Elementary Probability Theory

Yaokun Wu
Department of Mathematics
Shanghai Jiao Tong University
Shanghai, 200240, China
ykwu@sjtu.edu.cn

Course slides for ACM class, Fall 2006
http://www.math.sjtu.edu.cn/teacher/wuyk/acm06.pdf

It is not our job to make our computers easier for you to use. It is our job to teach the knowledge, skills and attitudes necessary to enable you to make computers easier to use by others. –inspired by John F. Kennedy

... And we hope some of the best of you will take on our job and do the same for your students.

Personal Home Page of S. Chaiken, `http://www.cs.albany.edu/~sdc/`

Classroom

E-323

Schedule

Every week: Tuesday 8:00 – 9:40

Every other week: Friday 10:00 – 11:40

HAVE A GOOD START IN THE NEW SEMESTER!

This course is mainly based on:

Emmanuel Lesigne, Heads or Tails: An Introduction to Limit Theorems in Probability, American Mathematical Society, 2005.

For those who wish to read a textbook in Chinese for this course, we suggest:

Jian-Gang Ying, Ping He, GAI LV LUN, Fudan University Press, 2005.

We recommend the following for further reading:

Sariel Har-Peled, Class notes for Randomized Algorithms, `http://valis.cs.uiuc.edu/~sariel/teach/notes/rand_alg/notes.pdf`, 2005.

Many actions have outcomes which are largely unpredictable in advance. To describe them, we often give the following basic information:*

- The set of all possible outcomes of the experiment.
- A list of all the events which may occur as consequences of the experiment.
- An assessment of the likelihoods of these events.

The word 'probability' is often used for likelihood. Recall the situation when tossing a coin, playing the lottery, making a weather forecast, giving birth to a child, taking an exam, ...

*G. Grimmett, D. Welsh, Probability: An Introduction, Clarendon, 1986.

Interpretations of Probability: 1. Long term frequency; 2. Measure of belief; 3. Information (Surprise) – The amount of information we get from the happening of an event with probability p is $-\log p$.*

All of us use the words probable and certain, and generally speaking, we all seem to agree as to what they mean. Nevertheless it will not be entirely useless to try to get as precise an idea of this meaning as possible. Émile Borel, Probability and Certainty, (translated by Douglas Scott from its 1950 French edition) Walker and Company, 1963.

*The less likely the event is to happen the more the information, the more the surprise, the greater the reduction in uncertainty. – Richard Hamming, The Art of Probability, Addison-Wesley, 1991.

Our first intuition about probability may be fraught with fallacies.

Example 1 (St Petersburg paradox) * *This was first stated and addressed in eighteenth century by a French expert on probability in his correspondence exchanged between Paris and St. Petersburg. Consider the following game. A fair coin is tossed repeatedly until the first time that it comes up tails. Let* X *be the (random) number of heads that come up before the first occurrence of tails. Suppose that the bank pays* 2^X *roubles depending on* X*. How much would you be willing to pay to enter this game? Do you agree to pay* $E(2^X)$ *roubles? Is* $(E(2^{\frac{X}{2}}))^2$*, as suggested by Bernoulli, more reasonable? There is a good discussion of this paradox in É. Borel, Probability and Certainty.*

*Olle Häggström, Finite Markov chains and algorithmic applications, Example 1.1, Cambridge University Press, 2002. Utility Theory is developed to explain this paradox.

Example 2 (Betrand's paradox) * *Throw at random straight lines onto the unit disc. What is the probability that the line intersects the disc with a length* $\geq \sqrt{3}$*, the length of the equilateral triangle inscribed in the circle?*

Example 3 (Monty Hall Problem) † *Suppose you're on a game show and you are given a choice of three doors. Behind one door is a car and behind the others are goats. You pick a door – say No. 1 – and the host, who knows what's behind the doors, opens another door – say, No. 3 – which has a goat. (In all games, he opens a door to reveal a goat.) He then says to you, "Do you want to pick door No. 2?" (In all games, he always offers an option to switch.) Is it to your advantage to switch your choice?*

*Oliver Knill, Probability and Stochastic Processes with Applications, `http://www.math.harvard.edu/~knill/teaching/math144_1994/probability.pdf`

†Oliver Knill, Probability and Stochastic Processes with Applications.

Example 4 *Put three balls randomly into two boxes. What is the probability that one box is empty?*

Bose-Einstein statistics: $\frac{1}{2}$

Maxwell-Boltzmann Statistics: $\frac{1}{4}$

Bose-Einstein statistics describe physical systems in which the particle number is nonconservative, particles are indistinguishable, and cells of phase space are statistically independent. Such systems allow two polarization states, and exhibit totally symmetric wavefunctions. In 1940, Pauli showed that Bose-Einstein statistics follows from quantum field theory. —`http://scienceworld.wolfram.com/physics/`

Maxwell-Boltzmann Statistics is the quantum statistics of particles that are distinguishable and any number of particles may exist in a given state. — `http://scienceworld.wolfram.com/physics/`

The abstract probability theory, consisting of axioms, definitions, and theorems, must be supplemented by an interpretation of the term "probability". This provides the correspondence rule by means of which the abstract theory can be applied to practical problems. There are many different interpretations of probability because anything that satisfies the axioms may be regarded as a kind of probability. – L.E. Ballentine, Probability Theory in Quantum Mechanics, in: The Concept of Probability: Fundamental Theories of Physics, (Eds., E.I. Bitsakis, C.A. Nicolaides) Kluwer Academic Publishers, 1989.

In each field we must carefully distinguish three aspects of the theory: (a) the formal logical content, (b) the intuitive background, (c) the applications. The character, and the charm, of the whole structure cannot be appreciated without considering all three aspects in their proper relation. – W. Feller

If the theory of probability is true to life, this experience must correspond to a provable statement. – W. Feller

Randomness is a mathematical concept, not a physical one. – Richard W. Hamming, The Art of Probability for Scientists and Engineers, Addison-Welsey, 1991.

The fundamental question is whether randomness really exists, or whether we use this term only to model objects and events with unknown lawfullness. Philosophers and scientists have disputed the answer to this question since ancient times. – Juraj Hromkovič, Design and Analysis of Randomized Algorithms, Springer, 2005.

Randomness and order do not contradict each other; more or less both may be true at once. The randomness controls the world and due to this in the world there are order and law, which can be expressed in measures of random events that follow the laws of probability theory. – Alfréd Rényi

The probabilistic method has recently been developed intensively and became one of the most powerful and widely used tools applied in combinatorics. One of the major reasons for this rapid development is the important role of randomness in theoretical computer science, a field that is recently the source of many intriguing combinatorial problems. – N. Alon, J.H. Spencer, The Probabilistic Method, Wiley, 1992.

We now present the mathematical model that describes those probabilistic experiment.*

A probability space is a triple (Ω, Σ, P), where Ω is a set, $\Sigma \subseteq 2^{\Omega}$ is a σ-algebra on Ω, namely a collection of subsets containing Ω and closed on complements, countable unions and countable intersections, and P is a countably additive measure on Σ with $P(\Omega) = 1$. Ω is called the sample space. The elements of Σ are called events and the inclusion minimal elements of $\Sigma \setminus \{\emptyset\}$ are called elementary events. For an event A, $P(A)$ is the probability of A.

Probability theory is the theory on probability spaces.

*We refer to R.W. Hamming, Models of Probability, Chapter 8 in The Art of Probability, Addison-Wesley, for many actual models of probability.

We call (Ω, Σ, P) a finite probability space if Ω is finite and call it a discrete probability space if Ω is finite or countably infinite. When $\Sigma = 2^{\Omega}$, we simply write (Ω, P) for (Ω, Σ, P). We will start with a study of finite (discrete) probability space.

Let $\mathcal{E}$ be the set of elementary events of a probability space (Ω, Σ, P). There is a natural correspondence between Σ and $2^{\mathcal{E}}$ and so, in some sense, we can view (Ω, Σ, P) the same with $(\mathcal{E}, P')$ for some obviously defined P'. This is the reason that in almost all textbooks a discrete probability space is defined directly to be a space (Ω, P).

For a discrete probability space (Ω, P), to give P we need only specify the probability of elementary events.

Example 5 *Let* $\Omega = \{0, 1\}$, $P(0) = 1 - p, P(1) = p$. *We say that the space* $\{0, 1\}$ *is equipped with the probability* $(1 - p, p)$. *You can think of the flip of a coin where the outcome heads, recorded as* 1, *has probability* p *to appear, and the outcome tails, recorded as* 0, *has probability* $1 - p$ *to appear.*

Example 6 (The space of random graphs) *The probability space of random graphs* $G(n, p)$ *is a finite probability space whose elementary events are all graphs on a fixed set of* n *vertices and where the probability of a graph with* m *edges is* $p^m(1 - p)^{\binom{n}{2} - m}$.

Example 7 *Let* $(\Omega^1, P^1), \ldots, (\Omega^n, P^n)$ *be a family of discrete probability space. Their product space is the discrete probability space* $(\prod_{i=1}^n \Omega^i, P)$ *where* P *is the probability satisfying* $P(\omega^1, \ldots, \omega^n) = \prod_{i=1}^n P^i(\omega^i)$ *for each elementary event* $\omega^i \in \Omega^i$.

We make the convention that $\Omega_i = \{0,1\}^i$. Let $(\Omega_1, (1-p, p))$ denote the space equipped with probability $(1-p, p)$. Let $(\Omega_n, (1-p, p)^{\bigotimes n})$ be the product space of n $(\Omega_1, (1-p, p))$ and say that this space is equipped with product probability $(1-p, p)^{\bigotimes n}$. $(\Omega_n, (1-p, p)^{\bigotimes n})$ is the mathematical model of flipping n times a coin whose probability of landing heads is p. Also observe that $G(n, p)$ is in essence no different from $(\Omega_{\binom{n}{2}}, (1-p, p)^{\bigotimes \binom{n}{2}})$.

The game of Heads or Tails, which seems so simple, is characterized by great generality and leads, when studied in detail, to the most sophisticated mathematics. – Émile Borel (1871–1956)

How does probability theory work to help us?

Given structure --→ Probability Space --→ Conclusion on the probability space --→ Conclusion on the given structure

The basic facts we learn from this course is something useful for establishing the second step. We also give some examples of how the first and the last steps are carried out. When dealing with a practical problem, it is your turn to invent the first and the last steps, possibly replacing the central part by 'Topological space --→ Conclusion on the topological space', or 'Linear space --→ Conclusion on the linear space', or some others.

Do not hurry to assert that Probability Theory (or Topology, Linear Algebra, ...) is useless. Do not just prove ourselves useless in failing to make probability theory useful.

One similarity is this: you have to solve a problem and you have certain tools that you are able to use and others that you are not allowed to use. And as in problem-solving there is the notion of elegance. The difference is that mathematicians have hundreds of years of tools whereas in magic you use whatever you can get. The similarity is especially so in applied mathematics in which the problem comes from somebody else. The chemist or biologist might have a question for you, and you don't have any ready tools. You have to start thinking about it and start using whatever tools you have or invent new ones. That's pretty similar to solving magic problems. – Persi Diaconis *

*Why not take a look at the whole interview of Persi Diaconis at `www.ims.nus.edu.sg/imprints/interviews/PersiDiaconis.pdf`

Persi W. Diaconis (born January 31, 1945) is an American mathematician and former professional magician. He is Mary V. Sunseri professor of statistics and professor of mathematics at Stanford University. He is particularly known for tackling mathematical problems involving randomness and randomization, such as coin flipping and shuffling playing cards. `http://www.answers.com/topic/persi-diaconis`

The last decade has witnessed a tremendous growth in the area of randomized algorithms. During this period, randomized algorithms went from being a tool in computational number theory to finding widespread application in many type of algorithms. Two benefits of randomization have spearheaded this growth: simplicity and speed. For many applications, a randomized algorithm is the simplest algorithm available, or the fastest, or both. – Rajeev Motwani, Prabhakar Raghavan, Randomized Algorithms, Cambridge University Press, 1995.

Probability theory is a central topic in mathematics. There are close relations and intersections with other fields like computer science, ergodic theory and dynamical systems, cryptology, game theory, analysis, partial differential equation, mathematical physics, economical sciences or statical physics. – Oliver Knill

THE ELEMENT OF CHANCE enters into many of our attempts to understand the world we live in. A mathematical theory of probability allows us to calculate the likelihood of complex events if we assume that the events are governed by appropriate axioms. This theory has significant applications in all branches of science, and it has strong connections with the techniques we have studied in previous chapter. – R.L. Graham, D.E. Knuth, O. Patashnik, Concrete Mathematics: A Foundation for Computer Sciences, China Machine Press, China Machine Press, 2002.

The great discovery of Erdös was that we can use probabilistic methods to demonstrate the existence of the desired graphs without actually constructing them. This phenomenon is not confined to graph theory and combinatorics: probabilistic methods have been used with great success in the geometry of Banach spaces, in Fourier analysis, in number theory, in computer science – especially in the theory of algorithms – and in many other areas. However, there is no area where probabilistic methods are more natural and lead to more striking results than in combinatorics. – Béla Bollobás, Modern Graph Theory, Beijing World Publishing Corporation, 2003.

The "probabilistic method" is a legacy of Paul Erdös that continues to grow and flourish and have powerful applications in all parts of the mathematical sciences. We will explore current results in discrete mathematics that use probabilistic existence arguments and require the use of sophisticated probability concepts. We will further explore the connection to modern issues in computer science by examining the possible implementations of probabilistic existence arguments by randomized or deterministic algorithms. –From Erdös to Algorithms (Applications of the "Probabilistic Method") A joint DIMACS - DIMATIA workshop, –`http://dimacs.rutgers.edu/Workshops/Erdos/announcement.html`

Advanced Course on Analytic and Probabilistic Techniques in Combinatorics, Dates: January 15 to 26, 2007 Place: Centre de Recerca Matemtica, Bellaterra.

The CRM offers a limited number of grants covering the registration fee and/or accommodation addressed to young researchers. The deadline for application is October 30, 2006.

`http://www.crm.es/Conferences/0607/ACProbabilisticTechniques/probabil`
`htm`

Assume $\Omega = \{\omega^1, \ldots, \omega^n\}$ and $P(\omega^i) = p_i$.*

Let $\mathbf{I}_A$ be the indicator variable of A; that is, $\mathbf{I}_A$ is the function mapping Ω to $\{0, 1\}$ that takes the value 1 on A and the value 0 on its complement A^c.† Thus

$$P(A) = \sum_{i=1}^{n} P(\omega^i)\mathbf{I}_A(\omega^i) = \sum_{i=1}^{n} p_i \mathbf{I}_A(\omega^i). \tag{1}$$

Exercise 8 *(i)* $\mathbf{I}_\Omega = 1, \mathbf{I}_\emptyset = 0$; *(ii)* $\mathbf{I}_{A\cap B} = \mathbf{I}_A\mathbf{I}_B$; *(iii)* $\mathbf{I}_{A\cup B} = \mathbf{I}_A + \mathbf{I}(B) - \mathbf{I}_A\mathbf{I}_B$. *(iv)* $\mathbf{I}_{A^c} = 1 - \mathbf{I}_A$. *(v)* $\mathbf{I}_{A\triangle B} = (\mathbf{I}_A - \mathbf{I}_B)^2 = \mathbf{I}_A + \mathbf{I}_B (\mod 2)$. *(vi)* $\mathbf{I}_{A\setminus B} = \mathbf{I}_A(1 - \mathbf{I}_B)$.

*n may be ∞.

†Note that two events are equal if and only if their indicator variables are the same. The support of $\mathbf{I}_A$ is just A.

Exercise 9 *Let A_1, A_2, A_3 be events. Prove that*

$$P(A_1 \cup A_2 \cup A_3) = \sum_i P(A_i) - \sum_{i \neq j} P(A_i \cap A_j) + P(\cap_i A_i). \quad (2)$$

A generalization of Eq. (2) for n events is called the Ramsey formula (also called the Inclusion-Exclusion Principle.). Please guess the form of the Ramsey formula and prove it. *

Exercise 10 †*Let A, B be two events. Prove that*

$$P(A \cup B) \leq P(A) + P(B). \quad (3)$$

Eq. (3) is known as Boole's inequality. State and prove a generalization of Boole's inequality for n events.

*$1 - \mathbf{I}_{\cup A_i} = \prod (1 - \mathbf{I}_{A_i})$.

†You can try to investigate what is the Kraft's inequality on instantaneous code and use this exercise to prove it.

Exercise 11 *Let A, B be two events. Prove that*

$$P(A \cap B) \geq P(A) + P(B) - 1. \tag{4}$$

Eq. (4) *is an instance of Bonferroni's inequality. State and prove the general Bonferroni's inequality for n events.*

Exercise 12 *Let A, B be two events. Prove that*

$$P(A \cup B) \geq 1 - P(A^c) - P(B^c). \tag{5}$$

State and prove a generalization of Eq. (5) *for n events.*

Exercise 13 *For any two events A, B, we define the distance $d(A, B)$ between A and B by $d(A, B) = P(A \triangle B)$. Prove that for any events A, B, C it holds $d(A, C) \leq d(A, B) + d(B, C)$.*

Exercise 14 *For any two events A, B, we define the distance $d(A, B)$ between A and B by $d(A, B) = \frac{P(A \triangle B)}{P(A \cup B)}$ if $P(A \cup B) \neq 0$ and $d(A, B) = 0$ otherwise. Prove that for any events A, B, C it holds $d(A, C) \leq d(A, B) + d(B, C)$.*

Events A and B are independent if $P(A \cap B) = P(A)P(B)$. If $P(B) = 0$ or 1, then B and any other event are independent.

For events A and B with $P(B) > 0$, the conditional probability of A, given that B occurs, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Except in the trivial case that $P(B) = 0$, A and B are independent just means that $P(A) = P(A|B)$.

Exercise 15 *Consider the space $G(n,p)$ of random graphs on n vertices (Example 6). Let ab and cd be different pairs of vertices. Show that the event consisting of all graphs containing the edge ab and the event consisting of all graphs containing the edge cd are independent events in $G(n,p)$.*

Exercise 16 *Suppose that $\Omega = \cup_{i=1}^{n} Y_i$ and $Y_i \cap Y_j = \emptyset$ for $i \neq j$. Show that $P(X) = \sum_{i=1}^{n} P(Y_i)P(X|Y_i)^*$. Hence prove Bayes' Theorem:*

$$P(Y_k|X) = \frac{P(Y_k)P(X|Y_k)}{\sum_{i=1}^{n} P(Y_i)P(X|Y_i)}.$$

*It is called the Partition Theorem, or The Theorem of Total Probability.

Exercise 17 * *You are travelling in a train with your sister. Neither of you has a valid ticket and the inspector has caught you both. He is authorized to administer a special punishment for this offence. He holds a box containing nine apparently identical chocolates, but three of these are contaminated with a deadly poison. He makes each of you, in turn, choose and immediately eat a single chocolate. (a) If you choose before your sister, what is the probability you survive? (b) If you choose first and survive, what is the probability that your sister survives? (c) If you choose first and die, what is the probability your sister survives? (d) Is it in your best interest to persuade your sister to choose first? (e) If you choose first, what is the probability that you survive, given that your sister survives?*

End of Lesson One 12/9/06

*Grimmett and Welsh, Probability: An Introduction, Clarendon Press, 1986.

The Ramsey number $R(k, \ell)$ is defined to be $\min\{n :$ any graph on n vertices contains a clique of size k or an independent set of size $\ell\}$.

Theorem 18 *For any $k \geq 2$, $R(k, k) > 2^{k/2-1}$.*

Proof. Consider a random graph G from $G(n, 1/2)$ (Example 6). For any fixed set of k vertices, the probability that they form a clique is $p = 2^{-\binom{k}{2}}$. The same goes for the occurrence of an independent set, and there are $\binom{n}{k}$ k-tuples of vertices where a clique or an independent set might appear. Recall that the probability of a union of events is at most the sum of their respective probabilities (Exercise 10). It then follows that $P(G$ contains a clique or an independent set of size $k) \leq 2\binom{n}{k}2^{-\binom{k}{2}}$.

It remains to choose n so that $2\binom{n}{k}2^{-\binom{k}{2}} < 1$. Using the simplest estimate $\binom{n}{k} \leq n^k$, we find that it is sufficient to have $2n^k \leq 2^{k/2}n^k < 2^{k(k-1)/2}$. This certainly holds whenever $n \leq 2^{k/2-1}$. Therefore, there are graphs on $\lfloor 2^{k/2} - 1 \rfloor$ vertices that contain neither a clique of size k nor an independent set of size k. This implies $R(k,k) > 2^{k/2-1}$. ∎

Question 19 *a) Improve the bound given in Theorem 18; b) Show explicitly that* $\sqrt[k]{R(k,k)} > 1 + \epsilon$ *for some constant* $\epsilon > 0$.

Complete disorder is impossible. – T.S. Motzkin

Let (Ω, Σ, P) be a discrete probability space. A function X defined from Ω to R is called a random variable provided $X^{-1}(r) \in \Sigma$ for each $r \in R$. We use the shorthand $(X \in F)$ for the event $\{\omega \in \Omega : \ X(\omega) \in F\}$. The probability distribution of the random variable X is given by the probabilities of the events corresponding to the values of X. If the random variable X takes values in a set $A \subseteq R$, then the events $(X = x)$ for $x \in A$ form a partition of Ω, called the partition associated with X, and the distribution of X is given by the pairs $(x, P(X = x))$ for x ranging in A.

The linear space spanned by all indicator variables for events just consists of all random variables.

Note that an asserted random variable makes sense only if the corresponding probability space has been specified. But, if the probability space can be understood from the context, we often do not describe it explicitly. When you have confusion with some statement on random variables, it is useful to first go back to check what is the correct probability space there.

Before Chebyshev the main interest in probability theory had been in the calculation of the probabilities of random events. He, however, was the first to realize clearly and exploit the full strength of the concept of random variables and their mathematical expectations. – A.N. Shiryayev, Probability, Springer, 1984.

The subject matter of probability theory is the mathematical analysis of random events, i.e. of those empirical phenomena which – under certain circumstances – can be described by saying that: They do not have deterministic regularity (observations of them do not yield the same outcome) whereas at the same time; They possess some statistical regularity (indicated by the statistical stability of their frequency). – A.N. Shiryayev, Probability, Springer, 1984.

The 'average value' of a random variable records its typical behavior and is an important parameter for its statistical regularity.

The expected value (expectation) $E(X)$ of a random variable X is given by the formula $E(X) = \sum_x xP(X = x)$. For finite probability space, every random variable has an expected value.

Exercise 20 *For finite probability space, the expected value function is the unique linear functional over the space of random variables satisfying $E(\mathbf{I}_A) = P(A)$ for each elementary event A. For any discrete probability space, the set of random variables whose expected value exist form a linear subspace and the expected value function is the unique linear functional over it satisfying $E(\mathbf{I}_A) = P(A)$ for each elementary event A.*

Exercise 21 *For any random variable X and function f, $E(f(X)) = \sum_x f(x)P(X = x)$.*

Exercise 22 *Verify the follow statements.*

1) If X *is a constant function, then* $E(X) = X$*; in particular,* $E(E(X)) = E(X)$ *for any random variable* X*.*

2) If $X \geq 0$*, then* $E(X) \geq 0$*; in particular,* $|E(X)| \leq E(|X|)$*.*

3) Cauchy-Schwarz inequality: $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$*. (hint: Use the linearity of expectation (Ex. 20) to show that* $\begin{pmatrix} E(X^2) & E(XY) \\ E(XY) & E(Y^2) \end{pmatrix}$ *is positive semidefinite and hence has a nonnegative determinant.)*

Exercise 23 (Law of iterated expectation) $E(X) = E(E(X|Y))$.*

*The density of a rectangle=the average of the density of its horizontal lines. This can be viewed as a discrete version of the Fubini's Theorem.

The linearity of expectation is a very useful tool. This guarantees that when computing the expectation of a complex random variable, we can try to decompose it into a sum of many simple random variables whose expectations are relatively easy to determine.

Here is a typical way of decomposing a random variable. Suppose we have a probability space (Ω, Σ, P) and a random variable X such that $X(A) = |A|$ for any $A \in \Sigma$. Then $X(A) = \sum_{\omega \in \Omega} \mathbf{I}_A(\omega)$ and hence $X = \sum_{\omega \in \Omega} X_\omega$, where X_ω is the random variable which counts the number of occurrence of ω in a experiment.

