p_i \log(p_i)$. Today, we would reformulate Boltzmann's law and say that it is the expectation $S = E[\log(W)]$ of the logarithm of the "Wahrscheinlichkeit" random variable $W(i) = 1/p_i$ on $\Omega = \{1, ..., n\}$. Entropy is important because nature tries to maximize it