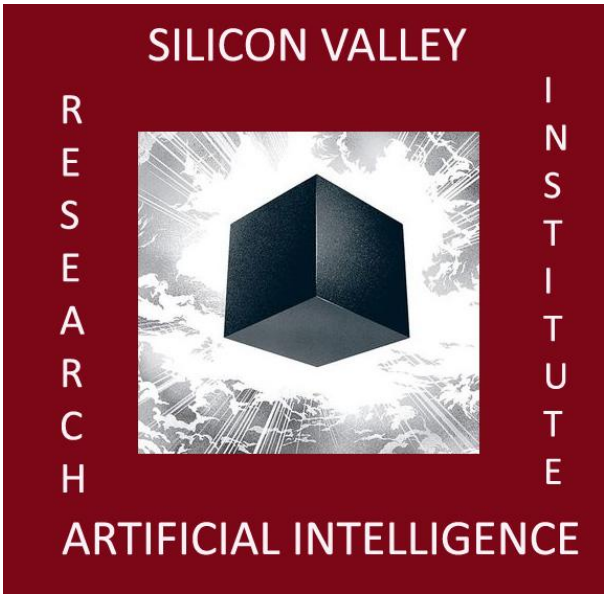

Probabilities, Markov Chains, Monte Carlo Methods: a brief introduction

In 1494 an Italian friar, Luca Pacioli published the first book ("Summa de Arithmetica, Geometria, Proportioni et Proportionalita") that mentions the "problem of points", considered by many the inspiration for the development of a mathematical theory of probability. (Pacioli, who later in Milan became a friend and housemate of Leonardo da Vinci, also published "De Viribus Quantitatis", the first book of recreational mathematics, including tricks of prestidigitation). In 1564 the Italian mathematician, inventor and physician Geronimo Cardano, son of another friend of Leonardo da Vinci in Milan, wrote a book ("Liber de Ludo Aleae") on games of chance, basically a gambling manual, that would be published only a century later. (Cardano published more than 200 books and invented dozens of mechanical devices, besides popularizing the magic trick of the magic coloring book, but, ironically, his son became an addicted gambler). The first person to give us a rigorous treatment of probability was the Dutch astronomer Christian Huygens, who in 1657 published "De Ratiociniis in Ludo Aleae". (Huygens discovered the rings of Saturn, invented the pendulum clock and built countless mechanical automata for the royal court). In 1689 the Swiss mathematician Jacob Bernoulli finished his treatise "Ars Conjectandi" (published posthumously), a collection of problems of probability that stands as the first textbook on the subject. (This Bernoulli, not to be confused with the Johann Bernoulli of infinitesimal calculus, with whom he founded the calculus of variations, and not to be confused with the Daniel Bernoulli of the Bernoulli principle in fluid dynamics, belonged to an incredible family of mathematicians). The French mathematician Abraham de Moivre, exiled in Britain, wrote an even more popular textbook, "The Doctrine of Chances" (1718), the first one not to be written in Latin. The British priest, theologian and amateur mathematician Thomas Bayes died in 1761 before he could publish his main achievement, a formula on how to calculate the probability of an event, a formula that today we know as Bayes' Theorem. (His "Essay towards Solving a Problem in the Doctrine of Chances" was read to the Royal Society 1763). The French mathematician Pierre Laplace had become a living legend applying Newtonian gravitation to the

solar system (Newton himself, unable to prove mathematically many details of the motion of the Sun, the planets and the moons, had conceded that God's intervention was periodically required to keep the planets from crashing). In 1796 Laplace had even discovered "black holes" (more than one century before Einstein's Relativity). In 1812 Laplace published his textbook "Analytical Theory of Probabilities" with the version of Bayes' theorem that we use today; and also called it what it is, i.e. "probabilities". (Two years later, in his "Philosophical Essay on Probabilities", Laplace articulated the principle of causal determinism: if one "at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed", that person would be able to predict the future and wouldn't need probabilities).

Markov chains extended the theory of probability in a new direction, to sequences of linked events.