Example 24 *There are k people in an elevator, each wanting off at a random floor of one of the n upper floors. By the linearity of expectation, it is easy to find that the expected number of elevator stops is $n(1 - (1 - 1/n)^k)$.*

Example 25 *Suppose we unwrap a fresh deck of n cards and shuffle it until the cards are completely random. How many cards do we expect to be in the same position as they were at the start? The probability space here is all permutations f of n cards each with probability $\frac{1}{n!}$. Let X be the random variable denoting the number of fixed points of the random permutation f. Let X_i be the random variable which equals 1 if the ith card is fixed by f and 0 otherwise. Clearly, $E(X_i) = \frac{1}{n}$ for all i. We have a decomposition $X = \sum_{i=1}^{n} X_i$ and so $E(X) = \sum_{i=1}^{n} E(X_i) = 1$.* *

To use the linearity of expectation is just to do double counting. Double counting is the most important tool in mathematics.

*Many beautiful results on card shuffling are established with tools from group representations. See, Persi Diaconis, Group Representations in Probability and Statistics, Institute of Mathematical Statistics, 1988.

Example 26 *Every set* $B = \{b_1, \ldots, b_n\}$ *of* n *nonzero integers contains a sum-free subset* A *of size* $|A| > \frac{n}{3}$.

Solution: (A double counting proof) Let $p = 3k + 2$ be a prime which is greater than $2 \max_{1 \leq i \leq n} |b_i|^*$ and take $C = \{k + 1, k + 2, \ldots, 2k + 1\}$. Observe that C is a sum-free subset of the cyclic group Z_p and that $\frac{|C|}{p-1} = \frac{k+1}{3k+1} > \frac{1}{3}$. Look at the $n \times (p-1)$ matrix M whose (i, j) entry is $jb_i \in Z_p \setminus \{0\}$. We count the number of positions which are occupied by elements from C in the matrix M. Since every row of M has more than $\frac{1}{3}$ elements coming from C, we conclude that there is at least a column, say column j, more than $\frac{1}{3}$ of whose positions are occupied by elements from C. It is trivial to see that $\{b_i : \ jb_i \in C\}$ is the required sum-free subset of A. ∎

*This means that B corresponds to n different nonzero elements in Z_p.

Exercise 27 *Rewrite (translate) the above double counting solution of Example 26 into a proof by probabilistic method.*

Exercise 28 *Suppose we have a bag with n red marbles and n blue ones, and we mate these by picking out couples at random. What is the expected number of red-blue couples?*

Indeed, probabilities are merely ratios of sizes of sets, and so probabilistic analysis is merely combinatorics. Yet, as is often the case in mathematics and science, using certain definitions and notations (rather than others) may simplify the analysis tremendously. Specifically, in many cases, the analysis is much easier to carry out in terms of probabilities than in terms of sizes. – Oded Goldreich

End of Lesson Two 15/9/06

By a drawing of a graph G on the plane, we mean a representation of G in the plane such that each vertex is represented by a distinct point and each edge by a continuous arc connecting the corresponding two points and no three arcs cross at the same point. For any graph G, the minimum number of crossings of its drawings on the plane is called its crossing number and denoted $cr(G)$.

Theorem 29 (Crossing Lemma) * *For every (simple) graph G with $e(G) \geq 4n(G)$, we have $cr(G) > \frac{e(G)^3}{64n(G)^2}$.*

*Ajtai, Chvátal, Newborn, Szemerédi '82, Leighton '83

Proof. From Euler's formula, we know that the number of edges of a planar graph on n vertices is at most $3n-6$. * Thus, it holds for any graph G that $cr(G) \geq e - 3n + 6 > e - 3n$.

Fix a drawing of G on the plane with $cr(G)$ crossings. Pick each $v \in V(G)$ with probability p. Let G' be the subgraph of G induced by the selected vertices and let c' denote the number of crossings of the fix drawing of G whose corresponding four vertices of G all remain in G'. By the linearity of expectation, we come to

$$p^4 cr(G) = E(c') \geq E(cr(G')) > E(e') - 3E(n') = p^2 e - 3pn.$$

Set $p = 4n/e$ in the above inequality and then we are done. ∎

*D.B. West, Introduction to Graph Theory, Theorem 6.1.23, China Machine Press, 2004.

Exercise 30 *There are at most* $4n^{\frac{4}{3}}$ *pairs of points at distance* 1 *among a set of* n *points in the plane.*

The next example gives another typical way of decomposing a random variable into simpler ones.

Example 31 (The Coupon Collector's Problem) * *There are* n *types of infinitely many coupons and that each time one buys a box of detergent one obtains a coupon at random. How many boxes must be bought, on average, to collect all* n *types of coupons?*

*For more on this problem, its analysis and applications, see: Donald E. Knuth, Stable Marriage and its Relation to Other Combinatorial Problems: An Introduction to the Mathematical Analysis of Algorithms, American Mathematical Society, 1997.

Solution: The sample space here is $\{x = (x_1, x_2, \ldots) : \ x_i \in [n]\}$ and the probability is given by $P(\{x_{i_1} = a_1, \ldots, x_{i_k} = a_k\}) = (\frac{1}{n})^k$. Let X_t be the random variable which denotes the minimum i such that $x_i = t$. What we want to estimate is $E(X_n)$. Note that there is a decomposition $X_n = \sum_{t=1}^{n}(X_t - X_{t-1})$. But $E(X_t - X_{t-1}) =$ * $1 + \frac{t-1}{n} + (\frac{t-1}{n})^2 + (\frac{t-1}{n})^3 + \cdots = \frac{n}{n-t+1}$. We then conclude that $E(X_n) = n(\sum_{t=0}^{n-1} \frac{1}{1+t}) = nH_n \sim n \ln n$. † ∎

Exercise 32 *Based on the lecture notes of Beier, Canzar, and Funke, which is available at* `http://www.mpi-inf.mpg.de/~funke/Courses/randalg/lecture1.pdf`, *write a notes and prepare a presentation on the analysis of the Randomized Quicksort algorithm.*

*By Exercise 23 and use double counting to deduce $E(X) = \sum_{i \geq 1} P(X \geq i)$ for any positive integer valued random variable X!

†H_n is usually referred to be the nth harmonic number. For more on these important numbers, see `http://mathworld.wolfram.com/HarmonicNumber.html`

A 3CNF Boolean formula is a Boolean formula written in conjunctive normal form where each claus contains exactly three different literals, each literal being either a Boolean variable or negation of a Boolean variable. An example of 3CNF Boolean formula is, say, $(x_1 \vee \neg x_2 \vee x_2) \wedge (\neg x_2 \vee x_4 \vee \neg x_5) \wedge (\neg x_3 \vee x_4 \vee x_5)$.

Exercise 33 *For each 3CNF Boolean formula ϕ there exists a truth assignment that satisfies at least $\frac{7}{8}$ of clauses in ϕ.*

Exercise 34 (Buffon's needle, 1733) *Suppose we toss a needle of unit length at random onto the plane with equally spaced parallel lines a unit distance apart. Show that the probability that the needle lands on a line is $\frac{2}{\pi}$. (Hint: An elegant proof making use of linearity of expectation can be found in this very nice booklet: Daniel A. Klain, Gian-Carlo Rota, Introduction to Geometric Probability, Cambridge University Press, 1997.)*

Theorem 35 (Markov's inequality) *Let X be a random variable taking only nonnegative values. Then, for each $a > 0$, $P(X \geq a) \leq \frac{E(X)}{a}$.*

Proof. Since $\frac{X}{a} \geq \mathbf{I}_{(X \geq a)}$, Exercise 22 tells us that $\frac{E(X)}{a} = E(\frac{X}{a}) \geq E(\mathbf{I}_{(X \geq a)}) = P(X \geq a)$. ∎

Exercise 36 (The first moment method) *Let X be a nonnegative integer valued random variable. Then $P(X > 0) \leq E(X)$.*

Example 37 *If $p = o(n^{-1})$, then in $G(n,p)$ we have $\lim_{n\to\infty} P(G \in G(n,p)$ is a forest $) \to 1$, namely G is asymptotically almost surely (a.a.s.) a forest.*

Solution. To prove this fact, let $X_{n,k}$ be the random variable which counts the number of k-cycles. Then, $1 - P(G \in G(n,p)$ is a forest $) = P(\cup_{k=3}^{n}(X_{n,k} > 0)) \leq \sum_{k=3}^{n} P(X_{n,k} > 0) \leq \sum_{k=3}^{n} E(X_{n,k}) = \sum_{k=3}^{n}\binom{n}{k}\frac{(k-1)!}{2}p^k < \sum_{k=3}^{n}(np)^k < \frac{np}{1-np} \to 0$. Here, the first inequality comes from Exercise 10 while the second one is due to Exercise 36. ∎

In addition to the expected value, significant numerical characteristics of a random variable X are $E(X^r)$ and $E((X - E(X))^r)$, which are known as the moment and central moment, respectively, of order r of X.

The variance $\mathbf{var}(X)$ of a random variable X is the second central moment of X, that is, $\mathbf{var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$. By Theorem 35, for each $a > 0$, we have $P(|X - E(X)| \geq a) = P((X - E(X)^2 \geq a^2)) \leq \frac{E((X-E(X))^2}{a^2}$. This leads to

Corollary 38 (Chebyshev's Inequality) *Let X be a random variable. Then, for each $a > 0$, $P(|X - E(X)| \geq a) \leq \frac{\mathbf{var}(X)}{a^2}$.*

Exercise 39 (The second moment method) *Deduce from the Chebyshev's inequality that for any $a > 0$ and any nonnegative integer valued random variable X, $P(X = 0) \leq \frac{var(X)}{(E(X))^2} = \frac{E(X^2)}{(E(X))^2} - 1$.*

With the help of the Cauchy-Schwarz inequality (Exercise 22), we can strengthen the conclusion of Exercise 39 as follows.

Theorem 40 (The strong second moment method) *If X is a nonnegative integer valued random variable, then $P(X = 0) \leq \frac{\mathbf{var}(X)}{E(X^2)} = 1 - \frac{(E(X))^2}{E(X^2)}$.*

Proof. Note that $X = X\mathbf{I}_{\{X>0\}}$. It then follows from the Cauchy-Schwarz inequality that $E(X)^2 = (E(X\mathbf{I}_{\{X>0\}}))^2 \leq E(\mathbf{I}^2_{\{X>0\}})E(X^2) = P(X > 0)E(X^2)$. Consequently, $P(X = 0) = 1 - P(X > 0) \leq 1 - \frac{(E(X))^2}{E(X^2)} = \frac{\mathbf{var}(X)}{E(X^2)}$, as desired. ∎

Exercise 41 (Cantelli Inequality) * *For any $a > 0$ and any random variable X, it holds $P(X > a + E(X)) \leq \frac{\mathbf{var}(X)}{a^2 + var(X)}$.*

*This is a one-sided version of Theorem 35.

Exercise 42 $\mathbf{var}(X) = E(\mathbf{var}(X|Y)) + \mathbf{var}(E(X|Y))$.

The probability generating function or pgf of a random variable taking only nonnegative integer values is defined to be $G_X(z) = \sum_{k\geq 0} P(X = k)z^k = E(z^X)$.

Exercise 43 $G_X(1) = 1, G'_X(1) = E(X), G''_X(1)+G'_X(1)-G'_X(1)^2 = var(X)$.

From the generating function alone we can get by formal differentiation both the mean and the variance of a distribution as well as many other important parameters for the statistical regularity. Thus the pgf is a fundamental tool in handling random variables and solving probability problems.

At Leningrad University, which Perelman entered in 1982, at the age of sixteen, he took advanced classes in geometry and solved a problem posed by Yuri Burago, a mathematician at the Steklov Institute, who later became his Ph.D. adviser. There are a lot of students of high ability who speak before thinking, Burago said. Grisha was different. He thought deeply. His answers were always correct. He always checked very, very carefully. Burago added, He was not fast. Speed means nothing. Math doesn't depend on speed. It is about deep. – S. Nasar, D. Gruber, MANIFOLD DESTINY: A legendary problem and the battle over who solved it, The Newyorker, Aug. 28, 2006.

End of Lesson Three 19/9/06

Let $\Omega = \cup_{j=1}^{k_i} A_{ij}$ be a partition into disjoint events for each $i = 1, \ldots, t$. We say that these t partitions are collectively independent (statistically independent) provided $P(\cap_{i=1}^{t} A_{i,s_i}) = \prod_{i=1}^{t} P(A_{i,s_i})$, where $1 \leq s_i \leq k_i$.

Two events A and B are independent * if and only if the partition (A, A^c) and the partition (B, B^c) are collectively independent (pairwise independent). This indicates that the set of equalities in defining the independence of a set of partitions are not independent.

Exercise 44 *For $t = 3, 4, 5$, determine a minimal set of equalities among those given in the definition of collectively independence of partitions of a sample space which already imply all those inequalities. (Compare with Example 47.)*

*Page 0.

A family of events $\{A_1, A_2, \ldots, A_k\}$ is called a family of independent events if $P(\cap_{i\in I} A_i) = \prod_{i\in I} P(A_i)$ for any $I \subseteq [k]$. We say that $\{A_1, A_2, \ldots, A_k\}$ is k-wise independent provided any k events among them are independent.

Exercise 45 *$\{A_1, A_2, \ldots, A_k\}$ is a family of independent events if and only if the partitions (A_i, A_i^c), $i = 1, \ldots, t$, are collectively independent.*

Example 46 (Bernstein) *Consider two tosses of a fair coin, with success corresponding to heads and failure corresponding to tails. We examine the following three events: $A_1 = (\omega_1 = 1)$, that is, in the first toss the coin lands heads; $A_2 = (\omega_2 = 1)$, that is, in the second toss the coin lands heads; and $A_3 = (\omega_1 = \omega_2)$, that is, the results of the two tosses are the same. Then, these three events are pairwise independent but not independent.*

A Quiz: For any $n \geq 2$, construct a sample space Ω and a set of $(n-1)$-wise independent events $A_1, \ldots, A_n$ which are not collectively independent and $P(A_i) = \frac{1}{2}$. * One More Quiz: Let $1 \leq k \leq n$. (a) Construct a probability space and n events which are k-wise independent; but no $k+1$ of these events are independent. (b) Solve part (a) under the additional assumption that each of these n events has probability $\frac{1}{2}$. †.

*Let Ω be the set of binary strings of length $n-1$, each having probability $\frac{1}{2^{n-1}}$. For $i \in [n-1]$, let A_i be the event consisting of all strings whose ith position is occupied by 1. Let A_n be the event of all strings having an even number of 1's.

†Let the sample space be the k-dimensional vector space over a finite field F of order $q > n$ and associate it with a uniform probability distribution. Select n different elements $t_1, t_2, \ldots, t_n \in F$ and let $x_i = \begin{pmatrix} 1 & t_i & t_i^2 & \cdots & t_i^{k-1} \end{pmatrix}$ for $i \in [n]$. For $i \in [n]$, take the event A_i to be $\{y \in F^k : x_i y^\top = 0\}$. Note that $\#\{y \in F : x_{i_1} y^\top = \ldots = x_{i_t} y^\top = 0\} = q^{k-dim(Span(x_{i_1}, \ldots, x_{i_t}))}$. For a solution to part (b), go to Michal Karoński, Twelve Lectures on Combinatorial Probability, Theorem 16.4., Com^{2}Mac Lecture Note Series 11, 2003.

Example 47 *Let $n \geq 3$. Consider the sample space $2^{[n]}$. Assign probability 0 to the elementary events $\{x \in 2^{[n]} : \mathbf{supp}(x) = [n] \setminus \{n-1\}\}$ and $\{x : \mathbf{supp}(x) = \{n-1\}\}$, assign probability $\frac{1}{2^{n-1}}$ to the elementary events $\{x : \mathbf{supp}(x) = [n-2]\}$ and $\{x : \mathbf{supp}(x) = \{n-1, n\}\}$, and assign probability $\frac{1}{2^n}$ to any other elementary event. For $i \in [n]$, let X_i be the event $\{x \in 2^{[n]} : x(i) = 0\}$. We find that 1) $X_1, \ldots, X_{n-1}$ are independent; 2) $P(\cap_{i=1}^n X_i) = \prod_{i \in [n]} P(X_i)$, $P(X_i) = \frac{1}{2}$, $i \in [n]$, 3) X_n and X_{n-1} are independent; 4) X_n and X_{n-i}, $i \geq 2$, are not independent.*

Given events $B_1, \ldots, B_n$ and subsets $S \subseteq [n]$, consider the possible 'product rules' $P(\cap_{i \in S} B_i) = \prod_{i \in S} P(B_i)$. There are $2^n - n - 1$ nontrivial cases in which S has at least two elements. Richard M. Dudley provided the following example to show that all these $2^n - n - 1$ cases must be checked in order to claim the independence of $B_1, \ldots, B_n$.*

*J.P. Romano, A.F. Siegel, Counterexamples in Probability and Statistics, Wadsworth & Brooks, 1986.

Example 48 *Let $n \geq k \geq 2$. Let $B_j^1 = B_j$ and let B_j^0 be the complement of B_j. Consider the 2^n 'atoms' $B^S = \cap_{j=1}^n B_j^{S(j)}$ where $S \in [2]^{[n]}$. Let each atom that intersects any B_j, where $j > k$, have probability $\frac{1}{2^n}$. Let the events A_1 be defined as in the answer to the first quiz in last slide. For each atom B^S such that $S(j) = 0$ for all $j > k$, set $P(B^S) = P(\cap_{j=1}^k A_j^{S(j)})2^{k-n}$. For any proper subset S of $[k]$ with r elements, where $0 \leq r < k$, $P(\cap_{j\in S} B_j) = 2^{-r}(1 - 2^{k-n}) + 2^{k-n}P(\cap_{j\in S} A_j) = 2^{-r}$. If a set $M \subset [n]$ intersects $[n] \setminus [k]$ and has m elements, then $P(\cap_{j\in M} B_j) = 2^{-m}$. Thus, all product rules hold except that $P(\cap_{j=1}^k B_j) = 2^{-k}(1 - 2^{k-n}) + 2^{k-n}P(\cap_{j=1}^k A_j) \neq 2^{-k}$, as stated.*

Exercise 49 *If we have n independent events whose probability is neither zero nor one, then the sample space has at least 2^n points.*

A family of random variables are independent if the partitions associated with them are collectively independent. This generalizes the concept of independence for events, which corresponds to indicator variables.

For two random variables X and Y, we know that $E(X+Y)$ is determined by $E(X)$ and $E(Y)$. But, generally, the probability distribution of $X+Y$ cannot be read from the probability distributions of X and Y. We need to know the relationship between the distributions of X and Y. The independence relation is such a relation from which we can determine the random variable $X+Y$ provided we know X and Y.

Exercise 50 *Verify that all statements on the independence of random variables as given by Emmanuel Lesigne are equivalent to our definition.*

The greatest gifts you can give your children are the roots of responsibility and the wings of independence. – Denis Waitley

In certain sense, the concept of independence, which we are now going to introduce, play a central role in probability theory: it is precisely this concept that distinguishes probability theory from the general theory of measure spaces. – A.N. Shiryayev, Probability, Springer, 1984.

It may be said that no one could have learned the subject properly without acquiring some feeling for the intuitive content of the concept of stochastic independence, and through it, certain degrees of dependence. – Kai Lai Chung, A Course in Probability Theory, Second Edition, Academic Press, 1974.

End of Lesson Four 26/9/06

Example 51 *Take* $[n]$ *as the sample space and give* $i \in [n]$ *a probability* p_i*. Suppose there are* k *pairwise independent events* $A_i, i \in [k]$*, whose probabilities are neither zero nor one. Put the vector* $a_i \in R^n$*, whose* j*th coordinate is equal to zero if* $j \notin A_i$ *or to* $\sqrt{p_j}$ *if* $j \in A_i$*, into correspondence with every event* $A_i \subseteq [n]$*. Since* $P(A_i) \neq 0, 1$*, we know that* $P(A_i) - P(A_i)^2 > 0, \forall i \in [n]$*. On the other hand, observe that*

$$\langle a_i, a_j \rangle = E(I_{A_i} I_{A_j}) = \begin{cases} E(I_{A_i})E(I_{A_j}) = p_i p_j & \text{if } i \neq j; \\ E(I_{A_i}) = p_i = p_i^2 + (p_i - p_i^2), & \text{if } i = j. \end{cases}$$

Therefore , for $B = \begin{pmatrix} a_1 & a_2 & \cdots & a_k \end{pmatrix}$*, the matrix* $B^\top B$ *turns out to be a sum of the diagonal matrix whose* i*th diagonal entry is* $P(A_i) - P(A_i)^2 > 0$ *and the semi-definite positive matrix* $\alpha^\top \alpha$*, where* $\alpha = \begin{pmatrix} p_1 & p_2 & \cdots & p_k \end{pmatrix}$*. This is possible only when* $k \leq n$*.*

Exercise 52 *Let $X_1, \ldots, X_k$ be a family of $2m$-wise independent non-constant random variables over a sample space of n elements. Prove that $n \geq \binom{k}{m}$. (Hint: Follow the 'inner product' argument in Example 51.)*

Exercise 53 *For $n = 2^k - 1$, construct a probability space of size $n+1$ with n pairwise independent events each of which of probability $\frac{1}{2}$. (b) Same for n a prime number of form $4k-1$.*

Exercise 54 *If there exist n pairwise independent nontrivial events in a probability space (Ω, P), then $|\Omega| \geq 1 + n$.*

Let $\Omega = \cup_{j=1}^{k_i} A_{ij}$, $i = 1, \ldots, t$, be t collectively independent partitions of the sample space into disjoint events. Suppose $\Omega = \cup_{j=1}^{k'_i} B_{ij}$ is a partition whose parts are some unions of the parts of $\Omega = \cup_{j=1}^{k_i} A_{ij}$. It is clear that $\Omega = \cup_{j=1}^{k'_i} B_{ij}$, $i = 1, \ldots, t$, are still collectively independent. This observation gives the following

Theorem 55 *Suppose that $(X_i)_{i \in I}$ is a finite family of independent random variables and that J and K are disjoint subsets of I. If Y is a real function defined on $(X_i)_{i \in J}$ and Z is a real function defined on $(X_i)_{i \in K}$, then Y and Z are independent random variables.*

Theorem 56 *For collectively independent random variables X_1, $\ldots$, X_n, it holds $E(\prod_i X_i) = \prod_i E(X_i)$.*

Proof. $E(\prod_i X_i) =^* \sum_{a_i}(\prod_i a_i P(X_i = a_i)) = \prod_i(\sum_j a_j P(X_i = a_j)) = \prod_i E(X_i)$. ∎

Exercise 57 *If X and Y are independent random variables each taking only nonnegative integer values, then $G_X(z)G_Y(z) = G_{X+Y}(z)$.*

Theorem 58 *For pairwise independent random variables $X_1, \ldots, X_n$, $\mathbf{var}(\sum_i X_i) = \sum_i \mathbf{var}(X_i)$.*

Proof. Assume, w.l.o.g., that $E(X_i) = 0$. Put $X = \sum_i X_i$. We have $\mathbf{var}(X) = E(X^2) = E(\sum_i X_i^2 + 2\sum_{i<j} X_i X_j) = \sum_i E(X_i^2) + 2\sum_{i<j} E(X_i X_j) = \sum_i \mathbf{var}(X_i)$. ∎

*Write $\prod_i X_i$ as a linear combination of some indicator variables and then use the linearity of expectations.

Notation: We write P_n for $(1-p,p)^{\bigotimes n}$ and E_n for the expected value associated with P_n. Let S_n be the random variable denoting the number of successes (1's) for each outcome of the product space (Ω_n, P_n).

Note that S_n is the sum of n independent random variables S_1. This gives $E(S_n) = E_n(S_n) = nE(S_1) = np$ and $\mathbf{var}(S_n) = n\mathbf{var}(S_1) = npq$. This observation generalizes to all product space.

A random variable X follows the Bernoulli distribution with parameter p if it only takes the values 0 and 1 and if $P(X = 1) = p$. If $X_1, \ldots, X_n$ are n independent random variables following a Bernoulli distribution with parameter p, then the distribution of $\sum_{i=1}^{n} X_i$ is called the binomial distribution with parameters n and p. S_1 follows Bernoulli distribution and S_n the binomial distribution. Please investigate the relationship between binomial distribution and binomial coefficients.

Theorem 59 (Weak Law of Large Numbers) * *For each* $\epsilon > 0$, $P_n(|\frac{S_n}{n} - p| > \epsilon) \to 0$ *as* n *approaches infinity and this convergence is uniform in* p.

