PROBABILITY THEORY

MATH 154

Unit 1: What is probability theory?

Introduction

1.1. Probability theory starts with a **probability space** $(\Omega, \mathcal{A}, \mathrm{P})$. It is given by a set Ω, a **σ-algebra** of subsets and a **probability measure** P. A **random variable** is a measurable map $X : \Omega \to \mathbb{R}$. It often has an **expectation** $\mathrm{E}[X]$ and a **variance** $\mathrm{Var}[X]$. The theory also studies **stochastic processes** like $S_n = X_1 + X_2 + \ldots X_n$. Every random variable X has a **cumulative distribution function** $F(x) = \mathrm{P}[X \leq x]$ and a **law** $f = F'$, which is a probability measure on the real line. Finite probability theory leads to **combinatorics**. The field is rooted on **measure theory**. It relates to **real analysis** and the **foundations of mathematics**. Geometry uses it in **integral geometry**. In **analysis**, it is helpful to study **partial differential equations**. In machine learning, **weights** define probability spaces on networks of tokens. Statistical decision theory uses probability spaces on states, actions and signals to tackle **decision problems** for **agents** trying to optimize **utility** under uncertainty. In physics, the language can model the **quantum nature** of the world.

1.2. Modern probability theory has a clean axiomatic foundation. It plays a role when probing the nature of **space and time**. Fundamental conundrums also appear in pure mathematics, like why the distribution of primes is so random or why why digits obtained by an expansion of fundamental constants like π pass statistical tests for being random. If we look at primes p_n of the form $p = x^2 + y^2$ with $0 < y < x$ for example, we see that the angles $X_n = \arctan(y_n/x_n)$ appear to be pretty random. The decimal digits X_n of π pass statistical tests that the digit data come from a random process. Already the **normality of** π, the statement that all decimal digits of π appear with the same frequency, is unproven. Probability theory is a play field also in **linear algebra**, where one can looks for example at properties of **random matrices** or in statistical physics, when looking at **random networks**. How are the eigenvalues of random matrices or Laplacians of random networks distributed? In complex analysis, the differences between successive roots of the Riemann zeta function using statistical tools.

1.3. Much of probability theory deals with **independent, identically distributed random variables**. This assumption, abbreviated IID, is very strong and allows to prove **theorems**. Given a sequence of X_k of such random variables, the **law of large numbers** assures that $\frac{1}{n}S_n$ converges to the expectation $\mathrm{E}[X]$. The average of a thousand random dice events is expected to be close to 3.5. The **law of large**

numbers assures that the normalized random variables $\overline{S_n}$ converge in law to the Gaussian distribution. The **law of iterated logarithms** refines this and assures that $\frac{1}{\sqrt{2n\log\log(n)}}S_n$ has its accumulation points in $[-1,1]$. Modern frontiers in the subject try to push these theorems to situations with less assumptions. The Birkhoff ergodic theorem for example generalizes the law of large numbers to an ergodic measure preserving transformation. An example is the irrational rotation $x \to x+\alpha = x+\sqrt{2} \bmod 2\pi$ and study $X_n(\theta) = \cos(\theta + n\alpha)$. These are identically distributed random variables but far from independent. Still, $\frac{1}{n}S_n \to \mathrm{E}[X]$ for $n \to \infty$.

1.4. Ergodic theory is a theory about measure preserving transformations of a probability space. It turns out that every sequence of random variables $X_1, X_2 \dots$ with identically distributed variables can be written as $X_k(x) = X(T^k x)$, where $T : \Omega \to \Omega$ is measure preserving, $T^k(x) = T(T^{k-1}(x))$ and X is a single random variable of that distribution. What distinguishes probability theory within the **theory of dynamical systems** is that in the former theory, random variables are often assumed to be independent. One can realize this by taking $\Omega = \prod_{k=1}^n \Delta$ to be a product probability spaces and to be $T(x)_k = x_{k+1}$ as the shift. Sometimes, the independence is not obvious. The map $z \to 4x(1-x)$ on $[0,1]$ preserves a measure that is smooth in $(0,1)$. One can pull out of this process IID random variables. For processes like the double pendulum, which is a measure preserving flow on 3-dimensional energy surfaces, we can not prove yet that there are random variables X such that $X_k = X(T^k x)$ are independent; despite the fact that physical experiments indicate this is true. Also for simpler systems like the measure-preserving map $T(x,y) = (2x - y + c\sin(x), y)$ on the 2-torus Ω, equipped with the area measure, there is numerical evidence that there should be random variables X on Ω such that $X_k = X(T^k)$ are IID. This would follow from $\frac{1}{n}\iint_\Omega \log(||dT^n(x,y)||)\, dxdy \geq \epsilon > 0$ for large enough n. It would imply that the Jacobean matrix dT^n grows exponentially with positive probability. This example illustrates that for matrix-valued random variables X_n, the growth rate of $S_n = X_1 X_2 \dots X_n$ is much tougher to understand.

1.5. More research is needed when studying stochastic processes which are correlated. In the real world, observations are hardly independent. Everything is connected, sometimes in a complicated way. There can be correlations between different things, sometimes unexpected. We also do not always can expect to have finite variance. Catastrophes like wildfires do not fit into probabilistic models. For example, look at the random variables $X_n(\theta) = \cot(\pi\theta + n\pi\alpha)$, where $\alpha = (1+\sqrt{5})/2$. This is very unusual. First of all, the random variables have Cauchy distribution which is an example of a distribution with infinite variance that models "high risk". Getting close to 0, the pole of cot, produces huge changes. For IID random variables there is also a central limit theorem. The Cauchy distribution plays the role of the Gaussian distribution now. But the random variables are also correlated. The covariance between successive variables $\mathrm{Cov}[X_0, X_1] = E[X_0 X_1] = \int_0^1 \cot(\pi x)\cot(\pi(x+\alpha))\, dx$ numerically evaluates to -1.12019 while $\mathrm{Cov}[X_0, X_2] = 1.60942$. The **cot**-process is interesting because the random walk S_n produces a growth rate can be computed explicitly and is self-similar.. We mention this example only to illustrate that we have hardly scratched the surface of the subject.