Markov chains were first described in 1906 by the Russian mathematician Andrey Markov. The goal of statistical classifiers and neural networks is to identify something (classify it). The goal of Markov chains is different: they use probability theory for guessing what the next element in a sequence could be. Markov analyzed Pushkin's lengthy poem "Eugene Onegin" using this method, trying to guess the next letter from the previous one. Little did he know that his method would become one of the most popular "guessing" methods of the century, applied to everything from physics to economics and from genetics to sociology and even to gambling (one can interpret the state of the Markov chain as the fortune of a gambler). Markov chains can only be used when the process is a Markov process: the current state of the system is always dependent on the immediate previous state only. Andrey Kolmogorov (1936) expanded the mathematics in such a way that Markov chains could be used for any continuous process, not just for sequences of discrete events. In 1948 Claude Shannon, the father of information theory, used Markov chains to produce a sentence, simply based on the distribution of alphabetical letters in English words, and came up with the sentence "in no ist lat whey cratict foure birs grocid". The similarity with real English convinced him that communication systems are Markov processes. In 1953 Shannon also built a machine based on Markov chains that could play a game against humans, the "mind-reading machine".

The popularity of Markov chains grew rapidly. The very first piece of computer music (music generated by software), the "Illiac Suite" premiered in August 1956 on the Illiac computer at the University of Illinois, was generated by Markov chains programmed by Lejaren Hiller and Leonard Isaacson. The Greek composer Iannis Xenakis, who at the time was still an engineer in the studio of architect Le Corbusier in Paris, used Markov chains to compose the electroacoustic piece "Analogique A-B (1958-59). In 1964 Hiroshi Kawano, a Japanese pioneer of computer art, started using Markov chains in his paintings.

Possibly the most influential application of the Markov chain is the Google search engine (1998).

A Markov chain prescribes only one action for each state, and there is no reward for the result of that action. When a Markov chain is equipped with multiple actions and rewards, it is called "Markov decision process". Finding a solution for a Markov decision process is not easy.

In 1957 Richard Bellman of RAND Corporation introduced the "value iteration" method (in his book "Dynamic Programming"), and in 1960 Ronald Howard of MIT published the "policy iteration" method (in his book "Dynamic Programming and Markov Processes").

Markov chains are useful in an ideal world in which you can know the states of the system. Markov chains are "observed Markov models". In the real world it is rarely possible to know the state. Instead, we get indirect information. For example, the same word can be pronounced in many different ways by different speakers. In this case there is a whole distribution of sounds that corresponds to that state. In the case of two homophones like "write" and "right" there is one sound that correspondes to two states. In such cases we have a Markov process with unobserved (or "hidden") states, states that can only be observed indirectly. Instead of a regular Markov chain we need to use a "hidden Markov model". We can directly compute the probability of the next state in a Markov chain because we can observe the current state, but, in the case of "hidden" states, the procedure is more complicated and consists in steps that get repeated to achieve a better and better estimate. A particular hidden Markov model is definited by "transition probabilities" and "emission probabilities". The model generates two sequences: a path of states determined by transition probabilities, and the observed sequence due to the emission probability of each state in the path. The state path is a hidden Markov chain. An observed sequence (e.g. a sequence of vocal sounds) can be due to many different state paths (e.g. to many different words). The inference consists in determining ("decoding") the underlying cause of the observed sequence, i.e. the most likely sequence of hidden states that account for those observations; The decoding is usually done with an algorithm proposed by Andrew Viterbi 1967; i.e. the Viterbi algorithm finds the most probable path that generates the observed sequence. Another possible use of the hidden Markov model is to predict the next observation in the sequence.

The hidden Markov model is a Bayesian network that has the sense of time and can model a sequence of events. It was invented in 1966 by Leonard Baum, a cryptographer working at the Institute for Defense Analyses in Princeton.

The obscure algorithms called Markov Chain Monte Carlo (MCMC) are among the most important mathematical discoveries of the 20th century because they have had literally thousands of practical applications, especially in physics, solving problems that were considered impossible to solve in a reasonable time. In fact, a special issue of Computing in Science & Engineering (January 2000), edited by Francis Sullivan, included one of them, the "Metropolis algorithm", among the ten most important algorithms of the 20th century. (Sullivan wrote in his introduction that "great algorithms are the poetry of computation").

Starting in 1934 the Italian nuclear physicist Enrico Fermi developed techniques of "statistical sampling" to model the motion of neutrons. (In 1942, after moving to the USA, Fermi built the world's first nuclear reactor). Given the practical impossibility of calculating what happens to every neutron, Fermi found a way to generalize to the whole population the results obtained from a sample. Fermi didn't give a name to the techniques that he was using. In 1946 the Polish-born mathematician Stanislaw Ulam came up with a method to simulate and estimate (not calculate exactly) a nuclear explosion, and he nicknamed it "Monte Carlo" method because it involves a form of (mathematical) gambling. This was an example of intractable mathematical problem, "intractable" because the time required to find the exact solution grows exponential. In

these cases an approximate solution is better than no solution. John Von Neumann, who was involved in the design of the ENIAC, understood that this method was perfectly suited for the electronic computer. In 1948 his wife Klara Von Neumann and Nicholas Metropolis of the Los Alamos national laboratory programmed the ENIAC to run the Monte Carlo method, first on fusion and then on fission problems. Ulam and Metropolis published the first paper on the Monte Carlo method in 1949 (titled "The Monte Carlo Method"). Over the years a variety of Monte Carlo algorithms have been developed.