Proof. By Chebyshev inequality, $P_n(|S_n - np| > n\epsilon) \leq \frac{\mathbf{var}(S_n)}{(n\epsilon)^2} = \frac{p(1-p)}{n\epsilon^2}$. ∎

Theorem 60 †*For all* $m \geq 1$, *we have* $\binom{2m}{m} \geq \frac{2^{2m}}{4\sqrt{m}}$.

*This result appeared for the first time in a posthumously published work by Jacob Bernoulli in 1713. This was named the 'weak law of large numbers' by Siméon Denis Poisson in a paper in 1837, who generalized Bernoulli's result to cases where the probability of success varies from trial to trial.

†Jiří Matoušek, Jan Vondrák, The probabilistic method, preprint.

Proof. Consider the binomial distribution S_{2m} with parameter $2m$ and $\frac{1}{2}$. We have $E(S_{2m}) = m$ and $\mathbf{var}(S_{2m}) = \frac{m}{2}$. The Chebyshev inequality gives $P(|S_{2m} - m| < \sqrt{m}) \geq \frac{1}{2}$. The probability of S_{2m} attaining a specific value $m + k$, where $|k| < \sqrt{m}$, is $\binom{2m}{m+k}2^{-2m} \leq \binom{2m}{m}2^{-2m}$. So we have $\frac{1}{2} \leq \sum_{|k|<\sqrt{m}} P(S_{2m} = m + k) \leq (2\sqrt{m} + 1)\binom{2m}{m}2^{-2m}$ and the result follows. ∎

Exercise 61 * *You and the bank play the following game: You flip n coins: If t of them come up "heads", you receive 2^t dollars. 1) You have to buy a ticket to play the game. What is the fair price of the ticket? 2) Prove that the probability that you break even is exponentially small; 3) Calculate the standard deviation of the variable 2^t. 5) State what the "weak law of large numbers" would say for the variable 2^t. Prove that the Law does not hold.*

*László Babai, Exercise 7.4.16, Discrete Mathematics, Lecture Notes, 2003.

Exercise 62 *Use Stirling's formula to show that* $\binom{2m}{m} \sim \frac{4^m}{\sqrt{\pi m}}$.

Exercise 63 * *Consider the space* (Ω_n, P_n) †. *Let* A_i *be the event that* $\omega^i = 0$ *and* $\omega^{i-1} = 1$ *for* $i \geq 2$. *If* U_n *is the number of times* A_i *occurs for* $2 \leq i \leq n$, *prove that* $E(U_n) = (n-1)p(1-p)$ *and find the variance of* U_n.

Exercise 64 *Let* $X_1, \ldots, X_n$ *be a family of independent random variables with the same probability distribution as that of* X. *Let* $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$. *Then* $\mathbf{var}(X) = E(\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2)$.

*A question from a probability course final exam at Oxford 1977.

†Recall that $P_n = (1-p, p)^{\bigotimes n}$!

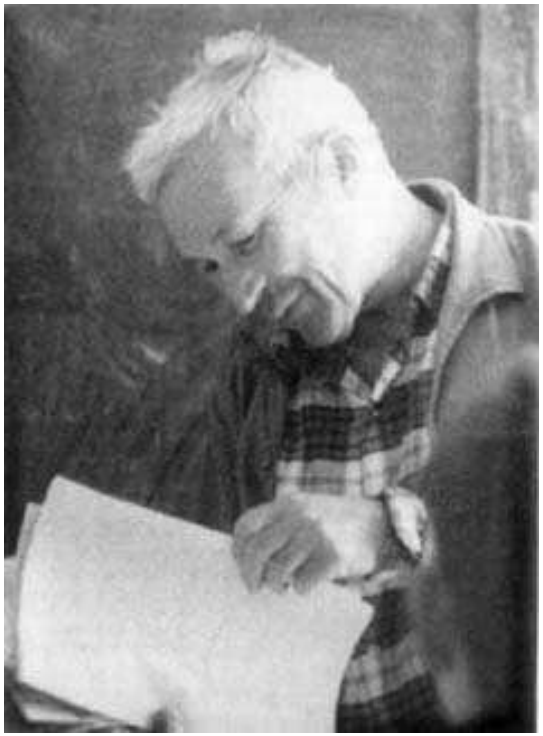
László Babai, Discrete Mathematics, Lecture Notes, 2003.

A binary BCH code construction of d-wise independent random variables: N. Alon, J.H. Spencer, Derandomization, Chap. 15, The Probabilistic Method, Wiley, 1992.

Dany Breslauer, Devdatt P. Dubhashi, Combinatorics for Computer Scientists, Chap. 24, BRICS, 1995, `http://www.brics.dk/LS/95/4/BRICS-LS-95-4/BRICS-LS-95-4.html`.

Exercise 65 *Choose one of the following Russian mathematicians and write an essay on his life and his mathematics: Pafnuty Lvovich Chebyshev, Andrei Andreyevich Markov, Sergei Natanovich Bernstein, Andrey Nikolaevich Kolmogorov. Please do not forget to indicate those references which you have used.*

Who is him?



http://www-groups.dcs.st-and.ac.uk/~history/Biographies/

Read some interesting stuff on probability theory and prepare to give a 20 minutes presentation. Here is a list of suggested readings for your National Day Holiday*:

Arnold Knopfmacher, Helmut Prodinger, A simple card guessing game revisited, The Electronic Journal of Combinatorics, Volume 8(2), 2001, http://www.combinatorics.org/Volume_8/Abstracts/v8i2r13.html.

Richard Ehrenborg, A bijective answer to a question of Zvonkin, Annals of Combinatorics, 4 (2000), 195-197. http://www.ms.uky.edu/~jrge/Papers/Normal.pdf#search=%22A%20bijective%20answer%20to%20a%20question%20of%22

*Find more relevant material from the internet!

Ilan Adler, Shmuel Oren, Sheldon M. Ross, The coupon-collector's problem revisited, Journal of Applied Probability, 40, (2003), 513–518.

Frank K. Hwang, Knock Down Tournament, http://episte.math.ntu.edu.tw/articles/mm/mm_03_2_01/index.html

Here is a reading report of a student in Taiwan: http://lib.fg.tp.edu.tw/research/%E7%AC%AC%E5%8D%81%E4%BA%8C%E8%BC%AF/E6%95%B8%E5%AD%B8/%E6%B7%98%E6%B1%B0%E8%B3%BD.doc

We are plunging down a cataract, and what's important is to call out. Not for help, there is no help. Not in despair – what can anyone do but shrug, look away? But to give a signal. A gesture of love and humor to acknowledge drawing so others who drown will know they are not alone. – Allen Wheelis, On Not Knowing How to Live, Harper and Row, 1975.

End of Lesson Five 29/9/06

Theorem 66 (Weierstrass's polynomial approximation theorem) *Let f be a continuous real function on $[0,1]$ and define the Bernstein polynomial $\{B_n\}$ as follows: $B_n(p) = \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k}p^k(1-p)^{n-k}$. Then B_n converges uniformly to f in $[0,1]$, namely*

$$\lim_{n\to\infty} \sup_{0\le p\le 1} |f(p) - B_n(p)| = 0.$$

Proof. Consider the probability space (Ω_n, P_n) and the random variable $f(\frac{S_n}{n})$. Bernstein's polynomial B_n turns out to be a meaningful construction regarding (Ω_n, P_n). In fact, we have $E_n(f(\frac{S_n}{n})) = B_n(p)$. Thus, $|f(p) - B_n(p)| = |E(f(p) - f(\frac{S_n}{n}))| \le E(|f(p) - f(\frac{S_n}{n})|) = E(f_1) + E(f_2)$, where $f_1 = |f(p) - f(\frac{S_n}{n})|\mathbf{I}_{\{|\frac{S_n}{n}-p|<\delta\}}$ and $f_2 = |f(p) - f(\frac{S_n}{n})|\mathbf{I}_{\{|\frac{S_n}{n}-p|\ge\delta\}}$*.

*This decomposition into a sum of f_1 and f_2 is induced by a partition of unity, which instead comes from a partition of the whole sample space. This strategy is used often in mathematics.

Given any $\epsilon > 0$, we need to choose a good δ so that for sufficiently large n (which should be independent of p) both $E(f_1)$ and $E(f_2)$ can be bounded by $\frac{\epsilon}{2}$. Since a continuous function on $[0,1]$ must be uniformly continuous, we can choose $\delta > 0$ such that for any $x, y \in [0,1]$ with $|x-y| < \delta$ we have $|f(x)-f(y)| < \frac{\epsilon}{2}$. With this choice of δ, we get $E(f_1) < \frac{\epsilon}{2}$. On the other hand, we have $E(f_2) \leq 2MP(\{|\frac{S_n}{n} - p| \geq \delta\})$, where $M = \max_{x\in[0,1]} |f(x)| < \infty$. Note that $E(\frac{S_n}{n} - p) = 0$ and $\mathbf{var}(\frac{S_n}{n} - p) = \frac{p(1-p)}{n}$. It then follows from Corollary 38 that $E(f_2) \leq 2M\frac{p(1-p)}{n\delta^2} \leq \frac{M}{2n\delta^2}$. Picking n large enough such that $\frac{M}{2n\delta^2} < \frac{\epsilon}{2}$, we arrive at $|f(p) - B_n(p)| < \epsilon$, finishing the proof. ∎

When estimating $E(f_2)$ in the above proof, we are repeating the proof of the weak law of large numbers (Theorem 59). We can also directly appeal to Theorem 59 instead of making use of the Chebyshev's inequality (Corollary 38).

Exercise 67 * *Let f be continuous and belong to $L^r(0,\infty)$ for some $r > 1$, and $g(\lambda) = \int_0^\infty \exp(-\lambda t) f(t) dt$. Then $f(x) = \lim_{n\to\infty} \frac{(-1)^{n-1}}{(n-1)!}(\frac{n}{x})^n g^{(n-1)}(\frac{n}{x})$, where $g^{(n-1)}$ is the $(n-1)$st derivative of g, uniformly in every finite interval.*

A simplicial complex is a set system closed under the taking subset operation.

Kleitman Lemma: Suppose $\mathcal{A} \subseteq 2^{\{1,\ldots,n\}}$ and $\mathcal{B} \subseteq 2^{\{1,\ldots,n\}}$ are two simplicial complexes. Then $|\mathcal{A} \cap \mathcal{B}| \geq \frac{|\mathcal{A}||\mathcal{B}|}{2^n}$.

We introduce a special case of the famous FKG inequality †, which is strong enough to imply Kleitman Lemma easily.

*Kai Lai Chung, A Course in Probability Theory, 2nd Ed., p. 140, Exercise 11, Academic Press, 1974.

†C.M. Fortuin, P.W. Kasteleyn, J. Ginibre, Correlation inequalities on some partially ordered sets, Commun. Math. Physics, 22 (1971), 89–103.

FKG inequality is first discovered when considering a problem in statistical mechanics.

In the realm of probability inequalities, the FKG inequality, due to Fortuin, Kasteleyn and Ginibre (1971), now occupies a position of fundamental importance because of its simplicity and widespread applications. – D.S.P. Richards, Algebraic methods toward higher-order probability inequalities II, The Annals of Probability, 32 (2004), 1509–1544.

For a discussion of the general FKG inequality as well as an application to percolation theory, see the notes of Dana Randall for the course "Combinatorial Methods for Statistical Physics Models" at `http://www.math.gatech.edu/~randall/topics2.html`

For a real vector $p = \begin{pmatrix} p_1 & \dots & p_n \end{pmatrix}$, where $p_i \in [0,1]$, put the probability Pr_p on Ω_n to be $Pr_p(\omega) = \prod_{\omega_i=1} p_i \prod_{\omega_i=0}(1-p_i)$. Note that (Ω_n, Pr_p) is just the product space of n spaces $\{0,1\}$ equipped with the probability $(1-p_i, p_i), i = 1, \dots, n$. This is a model of flipping n not necessarily identical coins independently.*

For any two $(0,1)$ vectors u and v of Ω_n, we write $u \succ v$ if $u - v$ is a nonnegative vector. This defines a partial order on Ω_n. A random variable X on (Ω_n, Pr_p) is monotone increasing (decreasing) if $X(u) \geq X(v)$ $(X(u) \leq X(v))$ whenever $u \succ v$. An event is monotone increasing (decreasing) if the corresponding indicator variable is monotone increasing (decreasing). $A \subseteq \Omega_n$ is monotone increasing iff $\{[n] - \mathbf{supp}(\omega) : \ \omega \in A\}$ is a simplicial complex. $A \subseteq \Omega_n$ is monotone decreasing iff $\{\mathbf{supp}(\omega) : \ \omega \in A\}$ is a simplicial complex.

*This is called a Poisson trial. When all p_i are equal, we go back to Bernoulli trial.

Two random variables X and Y are positively (negatively) correlated if $E(XY) \geq E(X)E(Y)$ ($E(XY) \leq E(X)E(Y)$). Two indicator variables $\mathbf{I}_A$ and $\mathbf{I}_B$ are positively correlated iff $P(A|B) \geq P(A)$.

Here comes a special FKG inequality.

Theorem 68 *Let X and Y be two monotone increasing random variables. Then X and Y are positively correlated.*

Corollary 69 *Let A, B be monotone increasing events of (Ω_n, Pr_n) and C, D monotone decreasing events of (Ω_n, Pr_p). Then it holds $Pr_p(A \cap B) \geq Pr_p(A)Pr_p(B)$, $Pr_p(C \cap D) \geq Pr_p(C)Pr_p(D)$, $Pr_p(A \cap C) \leq Pr_p(A)Pr_p(C)$.*

Proof. (of Theorem 68) We use induction on n.

For $n = 1$, $E(X) = X(1)p_1 + X(0)q_1, E(Y) = Y(1)p_1 + Y(0)q_1$ and $E(XY) = X(1)Y(1)p_1 + X(0)Y(0)q_1$. Note that $\begin{pmatrix} p_1 & \\ & 1-p_1 \end{pmatrix} = \begin{pmatrix} p_1 \\ 1-p_1 \end{pmatrix} \begin{pmatrix} p_1 & 1-p_1 \end{pmatrix} + \begin{pmatrix} p_1 \\ -p_1 \end{pmatrix} \begin{pmatrix} 1-p_1 & p_1-1 \end{pmatrix}$. It then follows $E(XY) - E(X)E(Y) = \begin{pmatrix} X(1) & X(0) \end{pmatrix} \begin{pmatrix} p_1 \\ -p_1 \end{pmatrix} \begin{pmatrix} 1-p_1 & p_1-1 \end{pmatrix} \begin{pmatrix} Y(1) \\ Y(0) \end{pmatrix} = p_1 q_1 (X(1) - X(0))(Y(1) - Y(0))$. By the monotone increasing property we have $X(1) \geq X(0)$ and $Y(1) \geq Y(0)$ and hence $E(XY) \geq E(X)E(Y)$ follows. *

*Deducing this inequality from the known conditions can be done with the middle school mathematics. We are still using that mathematics but we formulate it in the matrix theory language. When dealing with Exercise 71, you will realize that sometimes the correct language may be very important.

To finish the induction step, the key is double counting, as in many other places. More precisely, we need the law of iterated expectation, Exercise 23. If Z is a random variable on (Ω_n, Pr_p) and $v \in \Omega_{n-1}$, let Z_v denote the random variable on $(\Omega_1, (1 - p_n, p_n))$ such that $Z_v(0) = Z(v, 0)$ and $Z(v, 1) = Z_v(1)$. Note that setting $Z_{[n-1]}(v) = E(Z_v)$ defines a random variable $Z_{[n-1]}$ on $(\Omega_{n-1}, Pr_{p'})$, where p' is obtained from p be deleting its last entry. By the law of iterated expectation (Exercise 23), we find $E(Z) = E(Z_{[n-1]})$.

Assume $n \geq 2$ and the statement is already proved for smaller n.

For any fixed vector $v \in \Omega_{n-1}$, clearly X_v and Y_v are still monotone increasing. This gives $E(X_v Y_v) \geq E(X_v)E(Y_v)$, in view of the induction hypothesis for dimension one. Therefore,

$$E(XY) = E((XY)_{[n-1]}) \geq E(X_{[n-1]} Y_{[n-1]}). \tag{6}$$

Moreover, we observe that both $X_{[n-1]}$ and $Y_{[n-1]}$ are monotone increasing on $(\Omega_{n-1}, Pr_{p'})$. This enables us to utilize the induction hypothesis for dimension $n-1$ and deduce

$$E(X_{[n-1]}Y_{[n-1]}) \geq E(X_{[n-1]}) \times E(Y_{[n-1]}) = E(X)E(Y). \quad (7)$$

A combination of Eqs. (6) and (7) then completes the proof. ∎

Corollary 70 (Square root trick) *Suppose $A_1, \ldots, A_m$ are m monotone increasing events with the same probability α. Then $\alpha \geq 1 - (1 - P(\cup_{i=1}^m A_i))^{\frac{1}{m}}$.*

Proof. $1 - P(\cup_{i=1}^m A_i) = P(\cap_i A_i^c) \geq \prod_i P(A_i^c) = (1-\alpha)^m$. ∎

Exercise 71 *1) Let* $p = \begin{pmatrix} p_1 & \dots & p_n \end{pmatrix}$. *If* $\sum_i p_i = 1$, *then* $diag(p) - p^\top p = \sum_{i<j} p_i p_j (e_i - e_j)^\top (e_i - e_j)$, *where* e_i *is the* i*th unit column vector. 2) Consider a probability space on* $[n]$ *and two random variables* X *and* Y *satisfying* $(X(i) - X(j))(Y(i) - Y(j)) \geq 0$ *for all* $i, j \in [n]$. *Prove that* X *and* Y *are positively correlated and hence obtain another special FKG inequality. 3) Try to guess and prove more correlation inequalities.*

R.L. Graham, Applications of the FKG inequality and its relatives, Mathematical Programming: The State of the Art, (Eds., A. Bachem, M. Grötschel, B. Korte), pp. 115 – 131, Springer, 1983.

Graham R. Brightwell, William T. Trotter, A combinatorial approach to correlation inequalities, Discrete Mathematics, Kleitman and combinatorics: a celebration, 257, (2002) 311 – 327.

Jeffrey Steif, Percolation Theory, `http://www.math.chalmers.se/~steif/perc.html`

Gábor Lugosi, Concentration-of-measure inequalities, `http://www.econ.upf.es/~lugosi/surveys.html`

You cannot learn that which you already understand, so it is only by not understanding that you are capable of learning. Thus, only the intelligent admit ignorance and only the ignorant claim certainty. – Anonymous

It's kind of fun to do the impossible. – Walt Disney

I believe in an approach to research that complements knowing what is known with knowing what is not known. – Herbert Edelsbrunner, Geometry and Topology for Mesh Generation, Cambridge University Press, 2001.

End of Lesson Six 10/10/06

Example 72 *Let $\mathcal{F} \subseteq 2^{[n]}$ be a set system such that $A \cap B \neq \emptyset$ and $A \cup B \neq [n]$ (namely, $A^c \cap B^c \neq \emptyset$) for any $A, B \in \mathcal{F}$. Then $|\mathcal{F}| \leq 2^{n-2}$.*

Solution: 1. If $\mathcal{G} \subseteq 2^{[n]}$ is a set system satisfying $A \cup B \neq [n]$ for any $A, B \in \mathcal{G}$, then $|\mathcal{G}| \leq 2^{n-1}$. If $\mathcal{G} \subseteq 2^{[n]}$ is a set system satisfying $A \cap B \neq \emptyset$ for any $A, B \in \mathcal{G}$, then $|\mathcal{G}| = |\{A : \ A^c \in \mathcal{G}\}| \leq 2^{n-1}$.

2. Let $\mathcal{F}_1 = \{A : \ \exists B \in \mathcal{F}, [n] \supseteq A \supseteq B\}$, $\mathcal{F}_2 = \{A : \ \exists B \in \mathcal{F}, A \subseteq B\}$. Then, $|\mathcal{F}_1| \leq 2^{n-1}$, $|\mathcal{F}_2| \leq 2^{n-1}$.

3. By FKG inequality (Corollary 69), we get $\frac{|\mathcal{F}|}{2^n} \leq \frac{|\mathcal{F}_1 \cap \mathcal{F}_2|}{2^n} \leq \frac{|\mathcal{F}_1|}{2^n}\frac{|\mathcal{F}_2|}{2^n} \leq \frac{1}{4}$.

Setting $\mathcal{F} = \{A \cup \{n-1\} : \ A \subseteq [n-2]\}$, we see that the bound 2^{n-2} is sharp.

Let $p = \begin{pmatrix} p_1 & \dots & p_n \end{pmatrix}$, where $p_i \in [0,1]$. Consider the probability space (Ω_n, Pr_p). For $A \subseteq [n]$, define $A\uparrow = \{\omega \in \Omega_n : \mathbf{supp}(\omega) \supseteq A\}$. Clearly, for each $A \subseteq [n]$, $\mathbf{I}_{A\uparrow}$ is an increasing random variable. Let $A_1, \dots, A_t \subseteq [n]$, $X_i = \mathbf{I}_{A_i\uparrow}$, and $X = \sum X_i$.

Theorem 73 * $P(X = 0) \geq \exp(\frac{-E(X)}{1-\max p_i})$.

Proof. By FKG inequality and the fact that $1 - x \geq e^{\frac{-x}{1-x}}$ for any $x \in [0,1]$ †, we have $P(X = 0) = P(\prod(1 - X_i) = 1) = E(\prod(1-X_i)) \geq \prod(1-E(X_i)) \geq \exp(\frac{-E(X)}{1-\max_i E(X_i)}) = \exp(\frac{-E(X)}{1-\max p_i})$. ∎

*S. Janson, T. Luczak, A. Rucinski, Random Graphs, Corollary 2.13, Wiley, 2000.

†$\ln(\frac{1}{1-x}) \leq \frac{1}{1-x} - 1$

J. Nedelman, T. Wallenius, Bernoulli trials, Poisson trials, surprising variances, and Jensen's inequality, The American Statistician, 40 (1986), 286 – 289.

A function f is concave (convex downward) provided $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for any $\lambda \in [0,1]$. f is convex (convex upward) whenever $-f$ is concave.

Exercise 74 (Jensen's inequality) *For a concave function f and any random variable X, it holds $E(f(X)) \geq f(E(X))$.* Make use of the fact that $f(x) = x(1-x)$ is a concave function to reexamine Exercise 75. Use Jensen's inequality to show that arithmetic mean $\geq$ geometric mean $\geq$ harmonic mean.*

*Choose α such that $f(x) - f(E(X)) \geq \alpha(x - E(X))$. Then $E(f(X)) \geq E(\alpha(X - E(X)) + f(E(X))$.

Exercise 75 *Let $X_1, \ldots, X_n$ be n independent trials, where $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$. Let $X = \sum_i X_i$ and $\overline{p} = \frac{\sum_i p_i}{n}$. 1) Show that $E(X) = n\overline{p}$ and $\mathbf{var}(X) = n\overline{p}(1-\overline{p}) - \sum_i (p_i - \overline{p})^2$. 2) Fixing n and $E(X)$, $\mathbf{var}(X)$ is maximized by the binary distribution*. 3) Fixing n and $E(X)$, the minimum value of $\mathbf{var}(X)$ is $(n\overline{p} - k)(1 - n\overline{p} + k)$ and is attained when $p_1 = \cdots = p_{n-1-k} = 0$, $p_{n-k} = n\overline{p} - k$, $p_{n-k+1} = \cdots = p_n = 1$, where $k = \lfloor n\overline{p} \rfloor$. 4) (Hoeffding's Inequality)† Take any integer $b \in [0, n\overline{p} - 1]$ and $c \in [n\overline{p} + 1, n]$. Let g be a convex function. Fixing n and $E(X)$, $P(X \leq b)$, $P(X \geq c)$, and $\sum_{i=0}^{n} g(i)P(X = i)$ are all maximized by the binary distribution.*

*That is, Bernoulli trials is the best among all Poisson trials regarding the purpose of reducing variance. The lack of uniformity decreases the magnitude of chance fluctuations.

†W. Hoeffding, On the distribution of the number of successes in independent trials, Ann. Math. Stat. 35 (1964), 713–721.

Exercise 76 *Let* $(a_0, \ldots, a_n)$ *be a sequence of nonnegative real numbers with associated polynomial* $A(z) = \sum_{i=0}^{n} a_k z^k$ *such that* $A(1) > 0$. *The following conditions are equivalent:*

(i) the polynomial $A(z)$ *is either constant or has only real zeros;*

(ii) every minor of the infinite matrix $M = (a_{i-j})_{i,j=0,1,2,\ldots}$ *is nonnegative, where* a_t *should be viewed to be zero whenever* $t \neq 0, 1, \ldots, n$.*

(iii) The normalized sequence $(\frac{a_0}{A(1)}, \frac{a_1}{A(1)}, \ldots, \frac{a_n}{A(1)})$ *is the distribution of the number* X *of successes in* n *independent trials with probability* p_i *of success on the* i*th trial, for some sequence of probabilities* $0 \leq p_i \leq 1$. *The roots of* $A(z)$ *are given by* $\frac{p_i - 1}{p_i}$ *for* i *with* $p_i > 0$.

*This simply says that the sequence $(a_0, \ldots, a_n)$ is a Pólya frequency sequence.

L.H. Harper, Stirling behavior is asymptotically normal, Ann. Math. Stat. 38 (1966), 410–414.

S. Karlin, Total Positivity, Stanford Univ. Press, Stanford, 1968.

T. Ando, Totally positive matrices, Linear Algebra Appl. 90 (1987), 165–219.

F. Brenti, The application of total positivity to combinatorics, and conversely, in "Total Positivity and Its Applications" (M. Gasca and C.A. Micchelli, Ed.), Mathematics and Its Applications, Vol. 359, pp. 451–473, Kluwer, Dordrecht, 1996.

Jim Pitman, Probabilistic bounds on the coefficients of polynomials with only real zeros, Journal of Combinatorial Theory, Series A 77 (1997), 279–303.

Warm Up; Answer Questions

After that, I like to tell graduate students that, if they launch into a field and get frustrated because they didn't understand it, that is not necessarily time wasted. In this particular example – my launching into the Onsager solution – the reason I could understand the key idea during the 15 minutes ride with Luttinger was that I had "prelearned" the whole subject very well, though without true understanding. When the important point was revealed to me, I was able to immediately appreciate the whole strategy. – Chen Ning Yang

End of Lesson Seven 13/10/06

Warm Up; Answer Questions

Exercise 77 *If X is distributed in $B(n,p)$, then, for any $c \geq 0$ and any integer $j \geq 0$, $E((X+j)^c) \leq (j+np)^c$.*[*]

[*]H. Brönnimann, B. Chazelle, J. Matoušek, Product range spaces, sensitive sampling, and derandomization, SIAM J. Comput. 28 (1999), 1552–1575.

Student Presentation*

SHEN, Hongda, Randomized Quicksort Algorithm; HOU, Xiufeng, Knock Down Tournament

It is not possible to wait for inspiration, and even inspiration alone is not sufficient. Work and more work is necessary. Man blessed by genius can create nothing really great, not even anything mediocre, if he does not toil as hard as a slave. – Piotr Ilyich Tschaikovsky

End of Lesson Eight 17/10/06

*As listed by Tom Roby, you have to gain experience in the following skills of global importance within (and beyond) mathematics: constructing and carefully writing up rigorous proofs; solving problems for which templates are not given; presenting mathematics to others.

Let $(p_1, p_2 \ldots, p_n)$ be a finite probability distribution, namely, $p_i \geq 0$, and $\sum_i p_i = 1$. Shannon defined the entropy of this distribution to be $H = -\sum_i p_i \mathbf{log} p_i$ with $0 \cdot \mathbf{log}(0) = 0$. The entropy of a random variable X is defined to be the entropy of its probability distribution and denoted $H(X)$.

Information units: bits (binary digit) for base 2, Hartley for base 10 and nats for base e.

The words "uncertainty", "surprise", and "information" are related. Before the event (experiment, reception of a message symbol, etc.) there is the amount of uncertainty; when the event happens there is the amount of surprise; and after the event there is the gain in the amount of information. All these amounts are the same. ... The entropy function of a distribution summarizes one aspect of a distribution much as the average in statistics summarizes a distribution. The entropy has properties of both the arithmetic mean (the average) and the geometric mean.– R.W. Hamming, Coding and Information Theory, Prentice-Hall, 1980.

Conditional entropy $H(X|Y)$ is defined to be $E_Y(H(X_y))$ where for $y \in Im(Y)$, X_y is a random variable taking values in $Im(X)$ such that $P(X_y = x) = P(X = x|Y = y)$. That is, $H(X|Y)$ is the expectation of the random variable composed by two functions: $\Omega \rightarrow_Y y \in Im(Y) \rightarrow_{H(X_y)} R$.

Exercise 78 (Chain rule) $H(X,Y) = H(X) + H(Y|X)$. *Especially, when X and Y are independent, $H(X,Y) = H(X) + H(Y)$. By induction, we have $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|\, X_1, \ldots, X_{i-1})$.*

It follows from Exercise 78 that

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

This common value is referred to as the mutual information of X and Y, denoted $I(X,Y)$, which reflects the average amount of knowledge about X (Y) when we know Y (X).

Relative entropy is a measure of the distance between two distributions, also known as the Kullback-Leibler divergence*. For two probability distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ on the space $[n]$, $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$. Note that $I(X, Y) = D(p_{(X,Y)}||p_X p_Y)$ † and so the ensuing lemma tells us that conditioning reduces entropy.

Lemma 79 (Gibbs' Inequality) $D(p||q) \geq 0$. $D(p||q) = 0$ *if and only if* $p = q$.

Proof. W.l.o.g., we use logarithms to the base e. By $\ln x \leq x - 1$, we get $D(p||q) = -\sum_i p_i \ln \frac{q_i}{p_i} \geq -\sum_i p_i (\frac{q_i}{p_i} - 1) = 0$. ∎

*It is not symmetric. Neither does it satisfy the triangle inequality.

†If X and Y are independent, the joint distribution of X and Y is completely determined by that of X and that of Y and so $I(X, Y) = D(p_{(X,Y)}||p_X p_Y) = 0$ is easily understood – it is surely trivial to verify it by computation.

Proof. [Another proof of Lemma 79]* Consider the probability space $[n]$ with $P(i) = p(i)$. Consider the random variable $X(i) = \mathbf{log}(\frac{q(i)}{p(i)})$. Since $f(x) = \mathbf{log}(x)$ is convex, by Exercise 74, $D(p||q) = -E(f(X)) \geq -f(E(X)) = -\mathbf{log}(E(X)) = 0$.

Exercise 80 *The total correlation of the random variables $X_1, \ldots, X_n$ is defined to be $C(X_1, \ldots, X_n) = \sum_i H(X_i) - H(X_1, \ldots, X_n)$. This parameter represents the absolute information redundancy present in the given set of variables and is one of several generalizations of the mutual information. Show that $C(X_1, \ldots, X_n) = D(P_{(X_1,\ldots,X_n)} || \prod_i P_{X_i})$.*

*Since $\ln x \leq x-1$ means that $y = \ln(x)$ lies below the supporting line $y = x-1$, which, among many other methods, can be seen from the convexity of the ln function, these two proofs can be said to be essentially the same.

Notice the following correspondence between random variables and sets. Let μ be any additive set function.

$$H(X) \longleftrightarrow \mu(A)$$

$$H(X,Y) \longleftrightarrow \mu(A \cup B)$$

$$H(X|Y) \longleftrightarrow \mu(A \setminus B)$$

$$I(X,Y) \longleftrightarrow \mu(A \cap B)$$

$$\ldots$$

A theorem of Guoding Hu says that any relation involving those left-hand side terms is equivalent to a relation on set systems via the above correspondence. For example, $I(X,(Y,Z)) = I(X,Z) + I(X,Y|Z) \leftrightarrow \mu(A \cap (B \cup C)) = \mu(A \cap C) + \mu(A \cap (B \setminus C))$, $I(X,Y|Z) = H(X|Z) - H(X|Y,Z) \leftrightarrow \mu(A \cap (B \setminus C)) = \mu(A \setminus C) - \mu((A \setminus B) \cup C)$.

Exercise 81 *Let P and Q be the probability distributions $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_m)$, respectively. For $\lambda \in [0, 1]$, let $\lambda P + (1-\lambda)Q$ denote the probability distribution $(\lambda p_1, \ldots, \lambda p_n, (1-\lambda)q_1, \ldots, (1-\lambda)q_m)$. Show that the entropy function is convex up, namely $H(\lambda P + (1-\lambda)Q) \geq \lambda H(P) + (1-\lambda)H(Q)$.*

Exercise 82 *Let X be a random variable over (Ω, P). If the probability of no single value of X exceeds 2^{-t}, for some $0 \leq t \leq \log_2(|\Omega|)$, then $H(X) \geq t$.*

Exercise 83 *Prove that $H(X|Y) \leq H(X|Z) + H(Z|Y)$.*

Exercise 84 *Consider the probability distribution $P = (p, 1-p)$. If $p \leq \frac{1}{2}$, then $\sum_{0 \leq i \leq pn} \binom{n}{i} \leq 2^{nH(P)}$.*

There is no doubt that in a year to come the study of entropy will become a permanent part of probability theory. – A.I. Khinchin

Corollary 85 *If P is a probability distribution of size n, then $H(P) \leq \mathbf{log} n$, with equality occurring if and only if P is the uniform distribution U.*

Proof. It follows from Lemma 79 and $H(P) = H(U) - D(P||U)$. ∎

Proof. (Another, even simpler, proof) It is an easy consequence of the convexity of the function $f(x) = -x\mathbf{log}(x)$. ∎

The maximum entropy principle states that when there is not information to determine the whole probability distribution then you should pick the distribution which maximizes the entropy (compatible to the known information, of course). In a sense the maximum entropy principle is an extension of the principle of indifference, or of consistency, since with nothing known both principles assign a uniform distribution. – R.W. Hamming, The Art of Probability, Addison-Wesley, 1991.

Consider two alphabets $Y = \{y_1, \ldots, y_N\}$ and $A = \{a_1, \ldots, a_M\}$. A map ϕ from Y to A^*, the set of all words over A,* is called a prefix code, provided $\phi(x)$ is not a prefix of $\phi(y)$ for any $x \neq y \in Y$. For any $w \in A^*$, $|w|$ stands for the length of w. Let P be a probability distribution on Y, say $P(y_i) = p_i$. The expected length of the prefix code ϕ is $L = \sum_i p_i |\phi(y_i)|$. Let $\ell = \max_{y \in Y} |\phi(y)|$.

Theorem 86 (Shannon) † $L \geq \frac{H(P)}{\log(M)}$.

Proof. For $i \in [N]$, take $n_i = |\phi(y_i)|$ and $q_i = \frac{M^{-n_i}}{\sum_{j=1}^{N} M^{-n_j}}$.

*This form a free monoid under concatenation.

†Intuitively, each digit can carry $\log M$ information and so we need at least $\frac{H(P)}{\log(M)}$ digits to forward a total of $H(P)$ information.

Look at the uniform probability space on A^ℓ. Consider the event E_i consisting of all strings in A^ℓ having $\phi(y_i)$ as a prefix. Since ϕ is a prefix code, we know that E_i are pairwise disjoint events and hence, by virtue of Exercise 10, we get to the so-called Kraft's inequality $\sum_{j=1}^{N} M^{-n_j} \leq 1$, which implies that $q_i \geq M^{-n_i}$. By Lemma 79, we then conclude that $H(P) = -\sum p_i \mathbf{log}(p_i) \leq -\sum p_i \mathbf{log}(q_i) \leq -\sum p_i \mathbf{log}(M^{-n_i}) = L\mathbf{log}(M)$, completing the proof. ∎

He can compress the most words into the smallest ideas of any man I ever met. – Abraham Lincoln

End of Lesson Nine 24/10/06

Lemma 87 (Han's inequality) * *Let $X_1, \ldots, X_n$ be discrete random variables. Then*

$$H(X_1, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Proof. For any $i = 1, \ldots, n$, by the definition of the conditional entropy and the fact that conditioning reduces entropy, $H(X_1, \ldots, X_n) = H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \leq H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i | X_1, \ldots, X_{i-1})$. Summing these n inequalities and using the chain rule for entropy yield $nH(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_1, \ldots, X_n)$, which is exactly what we want. ∎

*T.S. Han, Nonnegative entropy measures of multivariate symmetric correlation, Information and Control, 36 (1978), 133–156.

A Quiz: Can you figure out any generalization of Han's inequality? (One such generalization is Lemma 91.)

Exercise 88 *Suppose n distinct points in R^3 have n_1 distinct projections on the XY-plane, n_2 distinct projections on the XZ-plane and n_3 distinct points on the YZ-plane. Make use of Han's inequality to deduce that $n^2 \leq n_1 n_2 n_3$.*

Example 89 *We have 25 virtually indistinguishable gold coins, 24 of them of the same weight and the remaining one lighter. We have a balance with two pans, but without weights. Accordingly, any measurement will tell us if the loaded pans weigh the same or, if not, which weighs more. How many measurements are needed to find the counterfeit?*

Answer: Let X be the random lighter coin, chosen uniformly among all 25 coins. Let X_i be the outcome of the ith measurement. If t weighings are enough for recovering X, then $X = f(X_1, \ldots, X_t)$. Thus, $\log(25) = H(X) = H(f(X_1, \ldots, X_t)) \leq H(X_1, \ldots, X_t) \leq \sum_i H(X_i) \leq t\log(3)$. Thus, the number of necessary weighings has to be at least $\lceil \frac{\log 25}{\log 3} \rceil = 3$. ∎

Quiz: Design an adaptive algorithm to find the counterfeit in three weighings.

Question: Assume that you have k coins. One of them is a counterfeit which weighs differently from the others, which are real and weigh equally. Isolate the counterfeit by using a balance three times and determine whether it is heavier or lighter.

Answer: n weighings are enough to isolate the counterfeit among k coins if and only if $k \leq (3^n - 1)/2$. See:

`http://www5a.biglobe.ne.jp/~sunomono/twelvecoins.html`

Truth never plays false roles of any kind, which is why people are so surprised when meeting it. Everyone must decide whether he wants the uncompromising truth or a counterfeit version of truth. Real wisdom consists of recommending the truth to yourself at every opportunity. − Vernon Howard

permanent of a $(0,1)$ matrix – number of perfect matchings of a bipartite graph

Example 90 (Brégman) * *Let $G = (A, B; E)$ be a bipartite graph with $|A| = |B| = n$. Then the number of perfect matchings in G is at most $\prod_{v \in A} (d(v)!)^{\frac{1}{d(v)}}$.*

Solution: First, it is obvious that the number of perfect matchings is at most $\prod_{v \in A} d(v)$. But, let us justify it using entropy.

Let Σ be the set of perfect matchings; let σ be a random element of Σ chosen with uniform distribution. Then $H(\Sigma) = \mathbf{log}(|\Sigma|)$. On the other hand, $\sigma \equiv (\sigma(v) : v \in A)$. Fix an ordering $v_1, v_2, \ldots, v_n$ for the vertices of A. We have $\mathbf{log}(|\Sigma|) = H(\sigma) = H(\sigma(v_1)) + H(\sigma(v_2)|\sigma(v_1)) + \cdots + H(\sigma(v_n)|\sigma(v_1), \ldots, \sigma(v_n)) \leq \sum_i H(\sigma(v_i)) \leq \sum_i \mathbf{log}(d(v_i)) = \mathbf{log}(\prod_{v \in A} d(v))$.

*J. Radhakrishnan, An entropy proof of Brégman's theorem, Journal of Combinatorial Theory A 77 (1997), 161–164.

In our entropy argument for getting an upper bound of $|\Sigma|$, we use the inequality $H(\sigma(v_i)|\sigma(v_1),\ldots,\sigma(v_{i-1})) \leq H(\sigma(v_i))$. Can we replace this rough estimate by a tighter one and thus establish a better bound for $|\Sigma|$?

Picking a random permutation $\tau : [n] \to A$, we have $H(\sigma) = H(\sigma(\tau(1))) + H(\sigma(\tau(2))|\sigma(\tau(1))) + \ldots + H(\sigma(\tau(n))|\ \sigma(\tau(1)),\ldots,\sigma(\tau(n)))$. Identifying the $\tau^{-1}(v)$ term of the RHS of this equality as a random variable $f_v(\tau,\sigma)$ for the uniform distributed τ, we see that $H(\sigma) = \sum_{v\in A} E(f_v(\tau,\sigma))$. Clearly, $f_v(\tau,\sigma) \leq \mathbf{log}(g_v(\tau,\sigma))$, where $g_v(\tau,\sigma)$ denotes the number of neighbors of v outside of $\{\sigma(\tau(1)),\ldots,\sigma(\tau(\tau^{-1}(v)-1))\}$. Since τ obeys the uniform distribution, $g_v(\tau,\sigma)$, as a random variable for the joint distribution (τ,σ), takes value $i \in [d(v)]$ with probability $\frac{1}{d(v)}$. This then gives $E(f_v(\tau,\sigma)) \leq \frac{1}{d(v)}\sum_{k=1}^{d(v)} \mathbf{log}(k) = \mathbf{log}((d(v)!)^{\frac{1}{d(v)}})$, and hence we are done. ∎

The entropy method is used to simplify an otherwise complicated counting argument. – B. Chazelle, The Discrepancy Method, Cambridge University Press, 2000.

Not everything that can be counted counts, and not everything that counts can be counted. – Albert Einstein

Chaos in physics corresponds to randomness in mathematics; Randomness in physics may correspond to uncomputability in mathematics. – K. Svozil, Randomness & Undecidability in Physics, World Scientific, Singapore, 1993.

End of Lesson Ten 27/10/06

Lemma 91 (Generalized Subadditivity) *Let $X = (X_1, \ldots, X_n)$ be a random variable and $\mathcal{A} = \{A_i\}_{i \in I}$ be a collection of subsets of $[n]$, such that each element of $[n]$ appears in at least k members of $\mathcal{A}$. For $A \subseteq [n]$, let $X_A = (X_j : \ j \in A)$. Then, $\sum_{i \in I} H(X_{A_i}) \geq kH(X)$.*

Proof. This is an easy consequence of the chain rule (Exercise 78) and the fact that conditioning reduces uncertainty (Lemma 79). ∎

Intuitively, a set cannot be too big if all its projections are small. As a generalization of Exercise 88, we now present a result in exactly this vein, which comes from "F. Chung, P. Frank, R. Graham, J. Shearer, Some intersection theorems for ordered sets and graphs, Journal of Combinatorial Theory Series A 43 (1986), 23–37," the same paper where Lemma 91 was first formulated.

Corollary 92 *Let $\mathcal{F}$ and $\mathcal{A} = \{A_i\}_{i \in I}$ be a collection of subsets of $[n]$, such that each element of $[n]$ appears in at least k members of $\mathcal{A}$. For $i \in I$, put $\mathcal{F}_i = \{f \cap A_i : \ f \in \mathcal{F}\}$. Then, $\prod_{i \in I} |\mathcal{F}_i| \geq |\mathcal{F}|^k$.*

Proof. Suppose $X = (X_1, \ldots, X_n)$ is a random variable taking values in $\{0,1\}^{[n]}$ such that for $x \in \{0,1\}^{[n]}$ we have $P(X = x) = \frac{1}{|\mathcal{F}|}$ if $x = \mathbf{I_F}$ for some $F \in \mathcal{F}$ and $P(X = x) = 0$ otherwise. For $A \in \mathcal{A}$, let X_A be the random variable obtained from X by removing all components not indexed by elements of A. By Lemma 91, we conclude that $\sum_i \mathbf{log}(|\mathcal{F}_i|) \geq \sum_i H(X_{A_i}) \geq kH(X) = k\mathbf{log}(|\mathcal{F}|)$.∎

Exercise 93 *Let $\mathcal{F}$ be a family of subsets of $[n]$ and let p_i be the fraction of set in $\mathcal{F}$ that contain i. Show that $\mathbf{log}(|\mathcal{F}|) \leq \sum H(p_i)$.*[*]

[*]Modify the proof of Corollary 92. Instead of appealing to Lemma 91, use the fact that $H((Z,Y)) \leq H(Z) + H(Y)$.

Exercise 94 * *Make use of Exercise 93 to give another proof of Exercise 84.*

A balanced graph is a graph the difference of whose maximum degree and minimum degree is at most one.

Theorem 95 †*Suppose $\mathcal{F}$ is a family of graphs having the vertex set $[m]$ such that for any $F, F' \in \mathcal{F}$ we can find a triangle as their common subgraph. Then, $|\mathcal{F}| < 2^{\binom{m}{2}-2}$.*

Proof. We identify each graph F with its edge set and hence a subset of $[n]$ for $n = \binom{m}{2}$.

*S. Jukna, Extremal Combinatorics: With Applications in Computer Science, Corollary 23.6, Springer, 2001.

†Compare with Example 72.

Take $\mathcal{A}$ to be the set of graphs whose complements in $[n]$ induce balanced complete bipartite graphs on $[m]$. Clearly, each member of $\mathcal{A}$ has a constant size $a = \binom{\lfloor \frac{m}{2} \rfloor}{2} + \binom{\lceil \frac{m}{2} \rceil}{2}$. Let $A \in \mathcal{A}$ and put $Y_A = \{A \cap F : \ F \in \mathcal{F}\}$. By assumption, the members of the family Y_A are pairwise intersecting. Thus, according to Example 72 (1), $|Y_A| \leq 2^{a-1}$. Observe that each $e \in [n]$ appears in exactly $k = \frac{a}{n}|\mathcal{A}|$ members of $\mathcal{A}$.

By now, Corollary 92 implies that $|\mathcal{F}| \leq \sqrt[k]{(2^{a-1})^{|\mathcal{A}|}} = 2^{n-\frac{n}{a}} < 2^{n-2} = 2^{\binom{m}{2}-2}$. ∎

Student presentation:

ZHANG, Hengli, FKG inequality

WANG, Dakan, Distance one point pairs; Buffon's needle

Whether random bits are truly needed or not is one of the major open problems in complexity theory today. – Bernard Chazelle, The Discrepancy Method: Randomness and Complexity, Cambridge University Press, 2001.

God not only plays dice, but sometimes throws them where they can not be seen. – Stephen Hawking

End of Lesson Eleven 31/10/06

Graph Entropy: A measure of the 'complexity' of the graph

Vertex packing polytope $VP(G)$ of a graph G: The convex hull of all the incidence vectors of stable sets of G in $\mathcal{R}^{V(G)}$.

Let G be a graph on the vertex set $[n]$ and let $P = (p_1, \ldots, p_n)$ be a probability distribution on $V(G)$. The entropy of G with respect to P is defined as $H(G, P) = \min_{a \in VP(G)}(-\sum_{i \in [n]} p_i \mathbf{log}(a_i))$.

Any point $a \in VP(G)$ can be viewed as a random variable taking values in the set of independent sets of G. Suppose that $VP(G)$ has the set of vertices $\{v_1, \ldots, v_m\}$ and $a = \sum t_i v_i$. Then a corresponds to the random variable Y which has probability t_i to take value v_i. We can thus rewrite the definition of $H(G, P)$ as $H(G, P) = \min_Y(-\sum_{i \in [n]} p_i \mathbf{log}(Pr(i \in Y)))$.

By Gibbs' inequality (Lemma 79), we know that $H(P) = H(K_n, P)$, where K_n denotes the complete graph on $[n]$. Thus, graph entropy is a generalization of the Shannon entropy.

Lemma 96 *Let F and G be two graphs on the same vertex set $[n]$ and $F \cup G = ([n], E(F) \cup E(G))$. For any probability distribution P, we have $H(F, P) + H(G, P) \geq H(F \cup G, P)$.*

Proof. Let $a \in VP(F)$ and $b \in VP(G)$ be the vectors achieving the minima for $H(F, P)$ and $H(G, P)$ in our definition of graph entropy. Notice that $a \circ b = (a_1 b_1, \ldots, a_n b_n)$ is in $VP(F \cup G)$. Hence, $H(F, P) + H(G, P) = \sum_i p_i \mathbf{log}(\frac{1}{a_i b_i}) \geq H(F \cup G, P)$. ∎

An immediate corollary of Lemma 96 is $H(G, P) + H(\overline{G}, P) \geq H(P)$.

We have a combinatorial problem that is translated in some appropriate way into a graph covering problem of the following type. Given a graph K and a class of graphs $\mathcal{G}$ where each $G_i \in \mathcal{G}$ has the same vertex set as K, the task is to cover the edge set of K with as few graphs from $\mathcal{G}$ as possible. The question is the minimal number of G_i's needed for the covering. Using the sub-additivity of graph entropy, one can obtain lower bound on this number. Indeed, if the graphs $G_1, \ldots, G_t \in \mathcal{G}$ are such that $\cup_{i=1}^{t} G_i$ covers K, then ... $t \geq \frac{H(K,P)}{\max_{G \in \mathcal{G}} H(G,P)}$. – Gábor Simonyi, Perfect graphs and graph entropy: An updated survey, in: Perfect Graphs (Eds., J.L.R. Alfonsin, B.A. Reed), pp. 293–328, John Wiley, 2001.