At the time the Los Alamos National Laboratory was working on the atomic bomb and its numerous physicists, including Von Neumann and Metropolis, were developing mathematical methods to deal with nuclear experiments, and using the first electronic computers to perform the calculations. Nicholas Metropolis was trying to compute the equilibrium state of a collection of atoms at a certain temperature, a problem that requires to calculate very complicated integrals. These integrals cannot be approximated even with the Monte Carlo method. Metropolis led the development of a new computer, nicknamed MANIAC (Mathematical Analyzer, Numerical Integrator and Computer), his team member Marshall Rosenbluth invented a new algorithm and his other team member Arianna Rosenbluth implemented it on the Maniac ("Equation of State Calculations by Fast Computing Machines", 1953). This algorithm, now erroneously known as the "Metropolis algorithm", was the first Markov Chain Monte Carlo (MCMC) method, a method combining Monte Carlo method and Markov chain: an MCMC algorithm builds a convergent Markov chain whose limit is the desired probability distribution. (For the record: Edward Teller posed the mathematical problem, Marshall Rosenbluth solved it, Arianna Rosenbluth implemented it, Metropolis was their boss and the fifth co-author did not do anything).

Meanwhile, mathematicians and philosophers had been arguing over the meaning of probabilities. You can view a probability as on objective number (the number of times that you get heads and tails when you flip a coin) or as a subjective number (my belief that Brazil will win the next world cup). The British economist John Maynard Keynes at Cambridge University published "A Treatise on Probability" (1921) in which he argued against the subjective approach. A little later two independent minds came up with the opposite view. Frank Ramsey, a mathematical prodigy at Cambridge University defended subjective probability in his essay "Truth and Probability" (1926) just before dying at the very young age of 26. The Italian statistician Bruno DeFinetti studied subjective probability aiming to use it for predictive inference, and explained his approach in a talk given in 1928 in Bologna at the International Congress of Mathematics. His famous motto was: "Probability does not exist", meaning that there is no objective probability, just subjective ones. John VonNeumann and Oskar Morgenstern mentioned in passing the need for a theory of subjective probability in their book "Theory of Games and Economic Behavior" (1944). The British mathematician Jimmie Savage (born Leonard Savage), who during World War II had worked with Von Neumann at the Institute for Advanced Study in Princeton, carried out that task at the University of Chicago and published the seminal book "Foundations of Statistics" (1954).

Others were working on the both the mathematical and philosophical foundations of probability. The Russian mathematician Andrey Kolmogorov in his book "Foundations of Probability Theory" (1933) did for probability what Frege and Peano had done for arithmetic: he

gave it logical foundations, based on just three axioms. Rudolf Carnap, a student of Gottlob Frege in his native Germany and a teacher of Walter Pitts at the University of Chicago, studied the philosophical foundations of probability and induction in "Logical Foundations of Probability" (1950), and was probably the first to explore the relation between probability and first-order predicate logic.

Bayes' theorem basically calculates the probability of belonging to a class given some facts. For example, what is the probability that you are a lawyer if you live in the USA? For example, if you test positive for a disease, what are the odds that you actually have that disease? Bayes' theorem says that the "conditional" probability of A occurring given B equals the probability that B occurs given A (a simple statistical fact) multiplied by the probability of A and divided by the probability of B (also simple statistical facts). The theorem is also used to calculate "posterior" probability from a "prior" probability: the conditional probability that event A occurs given that even B has occurred.

Bayesian models can represent complex problems by joining together a series of these theorems. Bayesian models have "latent" or "random" variables: variables that cannot be observed, such as a column of empty data. Layers of latent variables create a hierarchy of concepts.

This kind of hierarchical Bayesian reasoning was defended, at a time when Bayes' theorem was rather unpopular, in a paper titled "Rational Decisions" (1952) by the British mathematician Jack Good (born John Irving Good), who during World War II had worked at Bletchley Park with Alan Turing and who would later start singularity thinking with the essay "Speculations Concerning the First Ultraintelligent Machine" (1964).