J. Körner, K. Marton, New bounds for perfect hashing via information theory, European Journal of Combinatorics, 9 (1988), 523–530.

Other applications follow a somewhat different framework. There the problem given concerns the complexity of some algorithm. As the algorithm proceeds it produces some objects with higher and higher complexity. If these objects can be associated with some graph with an appropriate distribution on its vertex set, then its graph entropy may be used as a measure of complexity of the object corresponding to the graph. At the end the algorithm should produce some specified type of our projects. If the association with graphs is appropriate, this final object will correspond to a graph with high entropy. If we are able to bound from above the possible increase in entropy at each single step of the algorithm and also the entropy of the graph we initially had, then we obtain a lower bound for the number of steps needed. – Gábor Simonyi, Perfect graphs and graph entropy: An updated survey, in: Perfect Graphs (Eds., J.L.R. Alfonsin, B.A. Reed), pp. 293–328, John Wiley, 2001.

J. Radhakrishnan, Better bounds for threshold formulas, Proceedings of the IEEE FOCS (1991), 314–323.

A graph G is strongly splitting if for every probability distribution P we have $H(G,P) + H(\overline{G},P) = H(P)$.

CKLMS*: A graph is strongly splitting if and only if it is perfect.

Let $\widetilde{H}(G,P) = \min I(X,Y)$, where (X,Y) range over pairs of random variables with some joint distribution such that

1) X takes values in $V(G)$ with probability distribution P.

2) Y takes values in the set of independent sets of G.

3) $Pr(X \in Y) = 1$.

Theorem 97 (CKLMS) $H(G,P) = \widetilde{H}(G,P)$.

*C. Csiszár, J. Körner, L. Lovász, K. Marton, G. Simonyi, Entropy splitting for antiblocking corners and perfect graphs, Combinatorica, 10 (1990), 27–40.

Proof. $\widetilde{H}(G,P) \geq H(G,P)$

Assume that $\widetilde{H}(G,P) = I(X,Y)$ for two random variables X, Y satisfying all the above three conditions. Let $p_{i,J} = Pr(X = i, Y = J)$, $p_i = \sum_J p_{i,J}$ and $p_J = \sum_i p_{i,J}$.

Notice that $H(X) = -\sum_i p_i \mathbf{log}(p_i) = -\sum_{i,J} p_{i,J} \mathbf{log}(p_i)$, $H(X|Y) = -\sum_J p_J \sum_i \frac{p_{i,J}}{p_J} \mathbf{log}(\frac{p_{i,J}}{p_J})$, and hence

$$I(X,Y) = H(X) - H(X|Y) = -\sum_{i,J: p_{i,J}>0} p_{i,J} \mathbf{log}(\frac{p_i p_J}{p_{i,J}}). \qquad (8)$$

By now, we get $\widetilde{H}(G,P) = I(X,Y) = -\sum_i p_i \sum_{J:\ p_{i,J}>0} \frac{p_{i,J}}{p_i} \mathbf{log}(\frac{p_i p_J}{p_{i,J}})$ $\geq^{*} -\sum_i p_i \mathbf{log}(\sum_{J:\ p_{i,J}>0} p_J) \geq -\sum_i p_i \mathbf{log}(Pr(i \in Y)) \geq H(G,P)$.

*Jensen's inequality

$\widetilde{H}(G,P) \leq H(G,P)$

Suppose that $H(G,P)$ is attained at $a \in VP(G)$, which corresponds to a random variable Z taking values in the set of stable sets of G. Note that $Pr(i \in Z) = a_i$. Let $Pr(Z = J) = q_J$.

We now construct a pair of random variables (X,Y) fulfilling the three conditions listed on page 0. We only need to specify $p_{i,J} = P(X = i, Y = J)$ such that $\sum_J p_{i,J} = p_i$ and that $i \in J$ whenever $p_{i,J} > 0$ for any vertex i and independent set J. This is achieved by setting

$$p_{i,J} = \begin{cases} \frac{p_i q_J}{a_i}, & \text{if } i \in J, \\ 0, & \text{otherwise.} \end{cases}$$

$\widetilde{H}(G,P) \leq I(X,Y) =^{*} -\sum_{i,J:p_{i,J}>0} p_{i,J}\mathbf{log}(\frac{p_i p_J}{p_i q_J/a_i}) = -\sum_{i,J:p_{i,J}>0} p_{i,J}\mathbf{log}\ (\frac{a_i p_J}{q_J}) = -\sum_{i,J} p_{i,J}\mathbf{log}(a_i) - \sum_{i,J:p_{i,J}>0}\ p_{i,J}\mathbf{log}(\frac{p_J}{q_J}) = -\sum_i p_i\mathbf{log}\ (a_i) + \sum_J p_J\mathbf{log}(\frac{q_J}{p_J}) \leq {}^{\dagger} -\sum_i p_i\mathbf{log}(a_i) + \mathbf{log}(\sum_{J:p_J>0} q_J) \leq -\sum_i p_i\mathbf{log}(a_i) = H(G,P)$. ∎

Exercise 98 *Let $K_n(r)$ be the complete r-uniform hypergraph on n vertices. Let $H_1,\ldots,H_t$ be k-partite r-uniform hypergraphs such that $K_n(r) = \cup H_i$. Then, $t \geq \frac{\binom{n}{r-2}(n-r+2)\mathbf{log}(n-r+2)}{(k-r+2)(\frac{n}{k})^{r-1}\binom{k}{r-2}\mathbf{log}(k-r+2)}$.*

*Eq. (8)

†Jensen's inequality

G. Simonyi, Graph entropy – a survey, In: Combinatorial Optimization, DIMACS Series Discrete Math. Theoret. Comput. Sci. 20 (A.M.S., 1995), 399–441.

Gábor Simonyi, Perfect graphs and graph entropy: An updated survey, in: Perfect Graphs (Eds., J.L.R. Alfonsin, B.A. Reed), pp. 293–328, John Wiley, 2001.

Andrew Chi-Chih Yao, Graph Entropy and Quantum Sorting Problems, STOC 2004, 112–117.

Jeff Kahn, Jeong Han Kim, Entropy and sorting, Journal of Computer and System Sciences 51 (1995), 390–399.

In science one tries to tell people, in such a way as to be understood by everyone, something that no one ever knew before. But in poetry, it's the exact opposite . – Paul Dirac

Mathematics is about finding connections, between specific problems and more general results, and between one concept and another seemingly unrelated concept that really are related.

Poetry finds connections between moon and flowers, spring and autumn, orders and chaos, and happiness and sorrow, and weaves them into a fabric of many splendors.

In the eyes of a mathematician, In the eyes of a poet,

And through their eyes, In our eyes,

The world is a beautiful world, And life a beautiful life.

– LIU, Chung Laung, Chessboards, Hats, and Chinese Poetry: Some Rigorous and Not-So-Rigorous Mathematical Results, public lecture.

End of Lesson Twelve 7/11/06

Recall Exercises 84 and 93.

Suppose now the sample space is $\Omega = \{\omega : \ \omega = (a_1, \ldots, a_n), a_i \in [r]\}$ and the probability is given by $P(\omega) = \prod_i p_i^{v_i(\omega)}$, where $v_i(\omega)$ is the number of occurrences of i in the sequence ω and $(p_1, \ldots, p_r)$ is a probability distribution. This is to toss a coin with r faces independently n times.

Put $C(n, \epsilon) = \{\omega : \ |\frac{v_i(\omega)}{n} - p_i| < \epsilon, i = 1, \ldots, r\}$ and $H = H(p_1, \ldots, p_r)$ $= \sum_{k \in [r]} p_k \ln(\frac{1}{p_k})$. An elementary event (a path) $\omega \in C(n, \epsilon)$ is called a canonical path. Intuitively, almost all paths $\omega \in C(n, \epsilon)$ should be canonical when n goes to infinity. We use the notation $N(C(n, \epsilon))$ for the number of paths in $C(n, \epsilon)$.

Grasp the subject, the words will follow. – Cato the Elder

Theorem 99 (MacMillan) *Let* $p_i > 0$, $i \in [r]$, $\delta > 0$ *and* $0 < \epsilon < 1$. *Then (a)* $P(C(n,\epsilon_1)) = \sum_{\omega \in C(n,\epsilon_1)} p(\omega) \to 1, n \to \infty$; *(b)* $e^{-n(H+\frac{\epsilon}{2})} \le p(\omega) \le e^{-n(H-\frac{\epsilon}{2})}$ *for any* $\omega \in C(n,\epsilon_1)$; *(c)* $N(C(n,\epsilon_1)) \le e^{n(H+\frac{\epsilon}{2})}$; *there exists an* $n_0 = n_0(\epsilon; p_1, \ldots, p_r)$ *such that for all* $n > n_0$, $e^{n(H-\frac{\epsilon}{2}-\delta)} \le N(C(n,\epsilon_1))$, *where* ϵ_1 *is the smaller of* ϵ *and* $\frac{\epsilon}{2\sum \ln \frac{1}{p_k}}$.

Proof. By Exercise 12, to prove (a) we need only bound $1 - P(C_i(n,\epsilon))$, where $C_i(n,\epsilon) = \{\omega : \ |\frac{v_i(\omega)}{n} - p_i| < \epsilon\}$, $i \in [r]$. But $\lim_{n\to\infty} P(C_i(n,\epsilon)) = 1$ follows from the weak law of large numbers (Theorem 59) applied to the random variable

$$\xi_k(\omega) = \begin{cases} 1, & \text{if } a_k = i, \\ 0, & \text{otherwise.} \end{cases}$$

By now, (a) is verified.

For $\omega \in C(n,\epsilon_1)$, it holds $v_k(\omega) > np_k - n\epsilon_1, k \in [r]$. This leads to $p(\omega) = \exp(\sum_k v_k(\omega)\ln(p_k)) < \exp(n\sum p_k \ln(p_k) - n\epsilon_1 \sum \ln(p_k))$ $\leq \exp(-n(H - \frac{\epsilon}{2}))$, where the last inequality is due to $\epsilon_1 = \min(\frac{-\epsilon}{2\sum \ln p_k}, \epsilon)$. Similarly, for $\omega \in C(n,\epsilon_1)$, we have $np_k + n\epsilon_1 > v_k(\omega)$ and then $P(\omega) > \exp(-n(H + \frac{\epsilon}{2}))$ can be deduced. This completes the proof of (b).

It remains to prove (c). Since $P(C(n,\epsilon_1)) \geq N(C(n,\epsilon_1)) \min_{\omega \in C(n,\epsilon_1)} p(\omega)$, we arrive at $N(C(n,\epsilon_1)) \leq \frac{P(C(n,\epsilon_1))}{\min_{\omega \in C(n,\epsilon_1)} p(\omega)} \leq \frac{1}{\exp(-n(H+\frac{\epsilon}{2}))}$, where we make use of (b) in the last step. For the other direction of (c), we appeal to (b) as before and get $N(C(n,\epsilon_1)) \geq \frac{P(C(n,\epsilon_1))}{\max_{\omega \in C(n,\epsilon_1)} p(\omega)} \geq$ $P(C(n,\epsilon_1))\exp(n(H-\frac{\epsilon}{2}))$. Consequently, to obtain $N(C(n,\epsilon_1)) \geq$ $e^{n(H-\frac{\epsilon}{2}-\delta)}$, it suffices to show $P(C(n,\epsilon_1)) \geq \exp(-n\delta)$ holds for sufficiently large n. As $\delta > 0$, this is obviously true in view of (a). ∎

From Chebyshev's inequality (Corollary 38), we know that $P(|X - E(X)| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}$, where $\sigma = \sqrt{\mathbf{var}(X)}$. This provides a polynomial bound (in terms of λ) on the probability that the random variable X takes values in the tails of its distribution. As Serge Bernstein * remarked, this can be greatly improved. We now introduce a class of exponential large deviations estimates, the so-called 'Chernoff-type' bounds.

This estimate is often called Chernoff's inequality in the literature (although Chernoff proved a more general and less handy inequality in 1958 and the above theorem goes back to Bernstein's paper from 1924). – J. Matoušek, J. Vondrák, The Probabilistic Method, Lecture Notes, 2004.

*S. Bernstein, Sur une modification de l'inégalité de Techebychev, Ann. Sc. Instit. Sav. Ukraine, Sect. Math I, 1924.

Theorem 100 * *Let $X_1, X_2, \ldots, X_n$ be independent random variables, each attaining the values 1 and -1, both with probability $\frac{1}{2}$. Let $X = X_1 + \cdots + X_n$. Then, we have, for any $\lambda \geq 0$, $P(X \geq \lambda\sigma) \leq e^{-\frac{\lambda^2}{2}}$ and $P(X \leq -\lambda\sigma) \leq e^{-\frac{\lambda^2}{2}}$, where $\sigma = \sqrt{\mathbf{var}(X)} = \sqrt{n}$.*

Proof. We only prove the first inequality; the second one follows by symmetry. The key idea is to use the Laplace transform technique: Consider the auxiliary random variable $Y = e^{tX}$ where $t > 0$ is a yet undertermined real parameter and note that $P(X \geq \lambda\sigma) = P(Y \geq e^{t\lambda\sigma})$ and apply the Markov's inequality for some special t to get as good an estimate as possible.

*J. Matoušek, J. Vondrák, The Probabilistic Method, Theorem 7.1.1, Lecture Notes, 2004.

Let the machinery start grinding now. By Markov's inequality (Theorem 35), we have $P(X \geq \lambda\sigma) = P(Y \geq e^{t\lambda\sigma}) \leq \frac{E(Y)}{e^{t\lambda\sigma}}$. By Theorem 56 and the independence of X_i, we have $E(Y) = \prod E(e^{tX_i}) = (\frac{e^t+e^{-t}}{2})^n = \cosh(t)^n \leq e^{\frac{nt^2}{2}}$, where the last inequality follows by comparing the Taylor series of $\cosh(t)$ and $e^{\frac{t^2}{2}}$ for all real t. Therefore, we come to $P(Y \geq e^{t\lambda\sigma}) \leq \frac{E(Y)}{e^{t\lambda\sigma}} \leq e^{\frac{nt^2}{2}-t\lambda\sigma}$. The last expression is minimized by setting $t = \frac{\lambda\sigma}{n}$, which yields the value $e^{\frac{-\lambda^2}{2}}$. ∎

Exercise 101 *A football team wins each game with probability $\frac{1}{3}$ independently. Show that among n games the probability that this team wins at least $\frac{n}{2}$ games is less than $\frac{2^{3n/2}}{3^n}$.*

Exercise 102 *Assume that* $X_i, i \in [n]$ *are independent random variables with* $E(X_i) = 0$ *and* $|X_i| \leq 1$. *Set* $X = \sum X_i$ *and let* $\sigma = \sqrt{var(X)}$ *be the standard variation of* X. *Then for* $0 \leq \lambda \leq 2\sigma$, $P(X \geq \lambda\sigma) \leq e^{\frac{-\lambda^2}{4}}$ *and* $P(X \leq \lambda\sigma) \leq e^{\frac{-\lambda^2}{4}}$.

Thought is only a flash in the middle of a long night, but the flash that means everything. – Henri Poincaré

End of Lesson Thirteen 10/11/06

Let $\mathcal{H}$ be a subset of $2^{[n]}$. The discrepancy of $\mathcal{H}$ is $disc(\mathcal{H}) = \min_{f\in\{1,-1\}^{[n]}} \max_{A\in\mathcal{H}} |\sum_{x\in A} f(x)|$. The discrepancy measures how well we can color $[n]$ so that each member of $\mathcal{H}$ contains approximately the same number of red ($+1$) and blue (-1) points.

Theorem 103 (Erdös) *If $|\mathcal{H}| = n$, then $disc(\mathcal{H}) \leq \sqrt{2n\ln(2n)}$.*[*]

Proof. Let $f \in \{1,-1\}^{[n]}$ be a random coloring, the colors of points being chosen uniformly and independently. For any fixed set $A \in \mathcal{H}$, the quantity $f(A) = \sum_{x\in A} f(x)$ is a sum of $|A|$ independent random ± 1 variables. By Theorem 100, $P(|f(A)| > t) < 2e^{-t^2/2|A|} \leq 2e^{-t^2/2n}$. Therefore, defining $disc(f) = \max_{A\in\mathcal{H}} |\sum_{x\in A} f(x)|$, we have $P(disc(f) > t) = P(\cup_{A\in\mathcal{H}}\{|f(A)| > t\}) \leq \sum_{A\in\mathcal{H}} P(|f(A)| > t) < |\mathcal{H}| \cdot 2e^{-t^2/2n}$. Taking $t = \sqrt{2n\ln(2n)}$ yields $P(disc(f) > \sqrt{2n\ln(2n)}) < 1$. This immediately implies $disc(\mathcal{H}) \leq \sqrt{2n\ln(2n)}$.

∎

[*] Gy. Károlyi, Lectures on Extremal Set Systems and Two-Colorings of Hypergraphs.

A tournament is an orientation of a complete graph. Denote by T_n all tournaments with vertex set $[n]$. Given $T \in T_n$ and a permutation σ of $[n]$, define $fit(T, \sigma)$ to be # of arcs ab of T with $\sigma(a) < \sigma(b)$−# of arcs ab of T with $\sigma(a) > \sigma(b)$. Let $fit(T) = \max_\sigma |fit(T, \sigma)|$ and $F(n) = \min_{T \in T_n} fit(T)$.

Exercise 104 *Prove that* $F(n) \leq 2n^{\frac{3}{2}} \ln(n)$.

Recall the definition of a Bernoulli trial and a Poisson trial (Page 0).

Let $X_1, \ldots, X_n$ be independent coin tosses such that $P(X_i = 1) = p_i$ and $P(X_i = 0) = 1 - p_i$, $i \in [n]$. Such coin tosses are referred to as Poisson trials. When all p_i are equal, they are called Bernoulli trials.

Theorem 105 * *Let* $X_1, \ldots, X_n$ *be Poisson trials such that* $P(X_i = 1) = p_i$, *where* $0 < p_i < 1$ *for at least one* i. *Then, for* $X = \sum X_i$ *and any* $\delta > 0$, $P(X - E(X) \geq \delta E(X)) < (\frac{e^\delta}{(1+\delta)^{1+\delta}})^{E(X)}$.

Proof. For any positive real t, $P(X \geq (1+\delta)E(X)) = P(\exp(tX) \geq \exp(t(1+\delta)E(X)))$. Applying the Markov inequality to the right-hand side, we have $P(X \geq (1+\delta)E(X)) < \frac{E(\exp(tX))}{\exp(t(1+\delta)E(X))}$. Note that the inequality is strict, as we assume that p_i are not all identically 0 or 1 so that X is not a constant.

*R. Motwani, P. Raghavan, Randomized Algorithms, Theorem 4.1, Cambridge University Press, 1995.

Since X_i are collectively independent, $E(\exp(tX)) = \prod E(\exp(tX_i))$. So, we find that $P(X \geq (1+\delta)E(X)) < \frac{\prod(p_i e^t + 1 - p_i)}{\exp(t(1+\delta)E(X))}$. Now we use the inequality $1 + x < e^x$ with $x = p_i(e^t - 1)$ to obtain $P(X \geq (1+\delta)E(X)) < \frac{\prod \exp(p_i(e^t-1))}{\exp(t(1+\delta)E(X))} = \frac{\exp((e^t-1)E(X))}{\exp(t(1+\delta)E(X))}$.

We proceed to choose a particular t that gives the best possible bound. For this, we differentiate the last expression with respect to t and set to zero; solving for t now yields $t = \ln(1+\delta)$, which is positive for $\delta > 0$. Substituting this value for t, we finish the proof. ∎

Theorem 106 *Let $X_i, i \in [n]$ be independent $0,1$-valued random variables with $P(X_i = 1) = p_i$. Denote $X = \sum X_i$ and $p = \frac{E(X)}{n}$. Then, for $0 \leq \lambda < n - E(X)$, it holds $P(X \geq E(X) + \lambda) \leq \exp(nH_p(p + \frac{\lambda}{n}))$ and $P(X \leq E(X) - \lambda) \leq \exp(nH_{1-p}(1 - p + \frac{\lambda}{n}))$, where $H_p(x) = x \ln(\frac{p}{x}) + (1 - x) \ln(\frac{1-p}{1-x})$ is the relative entropy † of x with respect to p.*

Fan Chung, Linyuan Lu, Concentration inequalities and martingale inequalities – a survey, Internet Mathematics, to appear.

*H. Chernoff, Asymptotic efficiency for tests based on the sum of observations, Ann. Math. Stat., 23 (1952), 493–507. W. Hoeffding, Probability for sums of bounded random variables, J. of the American Statistical Association, 58 (1963), 13–30.

†Also called Kullback-Leibler divergence. See Page 0.

"There are two facts about the distribution of prime numbers which I hope to convince you so overwhelmingly that they will be permanently engraved in your hearts.

The first is that despite their simple definition and role as the building blocks of the natural numbers, the prime numbers grow like weeds among the natural numbers, seeming to obey no other law than that of chance, and nobody can predict where the next one will sprout.

The second fact is even more astonishing, for it states just the opposite: that the prime numbers exhibit stunning regularity, that there are laws governing their behaviour, and that they obey these laws with almost military precision."

– Don Zagier, Bonn University inaugural lecture.

End of Lesson Fourteen 14/11/06

At the end of a course on calculus, students should be good at algebra; at the end of a course on probability, students should be good at calculus. – T.M. Mills, Problems in Probability, p. 37, World Scientific, 2001.

The Central Limit Theorem

The Moderate Deviations Estimate

The Local Limit Theorem

The Arcsine Law

We do not have time to cover the above four chapters; you are required to read (at least) the statements of the results yourself.

We fix a parameter p between 0 and 1 to be the probability of success and then $1-p$ the probability of failure.

Let $\Omega = \Omega_1^{\mathbb{N}} = \{\omega = (\omega_n)_{n\geq 1} : \ \omega_n = 0 \text{ or } 1 \text{ for all } n \geq 1\}$. Each ω represents a possible outcome of the elementary experiment: the nth coordinate ω_n equals 1 if the outcome of the nth trial is success and 0 if the outcome is failure. We define $S_n(\omega) = \sum_{i\leq n} \omega_i$ to represent the number of successes observed after n trials.

Shift: $\mathscr{S}(\omega) = (\omega_2, \omega_3, \omega_4, \ldots)$.

D. Lind, B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, 1995.

B.P. Kitchens, *Symbolic Dynamics: One-sided, Two-Sided and Countable State Markov Shifts*, Springer, 1998.

A positive integer n is good for $A \subseteq \Omega$ provided there exists a subset $A(n)$ of Ω_n such that $A = \{\omega \in \Omega : \ \omega^{(n)} \in A(n)\}$, where $\omega^{(n)} := \ (\omega_1, \ldots, \omega_n)$. Note that $A = A(n) \times \Omega_1^{\mathbb{N}\setminus[n]} \subseteq \Omega_n \times \Omega_1^{\mathbb{N}\setminus[n]} = \Omega_1^{[n]} \times \Omega_1^{\mathbb{N}\setminus[n]} = \Omega$. By taking $A(n+k) = A(n) \times \Omega_1^k = A(n) \times \Omega_k$ for $k \geq 1$, we see that if n is good, all integers greater than n are good as well for the same $A \subseteq \Omega$.

A subset A of Ω is a finite type event * if there exists a good number n for it. Suppose n is the minimum good number for the finite type event A. By the Theorem of Total Probability (Exercise 16), we know that $P_m(A(m)) = \sum_{\omega^{(m)} \in A(m)} p^{S_m(\omega)} q^{m - S_m(\omega)}$ are all equal for $m \geq n$. This common number is defined to be the probability of the finite type event A and is denoted $P(A)$.

*For any finite set $K \subseteq \mathbb{N}$ and $x \in \Omega_1^K$, the set $x \times \Omega_1^{\mathbb{N}\setminus K}$ is called a cylinder, which is easily seen to be a finite union of several finite type events.

The set $\mathscr{E}$ of finite type events contains Ω and $\emptyset$, and is closed under taking complement and under finite union and intersection. This means that the $\mathscr{E}$ is a Boolean algebra of subsets of Ω. The probability P is a function from the Boolean algebra $\mathscr{E}$ to the interval $[0,1]$ such that $P(\Omega) = 1$, and $P(A \cup B) = P(A) + P(B)$ for every $A, B \in \mathscr{E}$ satisfying $A \cap B = \emptyset$. *

$N \subseteq \Omega$ is a negligible event if for every $\epsilon > 0$ there exists a countable set $\{A_k : \ k \geq 1\}$ of finite type events such that $N \subseteq \cup_{k \geq 1} A_k$ and $\sum_{k \geq 1} P(A_k) < \epsilon$.

$A \subseteq \Omega$ is an almost sure event if its complement A^c is negligible.

If A is an almost sure event, we say that "ω almost surely (a.s.) belongs to A".

*Compare with page 0.

Lemma 107 *Every subset of a negligible event is also negligible. Every countable union of negligible events is negligible. If p is not* 0 *or* 1*, then every countable subset of Ω is negligible.*

We now put $P(X) = 0$ if X is negligible and define $P(\cup_i X_i) = \sum_i P(X_i)$ and $P(\Omega \setminus (\cup_i X_i)) = 1 - \sum_i P(X_i)$, where X_i are pairwise disjoint and are either negligible or finite type events.

Whenever the book says "Lebesgue integral" or "Borel set", it does so for the sake of brevity and means, roughly, "the integral makes sense" and "the set is nice and behaves predictably" – Alexander Barvinok, A Course in Convexity, pp. x–ix, American Mathematical Society, 2002.

For any $A \subseteq \Omega$, put $\mathscr{S}(A) = \{\mathscr{S}(\omega) : \ \omega \in A\}$.

Invariance under shifting: $P(A) = P(\mathscr{S}(A))$

Lemma 108 *Events that are determined by coordinates with disjoint sets of indices are independent.*

Example 109 *Let b be a word constructed from the alphabet $\{0, 1\}$; that is, let b be a finite sequence of 0's and 1's. By Lemma 108, we find that the set of infinite sequences of 0's and 1's not including the word b is negligible.*

Theorem 110 (Borel's Strong Law of Large Numbers) *Almost surely,*

$$\lim_{n \to +\infty} \frac{S_n(\omega)}{n} = p. \tag{9}$$

Borel's Strong Law of Large Numbers (SLLN) is often rephrased as "the sequence $\frac{S_n(\omega)}{n}$, $n = 1, 2, \ldots$, converges to p "almost everywhere (a.e.)", meaning that the sequence $\frac{S_n(\omega)}{n}$, $n = 1, 2, \ldots$, converges to p for those ω outside of a negligible set.

A sequence of random variables $\{X_n\}$ is said to converge "in probability (in pr.)" to a random variable X if and only if for every $\epsilon > 0$ we have $\lim_{n \to +\infty} P(|X_n - X| > \epsilon) = 0$.

Weak Law of Large Numbers Theorem (WLLN), namely Theorem 59, says that "the sequence $\frac{S_n(\omega)}{n}$, $n = 1, 2, \ldots$, converges to p in pr.".

The next theorem tells us SLLN is indeed stronger than WLLN for a fixed p.

Theorem 111 *Convergence a.e. implies convergence in pr..*

Theorem 111 is a corollary of the following, because of the fact that the event $(\omega : |X_n(\omega) - X(\omega)| > \epsilon)$ is a subset of the event $(\omega : |X_m(\omega) - X(\omega)| > \epsilon$ for some $m \geq n)$.

Theorem 112 *The sequence $\{X_n\}$ converges a.e. to X if and only if for every $\epsilon > 0$ we have $\lim_{m \to +\infty} P(|X_n - X| \leq \epsilon$ for all $n \geq m) = 1$; or equivalently, $\lim_{n \to +\infty} P(|X_m - X| > \epsilon$ for some $m \geq n$.*

Proof. We first prove the forward direction. Take $\epsilon > 0$. It suffices to show that $\lim_{m \to \infty} P(|X_m - X| \leq \epsilon) = 1$.

Let $A_m(\epsilon) = \cap_{n=m}^{\infty}(|X_n - X| \leq \epsilon)$. Note that $A_m(\epsilon)$ is increasing with m. The assumption that the sequence $\{X_n\}$ converges a.e. to X means that $\cup_{m=1}^{\infty} A_m(\epsilon)$ is the complement of a negligible set. This gives $\lim_{m \to \infty} P(A_m(\epsilon)) = P(\cup_{m=1}^{\infty} A_m(\epsilon)) = 1$ and hence $1 \geq \lim_{m \to \infty} P(|X_m - X| \leq \epsilon) \geq \lim_{m \to \infty} P(A_m(\epsilon)) = 1$ follows.

We next establish the backward direction. We see above that $\lim_{m \to +\infty} P(|X_n - X| \leq \epsilon$ for all $n \geq m) = 1$ is equivalent to saying that $A(\epsilon) := \cup_{m=1}^{\infty} A_m(\epsilon)$ has probability equal to one.

For any $\omega \in A(\epsilon)$, there is $m(\omega, \epsilon)$ such that

$$n \geq m(\omega, \epsilon) \Rightarrow |X_n(\omega) - X(\omega)| \leq \epsilon. \tag{10}$$

Put $A = \cap_{n=1}^{\infty} A(\frac{1}{n})$. Then $P(A) = \lim_n P(A(\frac{1}{n})) = 1$.

For any $\omega \in A$, Eq. (10) is true for all $\epsilon = \frac{1}{n}$, hence for all $\epsilon > 0$. This means that $X_n(\omega)$ converges to $X(\omega)$ for all ω in A, a set of probability one. ∎

One grain of wheat does not constitute a pile, nor do two grains, nor three and so on. On the other hand, everyone will agree that a hundred million grains of wheat do form a pile. What then is the threshold number? Can we say that 325,647 grains of wheat do not form a pile, but that 325,648 grains do? If it is impossible to fix a threshold number, it will also be impossible to know what is meant by a pile of wheat; the words can have no meaning, although, in certain extreme cases everybody will agree about them. Émile Borel, Probability and Certainty.

Since the law of large numbers is widely misunderstood it is necessary to review some of the misconceptions. First, it does not say that a run of extra large (or small) values will be compensated for later on. ... Second, the law is not a statement that the average will be close to its expected value. ... Third, we have an inequality and not an approximation.

At first glance it is strange that from the repeated independent events we can (and do) find statistical regularity. It is worth serious consideration on your part to understand how this happens; how out of random independent events comes some degree of regularity. – R.W. Hamming, The Art of Probability, p. 87, Addison-Wesley Publishing Company, 1991.

Proof. (of Theorem 110) Let $R_n = \frac{S_n(\omega)}{n} - p$. The sequence $(R_n(\omega))_{n\geq 1}$ fails to approach zero if and only if there is an $m \geq 1$ such that for each $n \geq 1$ there exists a $k \geq n$ satisfying $|R_k(\omega)| > \frac{1}{m}$. In symbols, the set of ω not satisfying Eq. (9) is

$$\cup_{m\geq 1} \cap_{n\geq 1} \cup_{k\geq n} \{\omega \in \Omega : \ |R_k(\omega)| > \frac{1}{m}\}.$$

We want to show that this set is a negligible event. By Lemma 107, it suffices to show that

$$N_m := \ \cap_{n\geq 1} \cup_{k\geq n} \{\omega \in \Omega : \ |R_k(\omega)| > \frac{1}{m}\}$$

is negligible for each $m \geq 1$.

For each $k \geq 1$, let $A_{m,k} := \{\omega \in \Omega : |R_k(\omega)| > \frac{1}{m}\}$. By the large deviation estimate (Theorem 106), there exists a constant $c = c(p,m) > 0$ such that $P(A_{m,k}) \leq e^{-ck}$. Since the series $\sum_{k\geq 1} e^{-ck}$ converges, for every $\epsilon > 0$ there exists an $n \geq 1$ such that $\sum_{k\geq n} P(A_{m,k}) < \epsilon$. Because $N_m \subseteq \cup_{k\geq n} A_{m,k}$, this proves that each N_m is a negligible event. ∎

Corollary 113 *Let $(A_n)_{n\geq 1}$ be a sequence of equiprobable independent random events with probability $P(A)$. The asymptotic empirical probability that these events will occur is almost surely $P(A)$; that is, $\lim_{n\to+\infty} \frac{1}{n}\#\{k \in [n] : \omega \in A_k\} = P(A)$, a.s.*

Proof. Each point ω in the sample space corresponds to a point $(\omega_n) \in \Omega$ satisfying $\omega_n = 1$ if $\omega \in A_n$ and $\omega_n = 0$ otherwise. The result follows from Theorem 110. ∎

The Strong Law of Small Numbers was the title of a paper written by R.K. Guy, in which he asserts that "there are not enough small numbers to satisfy all the demands placed on them." In other words, many things are not true even though they are true for every number that you try, because the first counterexample has 20 digits, or maybe 1020 digits. We can come up with so many different "properties" for numbers to have that we come up with apparent patterns that hold for all the numbers we can conveniently do calculations with, but simply aren't true of all numbers. One example: $\gcd(n^{17}+9,(n+1)^{17}+9)$ seems to always be one. In fact, if you had your computer checking this for $n = 1, 2, 3, \ldots$ successively, it would never find a counter-example. That is because the first counter-example is 8424432925592889329288197322308900672459420460792433

Stolen from http://primes.utm.edu/glossary/page.php/LawOfSmall.html

This law explains why it's always necessary to prove things, rather than assuming them to be true (like that old urban legend about π being normal.)

– `http://www.everything2.com/index.pl?node_id=1306273`

End of Lesson Fifteen 21/11/06

Theorem 114 *Let A be a finite type event. For each integer $n \geq 1$ and each $\omega \in \Omega$, let $S(A, n, \omega)$ be the number of integers $k \in [n]$ such that $\mathscr{S}^{k-1}(\omega) \in A$. Then, $\lim_{n\to\infty} \frac{1}{n} S(A, n, \omega) = P(A)$, almost surely.* *

Proof. There is m and $A' \subseteq \Omega_m$ such that $A = A' \times \Omega_1^{\mathbb{N}\setminus[m]}$. For each $n > 0$ and $j \in [m]$, define $A_{j,n}$ to be $\{\omega : \mathscr{S}^{j+(n-1)m-1}(\omega) \in A\} = \{\omega \in \Omega : (\omega_{j+(n-1)m}, \omega_{j+(n-1)m+1}, \omega_{j+(n-1)m+2}, \ldots, \omega_{j+nm-1}) \in A'\}$. Let $S(A, j, t, \omega)$ be the number of integers $k \in [t]$ such that $\omega \in A_{j,k}$. For fixed j and varying t, the events $A_{j,t}$ are independent and each has probability $P(A)$. Correspondingly, Corollary 113 implies that $\lim_{t\to\infty} \frac{1}{t} S(A, j, t, \omega) = P_m(A') = P(A)$ almost surely. But, we have $\frac{S(A,tm,\omega)}{tm} = \frac{1}{m} \sum_{j=1}^{m} \frac{S(A,j,t,\omega)}{t}$. This then gives $\lim_{t\to\infty} \frac{1}{(t+1)m} S(A, tm, \omega) = \lim_{t\to\infty} \frac{1}{tm} S(A, (t+1)m, \omega) = P(A)$. Finally, we conclude the proof by noting that for every $k \in \{0, 1, \ldots, m\}$ it holds $\frac{1}{(t+1)m} S(A, tm, \omega) \leq \frac{1}{tm+k} S(A, tm + k, \omega) \leq \frac{1}{tm} S(A, (t + 1)m, \omega)$. ∎

*This means that, under the shifting map, the frequency of the trail of the point ω enters the set A converges to $P(A)$ almost surely.

Corollary 115 *In the sequence ω, every word b almost surely occurs with asymptotic frequency equal to its probability.*

What we proved above not only hold for the symbol set $\Omega_1 = \{0, 1\}$ but also hold in a more general setting, namely in case that there are any number of symbols.

We let $|I|$ be the length of the real interval I. A subset E of the real line R is Lebesgue negligible (or of measure zero) if for every $\epsilon > 0$ there exists a countable family $(I_k)_{k\geq 1}$ such that $E \subseteq \cup_{k\geq 1} I_k$ and $\sum_{k\geq 1} |I_k| < \epsilon$. If E is negligible, then we say that almost every real number belongs to the complement of E.

Let b represent an integer no less than 2. Let $\Omega' \subseteq \{0, 1, \ldots, b-1\} := \Omega$ consist of all those $\omega \in \Omega$ with infinitely many $\omega_n < b - 1$. Consider the function Φ from Ω' to the interval $[0, 1)$ defined by $\Phi(\omega) = \sum_{i=1}^{\infty} \omega_i b^{-i}$.

Lemma 116 *The function Φ is a bijection and a set $A \subseteq \Omega'$ is negligible in Ω if and only if $\Phi(A)$ is Lebesgue negligible.*

Proof. Each finite type event is a union of cylinders and for each cylinder A $\Phi(A)$ is an interval of length $P(A)$. ∎

For any real number α, there is a unique expansion in base b of the form

$$\alpha = \lfloor \alpha \rfloor + \sum_{n=1}^{\infty} a_n b^{-n}, \tag{11}$$

where $0 \leq a_n < b$ and $a_n < b-1$ infinitely often. For a fixed number α we write $A(d, b, N)$ to denote the number of occurrences of the integer d in the set $\{a_1, \ldots, a_N\}$ with the a_n given by Eq. (11). We say that a number α is simply normal to base b if $\lim_{N\to\infty} \frac{A(d,b,N)}{N} = \frac{1}{b}$ for every d with $0 \leq d \leq b-1$. We shall call a number entirely normal to base b if it is simply normal to all bases $b^n, n = 1, 2 \ldots$. Finally, we term a real number absolutely normal if it is entirely normal to every base b greater than 1.

Lemma 117 *[*] If α is entirely normal to base b, then each of $\alpha, \alpha b, \alpha b^2, \ldots$ is entirely normal to every base $b, b^2, b^3, \ldots$*

Theorem 118 (Borel) *Almost surely, every number is an absolutely normal number.*

Proof. Corollary 115 and Lemma 116. ∎

[*] Glyn Harman, Theorem 1.2, Metric Number Theory, Clarendon Press, 1998. S.S. Pillai, On normal numbers, Proc. Indian Acad. Sci. Sect. A, 10 (1939), 13–15; ibid., 12 (1940), 179–184

http://turnbull.mcs.st-and.ac.uk/history/Biographies/Borel.html

Probabilities must be regarded as analogous to the measurement of physical magnitudes; that is to say, they can never be known exactly, but only within certain approximation. Émile Borel, Probabilities and Life.

If a monkey hits keys on a typewriter at random for an infinite amount of time, he will almost surely produce the entire collected works of Shakespeare! Even better, he will almost surely do this infinitely often! *

Theorem 119 (Infinite-Monkey) *Consider an infinite-length string produced from a finite alphabet by picking each letter independently at random, uniformly from the alphabet (say the alphabet has n letters). Fix a string S of length m from the same alphabet. Let E_k be the event 'the m-substring starting at position k is the string S'. Then, with probability* 1 *infinitely many of the E_k occur.*

Proof. A result of Theorem 118 (Corollary 115). ∎

Edgardo Ugalde, An alternative construction of normal numbers, Journal de Théorie Nombres, 12 (2000), 165–177.

David H. Bailey, Richard E. Crandall, Random generators and normal numbers, `http://crd.lbl.gov/~dhbailey/dhbpapers/bcnormal.pdf`

*Muhammad Waliji, Monkeys and Walks, an undergraduate essay.

O. Häggström, G. Kalai, E. Mossel, A law of large numbers for weighted majority, Advances in Applied Mathematics 37 (2006), 112–123.

The preceding theorem makes a deep impression (at least on the older generation!) because it interprets a general proposition in probability theory at a most classical and fundamental level. If we use the intuitive language of probability such as coin-tossing, the result sounds almost trite. For it merely says that if an unbiased coin is tossed indefinitely, the limiting frequency of "heads" will be equal to $\frac{1}{2}$ – that is, its a priori probability. A mathematician who is unacquainted with and therefore skeptical of probability theory tends regard the last statement as either "obvious" or "unprovable", but he can scarcely question the authenticity of Borel's theorem about ordinary decimals. As a matter of fact, the proof given above, essentially Borel's own, is a lot easier than a straightforward measure-theoretic version, deprived of the intuitive content [see, e.g., Hardy and Wright, An Introduction to the Theory of Numbers, 3rd. ed., Oxford University Press, Inc., New York, 1954.] – Kai Lai Chung, A Course in Probability Theory, 2nd. ed., Academic Press, 1974.

If you know what you're doing, how long it will take, or what it will cost, it isn't research. – Anonymous

Your theory is crazy, but it's not crazy enough to be true. – Niels Bohr (1885 – 1962)

If a little dreaming is dangerous, the cure for it is not to dream less but to dream more, to dream all the time. – Marcel Proust (1871 – 1922)

The second half of a man's life is made up of nothing but the habits he has acquired during the first half. – Fyodor Dostoevsky (1821–1881)

Never despair, but if you do, work on in despair. – Edmund Burke (1729–1797)

End of Lesson Sixteen 24/11/06

Francesco Cantelli (1875 - 1966, Italian mathematician) extended the work of Borel on SLLR to reach the so-called Borel-Cantelli lemma which is critical in making the law of large numbers more precise. Exercise 41!

http://www-history.mcs.st-andrews.ac.uk/Mathematicians/Cantelli.html

A summary of the proof of Theorem 110: The large deviation estimate implies the convergence of a series of probabilities of events, which in turn implies that a certain event is negligible. The latter implication can be formulated as follows.

Lemma 120 *Let X_n be a sequence of random variables. If $\sum_{n=1}^{\infty} P(X_n > \epsilon)$ converges for all $\epsilon > 0$, then $\lim_{n\to\infty} X_n = 0$ a.s..*

It is time to introduce our tool to verify Lemma 120, the so-called Borel-Cantelli lemma.

F.P. Cantelli, Sulla probabilità come limite della frequenza, Rendiconti d. r., Acad. d. Lincei, 26 (1917), 39–45.

Accademia dei Lincei and Galileo - Father of Modern Science: "He was also made a member of the Accademia dei Lincei (in fact the sixth member) and this was an honour which was especially important to Galileo who signed himself 'Galileo Galilei Linceo' from this time on."

All truths are easy to understand once they are discovered; the point is to discover them. – Galileo Galilei (1564-1642)

What's in a name? That which we call a rose by any other name would smell as sweet. – William Shakespeare (1564 - 1616), Romeo and Juliet

If $(A_n)_{n\geq 1}$ is a sequence of subsets of $\Omega = \Omega_1^{\mathbb{N}}$, we call the event $\cap_{n\geq 1}\cup_{k\geq n} A_k$ A_n infinitely often (i.o.) and denote it by $\limsup_{n\to\infty} A_n$; Note that ω belongs to A_n i.o. if and only if $\omega \in A_n$ for infinitely many indices n and that $\overline{\lim}_{n\to\infty} \mathbf{I}_{A_n} = \mathbf{I}_{A_n \text{ i.o.}}$.

Lemma 121 (Borel-Cantelli Lemma) *Let $(A_n)_{n\geq 1}$ be a sequence of events. If $\sum_{n\geq 1} P(A_n)$ converges, then A_n infinitely often is a negligible event. Conversely, if $\sum_{n=1}^{\infty} P(A_n) = \infty$ and if the A_n's are independent, then $P(A_n \text{ i.o.}) = 1$.*

Proof. The first reading comes from the inequality $0 \leq P(A_n \text{ i.o.}) = P(\limsup_{n\to\infty} A_n) \leq \lim_{k\to\infty} P(\cup_{n\geq k} A_n) \leq \lim_{k\to\infty} \sum_{n=k}^{\infty} P(A_n) = 0$.

To prove the second reading, namely $P(\cap_{n\geq 1} \cup_{k\geq n} A_k) = 1$, it suffices to establish for any n that $P(\cup_{k\geq n} A_k) = 1$, or equivalently $\lim_{N\to\infty} P(\cap_{k=n}^{N} A_k^c) = 0$. This follows from $P(\cap_{k=n}^{N} A_k^c) = \prod_{k=n}^{N}(1 - P(A_k)) \leq \prod_{k=n}^{N} \exp(-P(A_k)) = \exp(-\sum_{k=n}^{N} P(A_k))$. ∎

Check that we are allowed to change the order of taking probability and taking limits in the proof of Borel-Cantelli Lemma.

Borel-Cantelli Lemma tells us that for a sequence of independent events A_n, the event A_n infinitely often is either negligible or almost sure. It is a simple one of various kinds of Zero-One Laws in mathematics!

Proof. (of Lemma 120) Set $A_n = (X_n > \epsilon)$. Lemma 121 implies that for every $\epsilon > 0$ and for $\omega \in \Omega$ almost everywhere there exists an $n_0(\omega, \epsilon) \geq 0$ such that $|X_n(\omega)| \leq \epsilon$ for every $n \geq n_0(\omega, \epsilon)$. Since a countable union of negligible sets is still negligible, we find that for ω almost everywhere it holds that for any positive integer m there is $n_0(\omega, \frac{1}{m})$ such that $X_n(\omega) \leq \frac{1}{m}$ for all $n \geq n_0(\omega, \frac{1}{m})$. This then demonstrates that (X_n) almost surely converges to zero. ∎

Example 122 *Let $(k_n)_{n\geq 1}$ and $(\ell_n)_{n\geq 1}$ be two sequences of positive integers. For each $n \geq 1$, let $A_n \subseteq \Omega$ be the event $\{\omega : \ \omega_{k_n} = \omega_{k_n+1} = \ldots = \omega_{k_n+\ell_n-1} = 1\}$. If $\sum_{n\geq 1} p^{\ell_n}$ converges, then it is almost sure that only finitely many A_n occurs. If the two sequences satisfy $k_n + \ell_n \leq k_{n+1}$ for all n and $\sum_{n\geq 1} p^{\ell_n}$ diverges, then the event A_n i.o. is almost sure.*

Example 123 *Let Φ_n be the function from Ω' to the interval $[0, 1)$ defined by $\Phi(\omega) = \sum_{i=1}^{n} \omega_i 10^{-i}$ (Recall what we define on page 0). Let (ϵ_n) be a sequence of real numbers such that $\sum_n \epsilon_n$ converges. Then it holds almost surely that $|\{n : \ \Phi_n(\omega) < \frac{\epsilon}{10^{n+1}}\}| < \infty$.*

Exercise 124 *1) Put $X(\omega_1, \omega_2) = |n \geq 1 : \ S_n(\omega_1) = S_n(\omega_2) = \frac{n}{2}|$, $(\omega_1, \omega_2) \in S_1^{\mathbb{N}} \times S_1^{\mathbb{N}}$. Then X is a.e. infinite. 2) Put $X(\omega_1, \Omega_2, \omega_3) = |n \geq 1 : \ S_n(\omega_1) = S_n(\omega_2) = S_n(\omega_3) = \frac{n}{2}|$, $(\omega_1, \omega_2, \omega_3) \in S_1^{\mathbb{N}} \times S_1^{\mathbb{N}} \times S_1^{\mathbb{N}}$. Then X is a.e. finite.*

Exercise 125 *Make use of Lemma 121 to give a direct proof of Theorem 119.*

For any mapping $f: \mathbb{N} \to \mathbb{R}$, we say $x \in [0,1]$ is f-good, if there are infinitely many $\frac{p}{q}, p, q \in \mathbb{Z}, p > 0$ such that $|x - \frac{p}{q}| < \frac{1}{qf(q)}$. Here is an application of Borel-Cantelli Lemma in the rational approximations of real numbers.

Theorem 126 *If $\sum_{q=1}^{\infty} \frac{1}{f(q)} < \infty$, then the set of f-good numbers is Lebesgue negligible.*

Proof. Put $B_q = \cup_{0 \le p \le q} (\frac{p}{q} - \frac{1}{qf(q)}, \frac{p}{q} + \frac{1}{qf(q)})$. Then x is f-good if and only if $x \in B_q \cap [0,1]$ for infinitely many q. Let $A_q = \Phi^{-1}(B_q \cap [0,1])$. Then, our task is to prove that $\cap_{N \ge 1} \cup_{k \ge N} (B_k \cap [0,1])$ is Lebesgue negligible; or equivalently, in view of Lemma 116, that $P(A_q \text{ i.o.}) = 0$. However, by Lemmas 116 and 121, this is a result of $\sum_{q=1}^{\infty} P(A_q) \le \sum_{q=1}^{\infty} \frac{2}{f(q)} < \infty$. ∎

Using the Pigeonhole Principle, it is easy to prove Dirichlet's Theorem which says that all real numbers are f-good for the function $f(x) = x$. Theorem 126 means that the result of Dirichlet on rational approximations cannot be improved substantially.

A. Blass, Y. Gurevich, V. Kreinovich, L. Longpré, A variation on the zero-one law, Infor. Process. Lett. 67 (1998), 29–30. available at `http://research.microsoft.com/~gurevich/Opera/132.pdf`

A. Blass, Y. Gurevich, A new zero-one law and strong extension axioms, Bulletin of the EATCS 72 (2000), 103–122.

Theorem 127 (BGKL) *For each decision problem, either almost all sequences of instances are easy, or almost all sequences of instances are hard.*

Proof. Consider a decision problem P on binary strings. Let $\mathscr{A}_p$ be the collection of all polynomial time-bounded algorithms such that $A(x)$ either outputs the correct answer that $x \in P$ or $x \notin P$ or outputs nothing, the latter meaning that A fails on x. Let X be the set of infinite sequences $\overline{x} = \{x_n : \ n \geq 1\}$ where x_n are binary strings of length n. Put $p_n(A) = P(A$ fails on the nth component x_n of $\overline{x})$.

Case 1: There exists $A \in \mathscr{A}_p$ with $\sum_n p_n(A) < \infty$. Use the first part of Lemma 121 to deduce that for almost all sequences $\overline{x}$ the problem P is solvable in polynomial time.

Case 2: For every $A \in \mathscr{A}_p$, we have $\sum_n p_n(A) = \infty$. Use the second part of Lemma 121 to get that for almost all sequences $\overline{x}$ the problem P is not solvable in polynomial time. ∎

Example 128 (Steinhaus)

Exercise 129 **Take a coin with $P(heads) = p$ repeatedly. Let A_k be the event that k or more consecutive heads occurs amongst the tosses numbered $2^k, 2^k + 1, \ldots, 2^{k+1} - 1$. Show that $P(A_k \ i.o.) = 1$ if $p \geq \frac{1}{2}$ and $P(A_k \ i.o.) = 0$ if $p < \frac{1}{2}$.*

Here is a partial converse of Theorem 111.

Exercise 130 *Suppose a sequence of random variables X_n converge to X in pr.. Then there is a subsequence $(X_{n_k})_{k=1}^{\infty}$ which converge to X a.e..*

K. Sutner, The Ehrenfeucht-Mycielski sequence, `http://www.cs.cmu.edu/~sutner/papers.html`

A. Ehrenfeucht, J. Mycielski, A pseudorandom sequence – how random is it? American Mathematical Monthly 99 (1992), 373–375.

T.R. McConnell, Laws of Large Numbers for some non-repetitive sequences.

*J. Jacod, P. Protter, Probability Essentials, Exercise 10.16, Springer, 2004.

Anyone who says that they can contemplate quantum mechanics without becoming dizzy has not understood the concept in the least. – Niels Bohr (1885-1962)

One had to be a Newton to notice that the moon is falling, when everyone sees that it doesn't fall. – Paul Valéry (1871-1945)

All of physics is either impossible or trivial. It is impossible until you understand it, and then it becomes trivial. – Ernest Rutherford (1871-1937)

End of Lesson Seventeen 28/11/06

As summarized on page 0, there are two key ingredients of the proof of Theorem 110. The first is the convergence of a series of probabilities of events, which we obtain by utilizing Chernoff's bound. We show that Markov's inequality, though much easier, can sometimes play the same role, if not stronger.

Corollary 131 *Let $(X_n)_{n\geq 0}$ be a sequence of random variables. If $\sum_{n=0}^{+\infty} E(|X_n|)$ converges, the the sequence (X_n) almost surely converges to zero.*

Proof. By Markov's inequality (Theorem 35), for any $\epsilon \geq 0$, we have $P(|X_n| > \epsilon) \leq P(|X_n| \geq \epsilon) \leq \frac{E(|X_n|)}{\epsilon}$. Consequently, Lemma 120 gives the result. ∎

Corollary 131 enables us to get a generalization of Theorem 110. Note that in the proof of Theorem 132, we repeatedly use the method of higher moments and Corollary 131.

Theorem 132 *Let $(X_i)_{i\geq 1}$ be a sequence of pairwise independent random variables such that $E(X_i) = 0$ for each i and that $M = \sup_{i\geq 1} E(X_i^2)$ is finite. Let $R_m = \sum_{i=1}^{m} X_i$. Then $\lim_{m\to+\infty} \frac{R_m}{m} = 0$ almost surely, namely $\frac{R_m}{m}$ converges to 0 a.s..*

Proof. Since X_i are pairwise independent, we get *

$$E((\frac{R_n}{n})^2) = \frac{\sum_{i=1}^{n} E(X_i^2)}{n^2} \leq \frac{M}{n}. \tag{12}$$

It is direct from Eq. (12) that $\sum_{n\geq 1} E((\frac{R_{n^2}}{n^2})^2) \leq M \sum_{n\geq 1} \frac{1}{n^2} < \infty$. Then, by means of Corollary 131, we arrive at $\lim_{n\to+\infty}(\frac{R_{n^2}}{n^2})^2 = 0$ a.s., which clearly implies that $\lim_{m\to+\infty} \frac{R_{\lfloor\sqrt{m}\rfloor^2}}{m} = \lim_{n\to+\infty} \frac{R_{n^2}}{n^2} = 0$ a.s.. Consequently, to finish the proof, it suffices to verify that $\lim_{m\to+\infty}(\frac{R_{\lfloor\sqrt{m}\rfloor^2}}{m} - \frac{R_m}{m}) = 0$, a.s., or equivalently, $\lim_{m\to+\infty}(\frac{R_{\lfloor\sqrt{m}\rfloor^2}}{m} - \frac{R_m}{m})^2 = 0$, a.s.. But, this, in virtue of Corollary 131 again, simply follows from $E((\frac{R_{\lfloor\sqrt{m}\rfloor^2}}{m} - \frac{R_m}{m})^2) = O(m^{-\frac{3}{2}})$, which in turn is a consequence of $(\frac{R_{\lfloor\sqrt{m}\rfloor^2}}{m} - \frac{R_m}{m})^2 = \frac{1}{m^2} E((\sum_{t=\lfloor\sqrt{m}\rfloor^2+1}^{m} X_i)^2) \leq \frac{M}{m^2}(m - \lfloor\sqrt{m}\rfloor^2)$. ∎

*By Markov's inequality (Theorem 35), for any $\epsilon \geq 0$, we have $P(|\frac{R_n}{n}| \geq \epsilon) \leq \frac{E((\frac{R_n}{n})^2)}{\epsilon^2}$. Therefore, Eq. (12) already implies that $\frac{R_n}{n}$ converge to 0 in pr., which, according to Theorem 111, is weaker than our assertion in the theorem.

We present a strengthened version of the divergence part of the Borel-Cantelli Lemma (Lemma 121). The method of proof deserves more attention than the result.

Lemma 133 * *Let* $(A_n)_{n=1}^{\infty}$ *be a series of pairwise independent events. If* $\sum_{n=1}^{\infty} P(A_n) = \infty$, *then* $P(A_n \text{ i.o.}) = 1$.

Proof. We use the shorthand I_n for $\mathbf{I}_{A_n}$, p_n for $E(I_n)$, and J_n for $\sum_{k=1}^{n} I_k$. It turns out that our present hypothesis is

$$\forall m \neq n, E(I_m I_n) = p_m p_n \tag{13}$$

and

$$\sum_{n=1}^{\infty} p_n = +\infty; \tag{14}$$

while our conclusion is

$$\lim_{n\to\infty} J_n = \infty, a.e.. \tag{15}$$

*K.L. Chung, A Course in Probability Theory, Theorem 4.2.5, Academic Press, 1974.

By Theorem 58, $\mathbf{var}(J_n) = \sum_{k=1}^{n} \mathbf{var}(I_k) = \sum_{k=1}^{n}(p_k - p_k^2)$.

Eq. (14) tells us that $E(J_n) = \sum_{k=1}^{n} p_k \to \infty$. This guarantees that $\sqrt{\mathbf{var}(J_n)} \leq E(J_n)^{\frac{1}{2}} = o(E(J_n))$.

Consequently, $\lim_{n\to\infty} P(J_n > \frac{E(J_n)}{2}) \geq \lim_{n\to\infty} P(|J_n - E(J_n)| \leq \beta\sqrt{\mathbf{var}(J_n)})$, which, according to Chebyshev's inequality (Corollary 38), gives $\lim_{n\to\infty} P(J_n > \frac{E(J_n)}{2}) \geq 1 - \frac{\mathbf{var}(J_n)}{\beta^2\mathbf{var}(J_n)} = 1 - \frac{1}{\beta^2}$.

Since $E(J_n) \to \infty$, we know that it holds for any $\beta > 0$ that $P(\lim_{n\to\infty} J_n = \infty) \geq 1 - \frac{1}{\beta^2}$. This means that $\lim_{n\to\infty} J_n = \infty$, a.e., as required. ∎

Exercise 134 **Let $(A_n)_{n=1}^{\infty}$ be a series of events for which $\sum_{n=1}^{\infty} P(A_n) = \infty$, and $\liminf_{n\to\infty} \frac{\sum_{k=1}^{n}\sum_{i=1}^{n} P(A_kA_j)}{(\sum_{k=1}^{n} P(A_k))^2} \leq C$, where $C > 1$ is a constant. Then $P(A_n \text{ i.o.}) \geq C^{-1}$.*

*F. Spitzer, Principles of Random Walk, Van Nostrand, Princeton, N.J. 1964.

To abstract is presumably to come down to essentials. It is to free oneself from accidental features and to focus one's attention on the crucial ones. Abstractly, the theory of 'heads or tails' ('fair' coin, independent tosses) is simply the study of ...

$t = \frac{\epsilon_1}{2} + \frac{\epsilon_2}{2^2} + \cdots$

$r_k = 1 - 2\epsilon_k$

$1 - 2t = \sum_{k \geq 1} \frac{r_k}{2^k}$

$\int_0^1 \prod_{k=1}^{\infty} \exp(ix\frac{r_k(t)}{2^k})dt = \prod_{k=1}^{\infty} \int_0^1 \exp(ix\frac{r_k(t)}{2^k})dt \Leftrightarrow \frac{\sin x}{x} = \prod_{n=1}^{\infty} \cos\frac{x}{2^k}$

Mark Kac, Statistical Independence in Probability, Analysis and Number Theory, Mathematical Association of America, 1959.

Gerald S. Goodman, Statistical independence and normal numbers: An aftermath to Mark Kac's Carus Monograph, American Mathematical Monthly, 106, (1999), 112–126.

I was always interested in problems rather than in theories. In retrospect the thing which I am happiest about, and it was done in cooperation with Erdös was the introduction of probabilistic methods in number theory. To put it poetically, primes play a game of chance. And also some of the work in mathematical physics. I am amused by things. Can one hear the shape of a drum? I also have a certain component of journalism in me, you see: I like a good headline, and why not? And I am pleased with the sort of thing I did in trying to understand a little bit deeper the theory of phase transitions. I am fascinated, also, with mathematical problems, and particularly the role of dimensionality: why certain things happen in 'from three dimensions on' and some others don't. I always feel that that is where the interface, will you pardon the expression, of nature and mathematics is deepest. To know why only certain things observed in nature can happen in the space of a certain dimensionality. Whatever helps understand this riddle is significant, I am pleased that I, in a small way, did something with it. − Mark Kac

Although the prime numbers are rigidly determined, they somehow feel like experimental data. − T. Gowers, Mathematics: A Very Short Introduction (Oxford Univ. Press, 2002), p.121.

It is evident that the primes are randomly distributed but, unfortunately, we don't know what 'random' means. − R.C. Vaughan (February 1990)

End of Lesson Eighteen 5/12/06

The simplest definition of random walk is an analytical one. It has nothing to do with probability theory, except insofar as probabilistic ideas motivate the definition. In other words, probability theory will "lurk in the background" from the very beginning. Nevertheless there is a certain challenge in seeing how far one can go without introducing the formal (and formidable) apparatus of measure theory which constitutes the mathematical language of probability theory. Thus we shall introduce measure theory only when confronted by problems sufficiently complicated that they would sound contrived if expressed as purely analytic problems, i.e., as problems concerning the transition function which we are about to define. – Frank Spitzer, Principles of Random Walks, Second Edition, Springer, 1976.

Random walk in random environment is one of the basic model of the field of disordered system of particles. In this model, an environment is a collection of transition probabilities ... – F. Rassoul-Agha, On the zero-one law and the law of large numbers for a random walk in a mixing random environment, Electron. Comm. in Probab., 10 (2005), 36–44.

`http://www.math.utah.edu/~firas/Research/`

A random walk is a mathematical model. As is so often the case in applied mathematics, a single mathematical model can be used in vastly different contexts – this is the power of mathematics. The random walk has been used to describe fluctutations on financial markets, random motion of particles subject to Brownian motion and the growth of certain populations: An Australian example of the application of random walks to financial markets can be found in Praetz. From a pure mathematical point of view, the subject of random walks is a beautiful theory with many problems which are still unsolved. – T.M. Mills, Problems in Probability, World Scientific, 2001.

The term random walk suggests stochastic motion in space, a succession of random steps combined in some way. .. We require the steps to be independent and to have the same probability distribution. The walk is then a succession of products of those steps. Later on we apply our results to slightly more general situations, e.g., cases where steps depend on each other in a Markovian way. Thus our study of random walk is synonymous with the study of products of independent identically distributed random elements of a semigroup. – Göran Högnäs, Arunava Mukherjea, Probability measures on semigroups: Convolution products, random walks, and random matrices, Plenum Press, 1995.

A random walk on $\mathbb{Z}^n$ composed of a finite number of possible steps can be described as follows: Let $(x_1, \ldots, x_d)$ be a finite family of elements of $\mathbb{Z}^n$ and $p_1, \ldots, p_d$ a family of real positive numbers that sum to one. Each x_i represents an allowed step and p_i the probability of taking the step represented by x_i. Consider the probability model that parallels what is described in Lecture Fifteen: The probability space is all sequences $(Y_i)_{i \geq 1}$, $Y_i \in \{x_1, \ldots, x_d\}$ and the probability is defined by specifying all probabilities of those cylinder sets in the most obvious way.

If a drunkard stands at the origin at time zero, then each random element in the above random walk model does specify a drunkard's walk $(M_i)_{i \geq 1}$, where $M_i = \sum_{j \leq i} Y_j$. We often identify $(Y_i)_{i \geq 1}$ with $(M_i)_{i \geq 1}$ and directly say that the later sequences with corresponding probability distribution form a random walk.

A drunkard's walk is recurrent if this walk hits the origin infinitely many times and is transient otherwise.

A random walk is recurrent if almost surely every drunkard's walk is recurrent and is transient if almost surely every drunkard's walk is transient in the above mathematical model.

The equivalence between (i) and (ii) in Theorem 135 is again an all-or-nothing principle, namely a Zero-One Law. It says that either all drunkards returns to the origin infinitely often or all drunkards return to the origin only finitely many times, with a negligible set of possible exceptions.

Theorem 135 *For every random walk $(M_n)_{n\geq 1}$, the following four statements are equivalent:*

(i) *The random walk is recurrent.*

(ii) *The random walk is not transient.*

(iii) $\sum_{n=1}^{+\infty} P(M_n = 0) = +\infty$.

(iv) $\lim_{n\to+\infty} P(\textit{there exists } k \leq n \textit{ such that } M_k = 0) = 1$.

Proof. (i)⇒(ii): This is trivial.

(ii)⇒(iii): If $\sum_{n=1}^{+\infty} P(M_n = 0)$ converges, then by Borel-Cantelli Lemma (Lemma 121), $M_n = 0$, i.o., is a negligible set. But a drunkard's walk (M_n) is recurrent if and only if $M_n = 0$ happens infinitely often. This then shows that the random walk is transient. ∎

To finish the proof of Theorem 135, we need a lemma.

If m, s, t are positive integers with $s \leq t$, let $A^m_{s,t}$ be the event consisting of those drunkard's walks that return to the starting point at least m times between steps s and t. For a drunkard's walk ω, let $t_1(\omega), \ldots, t_k(\omega)$ be the first k positive numbers at which time the drunkard returns to the origin.

Lemma 136 *For every $m, t > 0$, $P(A^m_{1,t}) \leq (P(A^1_{1,t}))^m \leq P(A^m_{1,mt})$.*

Proof. $P(A^m_{1,t}) = \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq t} P(\omega : t_1(\omega) = i_1, \ldots, t_m(\omega) = i_m) = \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq t}$
$[P(t_1(\omega) = i_1) P(t_1(\omega) = i_2 - i_1) \cdots P(t_1(\omega) = i_m - i_{m-1})] \leq (P(A^1_{1,t}))^m$.

$P(A^m_{1,mt}) = \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq mt} P(\omega : t_1(\omega) = i_1, \ldots, t_m(\omega) = i_m) \geq \sum_{1 \leq i_1 \leq t, 1 \leq i_2 - i_1 \leq t, \cdots, 1 \leq i_m - i_{m-}$
$[P(t_1(\omega) = i_1) P(t_1(\omega) = i_2 - i_1) \cdots P(t_1(\omega) = i_m - i_{m-1})] = (P(A^1_{1,t}))^m$. ∎

Proof. (of Theorem 135 continued)

(iii)⇒(iv): Set $\rho = \lim_{n\to+\infty} P(A^1_{1,n})$. To prove (iv) is to deduce $\rho = 1$.

By linearity of expectation, $\sum_{k=1}^n P(M_k = 0) = E(\#\{j : \ j \in [n], M_j = 0\}) = \sum_{j=1}^n jP((M_k)_{k\in[n]} \text{ include } j \text{ zeros }) = \sum_{j=1}^n j(P(A^j_{1,n}) - P(A^{j+1}_{1,n})) = \sum_{j=1}^n P(A^j_{1,n})$. The first inequality of Lemma 136 now enables us write $\sum_{k=1}^n P(M_k = 0) \leq \sum_{j=1}^n (P(A^1_{1,n}))^j \leq \sum_{j=1}^n \rho^j$. On account of (iii), namely $\sum_{n=1}^{+\infty} P(M_n = 0) = +\infty$, this gives $\sum_{j=1}^{\infty} \rho^j = \infty$, which proves that $\rho = 1$, as desired.

(iv)⇒(i): We have $\rho = 1$ according to (iv). Let m be a positive integer and let $A^m_{1,\infty}$ be the set of ω such that the sequence $(M_n(\omega))_{n\geq 1}$ contains at least m zeros. For each $t > 0$, $A^m_{1,\infty} \supseteq A^m_{1,mt}$. Thus, by the second inequality of Lemma 136, $P(A^m_{1,\infty}) \geq \lim_{t\to\infty} P(A^m_{1,mt}) \geq \lim_{t\to\infty} (P(A^1_{1,t}))^m = \rho^m = 1$ and hence $(A^m_{1,\infty})^c$ is negligible. Note that a drunkard's walk ω is transient if any only if $\omega \in \cup_{m\geq 1}(A^m_{1,\infty})^c$ while the union of countably negligible sets is still negligible. We then arrive at the conclusion that the random walk is recurrent. ∎

Let $F_n = P(M_n = 0, M_i \neq 0, \forall 0 < i < n)$. Clearly, $\sum_{n=1}^{\infty} F_n \leq 1$.

Corollary 137 *Put $F = \sum_{n=1}^{\infty} F_n$. The random walk is recurrent if and only if $F = 1$.*

Proof. By Theorem 135, the result follows from the fact that $\sum_{n=1}^{+\infty} P(M_n = 0) = \sum_{n=1}^{+\infty} \prod_{(i_1,\ldots,i_t)\vdash[n]} F_i = \sum_{t=1}^{\infty} (\sum_{i=1}^{\infty} F_i)^t = \frac{F}{1-F}$. ∎

I remember meeting a young Frenchman years ago, and he had been trying to do research for several years. He asked me, "How do you do research? How do you start on a problem?" I said, "Well, sometimes it happened to me that I read a paper and I didn't like the proof. So I started to think about something that might be more natural, and very often this led to some new work." Then I asked him, "What about your case?" He said, "I never found a proof I didn't like." I thought, "This is hopeless!" – Louis Nirenberg

Try to learn something about everything and everything about something. − Thomas Henry Huxley (1825-1895)

Copy from one, it's plagiarism; copy from two, it's research. − Wilson Mizner (1876-1933)

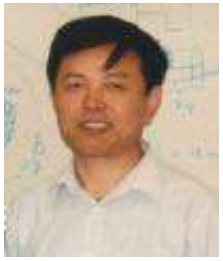
The first examinee is saying: Sir, I did not have time enough to study everything but I have learned very carefully the first chapter of your handout.

Very good − says the professor − you will be a great specialist. You know what a specialist is. A specialist knows more and more about less and less. Finally he knows everything about nothing.

The second examinee is saying: Sir, I did not have time enough but I read your handout without taking care of the details. Very good − answers the professor − you will be a great polymath. You know what a polymath is. A polymath knows less and less about more and more. Finally he knows nothing about everything.

− Pál Révész, Random Walk in Random and Non-Random Environments, Second Edition, World Scientific, 2005.

End of Lesson Nineteen 8/12/06



Two guest talks taking place in the mathematics building: 16:00 - 16: 50, Andreas Dress (CAS-MPG Partner Institute for Computational Biology), A New Approach towards Detecting Community Structure in Networks; 17:00 - 17: 50, Ding-Zhu Du (Department of Computer Science and Engineering, University of Texas at Dallas), Non-Unique Probe Selection.

End of Lesson Twenty 12/12/06

We will talk more about random walks, especially its connection with electrical network. We will mostly follow:

Russell Lyons, Yuval Peres, Probability on Trees and Networks, A book in progress, available at `http://www.stat.berkeley.edu/~peres/`.

Probability theory, like much of mathematics, is indebted to physics as a source of problems and intuition for solving these problems. Unfortunately, the level of abstraction of current mathematics often makes it difficult for anyone but an expert to appreciate this fact. – Peter G. Doyle, J. Laurie Snell, Random Walks and Electric Networks, Carus Math. Monogr., vol. 22, Mathematical Assoc. of America, Washington, 1984.

The history of random walk goes back to two classical scientific recognition. In 1827 Robert Brown, the English botanist published his observation about the irregular movement of small pollen grains in a liquid under his microscope. He not only described the irregular movement but also pointed out that it was caused by some inanimate property of Nature. The irregular and odd series produced by gambling, e.g., while tossing a coin or throwing a dice raised the interest of the mathematicians Pascal, Fermat and Bernoulli as early as in the mid-16^{th} century. – András Telcs, The Art of Random Walk, Springer, 2006.

Some other good references:

D. Aldous, P. Diaconis, Shuffling Cards and Stopping Times, Amer. Math. Monthly 93 (1986), 333–348

Béla Bollobás, Modern Graph Theory, Springer, 1998.

N.L. Biggs, Algebraic potential theory on graphs, Bull. London Math. Soc. 29 (1997), 641–682.

Kai Lai Chung, Green, Brown, and Probability & Brown Motion on the Line, World Scientific, 2002.

E.B. Curtis, J.A. Morrow, Inverse Problems for Electrical Networks, World Scientific, 2000.

Peter G. Doyle, J. Laurie Snell, Random Walks and Electric Networks, available at `http://arxiv.org/abs/math.PR/0001057`.

David A. Levin, Yuval Peres, Elizabeth L. Wilmer, Markov Chains and Mixing Times, available at `http://www.oberlin.edu/markov/`.

L. Lovász, Random walks on graphs: A survey, available at `http://research.microsoft.com/users/lovasz/erdos.ps`.

L. Lovász, Eigenvalues and Geometric Representations of Graphs, lecture notes, `http://research.microsoft.com/users/lovasz/course.htm`.

A locally finite weighted graph is a graph together with a map $C: E(G) \to \mathbb{R}_+$ such that for all $v \in V(G)$ we have $D(v) = \sum_{\substack{e \in E(G) \\ e \sim v}} C(e) \in (0, \infty)$. For each $e \in E(G)$, we can think of $C(e)$ as the conductance of the edge e and thus identify (G, C) with an electrical network.

For any $x, y \in V(G)$, define the transition function $P_{xy} = \frac{\sum_{e \sim x, e \sim y} C(e)}{D(x)}$. A random walk on the weighted graph (G, w) is a sequence of random variables $X_0, X_1, X_2, \ldots$ each taking values in $V(G)$ and satisfy $P(X_{t+1} = x_{t+1} | \ X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = P_{x_i x_{i+1}}$. The special random walk we studied before corresponds to the case that G is the lattice in an Euclidean space, namely the nearest-neighbor graph of $\mathbb{Z}^n$, and C takes constant value.

When the graph G is finite, the study of the random walk on G under the transition rule P is the main topic of finite Markov chains, for which we refer to:

O. Häggström, Finite Markov Chains and Algorithmic Applications, London Mathematical Society Student Texts, 52, Cambridge University Press, Cambridge, 2002.

Exercise 138 *Show that $\sqrt{D} P \sqrt{D}^{-1}$ is a symmetric matrix.*

Exercise 139 *Show that* $D(x_1)\prod_{i=1}^{n-1} P_{x_i x_{i+1}} = D(x_n)\prod_{i=1}^{n-1} P_{x_{n+1-i}x_{n-i}}$. *This means that the chain is equally likely to make a tour through the states in forwards as in backwards order.*

For any $f \in \mathbb{R}^{V(G)}$, define $d(f) \in \mathbb{R}^{E(G)}$ * to be the antisymmetric function satisfying $d(f)(xy) = \sqrt{C(xy)}(f(x) - f(y))$, while for any antisymmetric function $g \in \mathbb{R}^{E(G)}$, define $d^*(g)(x) = \sum_{y\sim x} \sqrt{C(xy)}g(xy) = \sum_{\substack{e\in E(G)\\ x=e^-}} \sqrt{C(e)}g(e)$. Note that d and d^* are adjoints of each other and are called the coboundary operator and the boundary operator of the electrical network, respectively.

Let $L = d^*d$ be the vertex Laplacian on (G, C) and let $\Delta = -L$. Clearly, $\Delta = C - D$, namely $\Delta(f)(x) = \sum_{x\sim y} C(xy)f(y) - D(x)f(x) = \sum_{x\sim y} C(xy)(f(y) - f(x))$. Expressing in matrix language, this is simply $\Delta f = (C - D)f$.

*We assign a direction to each edge $e \in E(G)$, say from x to y, and denote this oriented edge by $a = xy$ and use the notation $a^- = x$, $a^+ = y$. We think of $R^{E(G)}$ as a space spanned by all these oriented edges and identify the oriented edge yx with -1 times xy, namely $-xy$. Correspondingly, a linear function f on $E(G)$ is given by a function on all oriented edges, which by its linearity, must satisfy $f(xy) = f(-yx) = -f(yx)$. Sometimes, we can fix an orientation of the graph and denote the set of such oriented edges by $A(G)$ and use $\mathbb{R}^{A(G)}$ instead of $\mathbb{R}^{E(G)}$.

Let B be the incidence matrix of the oriented graph $G = (V(G), A(G))^*$, namely $B = B^- - B^+$, where

$$B^+(a,v) = \begin{cases} 1, & \text{if } v = a^+, \\ 0, & \text{otherwise,} \end{cases} \quad B^-(a,v) = \begin{cases} 1, & \text{if } v = a^-, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\sqrt{C}$ designate the diagonal matrix whose (a,a)-entry is $\sqrt{C(a)}$. Observe that

$$df = \sqrt{C}Bf, \forall f \in \mathbb{R}^{V(G)}. \tag{16}$$

Let i stand for the current flowing in $A(G)$, v the potential on $V(G)$, and $j = \sqrt{C}^{-1}i$. Here are mathematical interpretations of two facts of physics.

Lemma 140 (Ohm's Law) $dv = \sqrt{C}^{-1}i = j$.

Lemma 141 (Kirchhoff's Node (Current) Law) $d^*j(x) = B^\top i = 0$ *if no battery is connected at* x.

*Remember that the orientation is nothing but to fix a basis of the edge space of the undirected graph G.

A function f on $V(G)$ is harmonic at x if $\Delta(f)(x) = 0$. In other words,

$$0 = \frac{\Delta(f)(x)}{D(x)} = \sum_{x \sim y} P_{xy}(f(y) - f(x)) = \sum_{x \sim y} P_{xy} f(y) - f(x), \qquad (17)$$

saying that $f(x)$ is the weighted average of the values of f at the neighbors of x.

Lemma 142 *The potential function v of an electrical network is harmonic at any node which is not connected to battery.*

Proof.
$$\begin{aligned} \Delta v(x) &= -d^* dv(x) && \text{by definition} \\ &= -d^* j(x) && \text{by Ohm's Law} \\ &= 0. && \text{by Kirchhoff's Node Law} \end{aligned}$$
∎

A **harmonic function** *is a function that satisfies the* **Laplace equation**

$$\nabla^2 u = 0.$$

The problem of finding a harmonic function subject to its boundary values is called the **Dirichlet problem**. *The harmonic function that satisfies the boundary conditions minimizes the Dirichlet integral since the Laplace equation is the Euler-Lagrange equation for the Dirichlet integral. – Leo Grady, Random walks for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (2006), 1768–1783.*

For a graph G and $W \subseteq V(G)$, a boundary point of W is a point of W which is adjacent to some point outside of W. The set of all boundary points of W is designated as ∂W.

Lemma 143 (Maximum Modulus Principle) *Let (G, C) be a connected electrical network and $w \in W \subsetneq V(G)$. Suppose $f \in \mathbb{R}^{V(G)}$ is harmonic on W and $f(w) = \max\{f(v) : v \in W\}$ ($f(w) = \min\{f(v) : v \in W\}$). Then there is $w' \in \partial(V(G) \setminus W)$ such that $f(w') \geq f(w)$ ($f(w') \leq f(w)$).*

Lemma 144 (Uniqueness Principle) *Let (G, C) be a connected electrical network and W a finite proper subset of $V(G)$. If $f, g \in \mathbb{R}^{V(G)}$ are both harmonic on W and $f \mid_{\partial W} = g \mid_{\partial W}$, then $f \mid_W = g \mid_W$.*

Corollary 145 (Superposition Principle) *Let (G, C) be a connected electrical network and W a finite proper subset of $V(G)$. If $f, g, h \in \mathbb{R}^{V(G)}$ are all harmonic on W and if $f = ag + bh$ holds on $V(G) \setminus W$, then $f = ag + bh$ everywhere.*

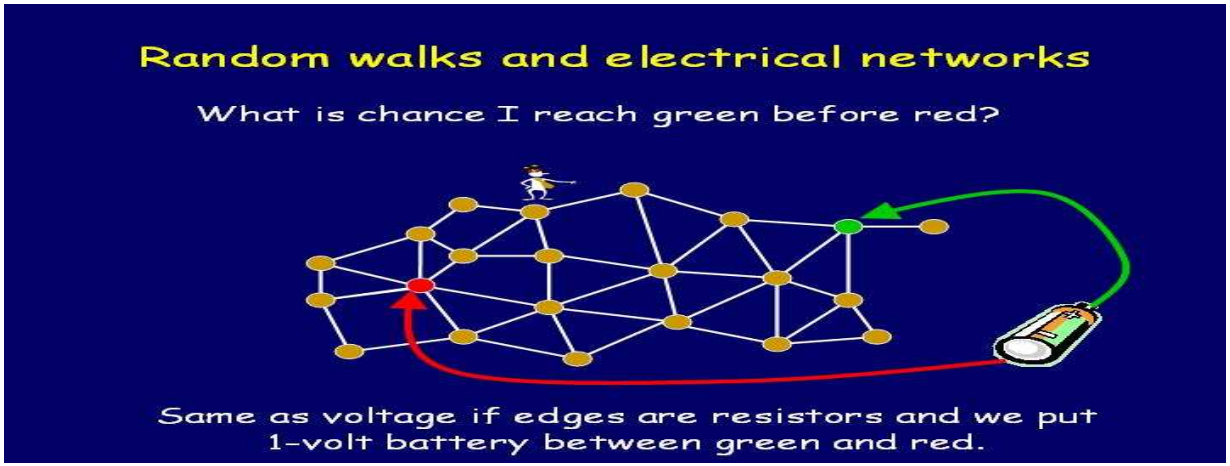
Let $\tau_A(x)$ be the time of the first hit at A by the random walk starting from x and $\tau_A^+(x)$ the first time after 0 that the random walk hits A. We set $\tau_A = \infty$ ($\tau_A^+ = \infty$) when there is not such a hit at all.

Lemma 146 (Existence Principle) *Let (G, C) be any electrical network and $W \subseteq V(G)$. For any bounded function $g \in \mathbb{R}^{V(G)\setminus W}$, there is $f \in \mathbb{R}^{V(G)}$ such that $f\,|_{V(G)\setminus W} = g$ and that f is harmonic on W.*

Proof. Let $x \in V(G) = V$. Take $X = (X_i)_{i\geq 0}$ to be the random walk starting from x, namely $X_0 = x$. Let $Y = g(X_{\tau_{V\setminus W}})$ provided $\tau_{V\setminus W} < \infty$ and $Y = 0$ otherwise. Put $f(x) = E_x(Y)$. Since g is bounded, f is well-defined. Clearly, $f\,|_{V(G)\setminus W} = g$ is trivially true. Moreover, for any $x \in W$, $f(x) = E_x(Y) = \sum_{x\sim y} P_{xy} E_y(Y) = \sum_{x\sim y} P_{xy} f(y)$, establishing Eq. (17) and yielding the assertion that f is harmonic at x, as required. ∎

Another proof: By connecting to battery, we set the potential of the nodes $x \in V \setminus W$ to be $v(x) = g(x)$ and look at the potential distribution v thus caused in the electrical network *. Lemma 142 implies that $f = v$ is the required extension of g, proving Lemma 146. ∎

*Apart from physics intuition, this existence of the potential function satisfying Ohm's law and Kirchhoff's law and the given boundary condition can be proved easily using linear algebra and is referred to as a Dirichlet problem.



Stolen from `http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15251-s04/Site/Materials/Lectures/Lecture24/`

Not many mathematicians believe that there are close relations between the classical Dirichlet problem and algebraic-topological invariants of spaces. Such relations appeared, more than half a century ago, in one of my papers [E1], with concrete applications to coverings of closed manifolds [E2]. They were restricted to spaces which admit a finite cell structure, like triangulated manifolds. The motivation came from two sources, the discrete analog of the Dirichlet problem for arbitrary (finite) graphs; and from the Hodge-de Rham decomposition of differential forms on closed manifolds.

We recall the discrete Dirichlet problem. A real function f of the vertices of P is given on the boundary of P. It has to be determined in the interior vertices of P in such a way that the "mean value theorem" holds; i.e that in any interior point p the value $f(p)$ is the arithmetic mean of the four neighboring values. It is easily seen that this is a problem of linear algebra which always has a well-determined solution. There exist many procedures for getting the solution by approximation. And there are also many interpretations in terms of random walks leading to the boundary, and of probability problems.

– Beno Eckmann, Harmonic Functions and Topological Invariants – any Relation? in: Mathematical Miniatures, Swiss Federal Institute (ETH), `http://www.fim.math.ethz.ch/preprints/2005`

Corollary 147 *Let (G, C) be a connected electrical network and W a finite proper subset of $V(G)$. Let $V(G)\backslash W$ be the disjoint union of S and T. Then the function $f(x) = P_x(\tau_S < \tau_T)$ is the unique function satisfying $f|_S = 1, f|_T = 0$ and f is harmonic on W. Moreover, let v be the potential in the electrical network when the nodes in S have been assigned potential 1 and the nodes in T potential 0. Then $v(x) = P_x(\tau_S < \tau_T)$.*

Proof. It is immediate from Lemma 144 and the two proofs of Lemma 146.∎

Remark 148 *Note that on the condition that W is of finite size, we have $P_x(\tau_S = \tau_T) = P_x(\tau_S = \tau_T = \infty) = 0$.*

We build understanding using figures and geometric arguments, and then translate ideas into algebraic formulas and algorithms. – D. Bertsimas, R. Weismantel, Optimization over Integers, Dynamic Ideas, 2005.

Let (G, C) be an electrical network in which only s and t are connected to battery. By Kirchhoff's Node Law (Lemma 141), we have $B^\top i(s) = \sum_{x:\ s\sim x} i(s,x) = \sum_{x:\ s\sim x} i(s,x) + \sum_{v\neq s,t} \sum_{x:\ v\sim x} i(v,x) = \sum_{\{x,y\}:\ t\notin\{x,y\}} (i(x,y) + i(y,x)) + \sum_{x:\ t\sim x} i(x,t) = \sum_{x:\ t\sim x} i(x,t) = -B^\top i(t)^*$. This common value is called the value of the current from s to t, or the amount of current flowing into the network at s and flowing out of the network at t.

Since $\sum_x i(s,x)$ is the total number of current flowing into the circuit at s, we may regard the entire circuit between s and t as a single conductor of effective conductance $C_{eff}(s \leftrightarrow t) \doteq \frac{\sum_{x:\ s\sim x} i(s,x)}{v(s)-v(t)} = \frac{d^* j(s)}{v(s)-v(t)} = \frac{B^\top i(s)}{v(s)-v(t)}$. The effective resistance is defined to be $R_{eff}(s \leftrightarrow t) = \frac{1}{C_{eff}(s\leftrightarrow t)}$, the potential difference between s and t ensuring a current of size 1 from s to t.

Exercise 149 *Let (G, C) be a connected electrical network. Pick $s \neq t \in V(G)$. Set s at potential 1 and set t at potential 0. Show that for the resulting potential function v we have $-\Delta v(s) = \Delta v(t) = C_{eff}(s \leftrightarrow t)$ and $\Delta v(x) = 0$ for $x \neq s, t$.*

*This result is essentially due to the fact that $0 = \sum_{x\in V(G)} B^\top i(x)$, which corresponds to the geometric fact that the boundary of a boundary is empty.

The escape probability from s to t of an electrical network (G, C) is the chance that a random walk starting at s will hit t before it returns to s and is denoted by $P_{esc}(s \to t)$. Observe that $P_{esc}(s \to t) = P_s(\tau_t < \tau_s^+)$. It is easy to see that

$$\sum_{x:\ s\sim x} P_{sx} P_x(\tau_s > \tau_t) = P_{esc}(s \to t). \tag{18}$$

Imposing voltage 1 at s and 0 at t, by Corollary 147, Remark 148 and Eq. (18), we get

$$\begin{aligned} C_{eff}(s \leftrightarrow t) &\doteq \frac{\sum_{x:\ s\sim x} i(s,x)}{v(s)-v(t)} \\ &= \sum_{x:\ s\sim x} (v(s) - v(x)) C(sx) \\ &= \sum_{x:\ s\sim x} (1 - v(x)) P_{sx} D(s) \\ &= D(s) \sum_{x:\ s\sim x} P_{sx}(1 - P_x(\tau_s < \tau_t)) \\ &= D(s) P_{esc}(s \to t). \end{aligned} \tag{19}$$

This gives

Theorem 150 $C_{eff}(s \leftrightarrow t) = D(s) P_{esc}(s \to t)$. *In particular,* $D(s) P_{esc}(s \to t) = D(t) P_{esc}(t \to s)$ *.

*Compare with Exercise 139.

Suppose that the diameter of the graph G is infinite. Fix $x \in V(G)$. Let G_r be the graph obtained from G by merging all vertices of G whose distance to x is larger than r into one vertex z_r. We define *

$$C_{eff}(G : \ x \leftrightarrow \infty) \doteq \lim_{r \to \infty} C_{eff}(G_r : \ x \leftrightarrow z_r). \tag{20}$$

Note that when we mention ∞ hereafter, its meaning should be understood in the same way as we do in Eq. (20).

Write $S_x(s \leftrightarrow t)$ for the expected sojourn time at x in a random round-trip from s to t. Namely, $S_x(s \leftrightarrow t)$ is the expected number of times the random walk pass by x if we start at s, continue till we get to t, and then stop when we are next at s. Let $S_{xy}(s \leftrightarrow t)$ be the expected number of times we traverse the edge xy from x to y during a round-trip from s to t. Let $S_x(s \to t)$ be the expected sojourn time at x if we start at s and stop when arrive at t.

Let $F_n(x, y)$ denote the probability that a random walk starts at x at time 0 and its first arriving time at y after time 0 is time n. † We can in some case consider a point to escape to infinity if it never returns to the origin.

*Theorem 165 guarantees that this limit exists, though may be infinite.

†This concept already appears in Corollary 137.

Lemma 151 $\frac{1}{P_{esc}(x \to \infty)} = S_x(x \to \infty)$.

Proof. Note that

$$1 - P_{esc}(x \to \infty) = \sum_{n=1}^{\infty} F_n(x, x). \tag{21}$$

Let $p = \sum_{n=1}^{\infty} F_n(x, x)$. Then, $\frac{1}{P_{esc}(x \to \infty)} = \frac{1}{1-p} = (1+2p+3p^2+4p^3+\cdots)(1-p) = S_x(x \to \infty)$. ∎

Remark 152

$$\begin{aligned}
Recurrent &\Leftrightarrow \textstyle\sum_{n=1}^{\infty} P_n(x, x) = \infty && \textit{By Theorem 135} \\
&\Leftrightarrow \textstyle\sum_{n=1}^{\infty} F_n(x, x) = 1 && \textit{By Corollary 137} \\
&\Leftrightarrow P_{esc}(x, \infty) = 0 && \textit{By Eq. (21)} \\
&\Leftrightarrow S_x(x \to \infty) = \infty && \textit{By Lemma 151} \\
&\Leftrightarrow C_{eff}(x \leftrightarrow \infty) = 0 && \textit{By Theorem 150} \\
&\Leftrightarrow R_{eff}(x \leftrightarrow \infty) = \infty
\end{aligned}$$

I consider the topic of this book to be the intersection of two topics: $random\ graphs \cap statistical\ pattern\ recognition$. In this way I agree with Knuth that disparate fields of knowledge can be used to shed light on each other, and to bridge the gap between them. – David J. Marchette, Random Graphs for Statistical Pattern Recognition, Wiley, 2004.

Exercise 153 *(i)* $\frac{1}{P_{esc}(s\to t)} = S_s(s \to t)$; *(ii)* $R_{eff}(s \leftrightarrow t) = \frac{S_s(s\to t)}{D(s)}$.

Exercise 154 *Show that the infinite random walk on the integer line is recurrent.* *

Definition 155 *Motivated by Kirchhoff's Current Law (Lemma 141), a function* $f \in \mathbb{R}^{A(G)}$ *satisfying* $B^\top f = 0$ *is called a flow in the graph* G. *For any* $s, t \in V(G)$, *we call* $f \in \mathbb{R}^{A(G)}$ *a flow from* s *to* t *of value* h *in* G *provided* $d^* f(v) = B^\top f(v) = 0$ *for any* $v \neq s, t$ *and* $B^\top f(s) = -B^\top f(t) = h$. *A unit flow from* x *to* ∞ *is an element* $f \in \mathbb{R}^{A(G)}$ *satisfying* $B^\top f(x) = 1$ *and* $B^\top f(y) = 0$ *for any* $y \neq x$.

*By Theorem 150: Calculate $R_{eff}(0 \leftrightarrow \infty)$; By Lemma 151: Deduce from $\sum_{i=1}^{\infty} \frac{\binom{2i}{i}}{2^{2i}} \geq \sum_{i=1}^{\infty} \frac{1}{2\sqrt{i}} = \infty$ that the expected number of times that such a random walk visits its starting point is unbounded.

Theorem 156 * *Let $N = (G, C)$ be a connected electrical network with $s, t \in V(G), s \neq t$. Setting s at abstract potential $R_{eff}(s \leftrightarrow t)$ and t at absolute potential* 0, *so that there is a current of size* 1 *from s to t through N, the distribution of potential is given by $v(x) = \frac{S_x(s \to t)}{D(x)}$ for $x \in V(G)$ and the distribution of current is given by $i(a) = E(\#\{\tau_t(s) > n \geq 0 : X_n = a^-, X_{n+1} = a^+\} - \#\{\tau_t(s) > n \geq 0 : X_n = a^+, X_{n+1} = a^-\})$, where (X_n) denotes the random walk starting from $X_0 = s$, for $a \in A(G)$.*

Proof. Consider the set of all random walks w beginning from s and ending at time τ_t. Send a unit flow through the walk w and denote this flow by i_w. The set of these unit flows can be made into a probability space by using the corresponding measures defined on the random walks. Clearly, $i = E(i_w)$ and hence i is still a unit flow from s to t.

To prove the theorem, it remains to verify that Ohm's law (Lemma 140) is satisfied for the unit flow i and the asserted potential function v. This is demonstrated by the following computation: For any $a \in A(G)$, $(v(a^-) - v(a^+))C(a) = (\frac{S_{a^-}}{D(a^-)} - \frac{S_{a^+}}{D(a^+)})C(a) = S_{a^-}P_{a^-a^+} - S_{a^+}P_{a^+a^-}{}^\dagger = E(i_w)(a) = i(a)$. ∎

Infer from Theorems 150 and 156 the assertion in Exercise 153!

*B. Bollobás, Modern Graph Theory, p. 306, Theorem 9.

†Note that $S_t = 0$!

Theorem 157 $R_{eff}(s \leftrightarrow t) = S_{xy}(s \leftrightarrow t) = \frac{S_x(s \to t)}{D(x)} + \frac{S_x(t \to s)}{D(x)}$.

The history of mathematics shows that "point of view" can be very important. What is difficult from one point of view may become easy from another. ... Generally, the more varied and effective the points of view which a subject admits, the more profound and useful it becomes.

Applications are the touchstone of mathematics. The author started solving combinatorial isoperimetric problems as a research engineer in communications at the Jet Propulsion Laboratory. Since then, as a mathematician at the Rockefeller University and the University of California at Riverside, applications to science and engineering have continued to motivate the work. A good application for the solution of a hard problem doubles the pleasure, and every other benefit, from it. Global methods are by nature abstract and might easily degenerate into what von Neumann called "baroque mathematics" if not guided by real applications. On several occasions over the years, promising technical insights were left undeveloped until the right application came along. We would recommend that same caution to others developing global methods.

– L.H. Harper, Global Methods for Combinatorial Isoperimetric Problems, Cambridge University Press, 2004.

End of Lesson Twenty One 19/12/06

For $f, g \in \mathbb{R}^{A(G)}$, define their inner product on an edge (x, y) to be $\langle f, g \rangle_e = (f(x) - f(y))(g(x) - g(y))$, which, for any orientation a of this edge, can be written as $(f(a^+) - f(a^-))(g(a^+) - g(a^-))$, and define their inner product on the graph to be $\langle f, g \rangle = \sum_{e \in E(G)} \langle f, g \rangle_e = \sum_{a \in A(G)} (f(a^+) - f(a^-))(g(a^+) - g(a^-)) = \frac{1}{2} \sum_{x \in V(G)} \sum_{y \sim x} (f(y) - f(x))(g(y) - g(x)) = \sum_{x \in V(G)} \sum_{\substack{a \in A(G) \\ x = a^-}} (f(a^+) - f(x))(g(a^+) - g(x))$.

For $f, g \in \mathbb{R}^{V(G)}$, define their inner product to be $\langle f, g \rangle = \sum_{v \in V(G)} f(v) g(v)$.

Exercise 158 *[Kirchhoff's Cycle (Potential) Law] Deduce from Ohm's Law (Lemma 140) that it holds $\langle f, C^{-1} i \rangle = 0$ for any flow f of the graph G and any current i in the electrical network (G, C).*

Green's Formula:

$$\int_\Omega (\nabla f \cdot \nabla g + f\Delta g)dS = \int_{\partial\Omega} f\frac{\partial g}{\partial n}ds.$$

Lemma 159 (Discrete Gauss-Green Theorem) * $\langle -\Delta(f), g\rangle = \langle d(f), d(g)\rangle$.

Proof. $\langle -\Delta(f), g\rangle = f^\top(D-C)g = f^\top(d^*d)g = (df)^\top(dg) = \langle d(f), d(g)\rangle$. ∎

It is noteworthy that $f^\top Lf = -f^\top\Delta f = \langle df, df\rangle$ represents the energy dissipation in the electrical circuit when the potential distribution is given by f.

Quiz: Use Lemma 159 to give another proof of Lemma 144. †

Exercise 160 (Conservation of energy dissipation) *The energy dissipation of an electrical network is* $C_{eff}(s \leftrightarrow t)(v(s)-v(t))^2$.‡

*This is a very useful double counting formula.

†The support of $f-g$ and $\Delta(f-g)$ are disjoint.

‡Eq. (26)

We call $E(\tau_t(s))$ the hitting time from s to t and denote it by $H(s,t)$. The parameter $k(s,t) = H(s,t) + H(t,s)$ is said to be the commute time between s and t.

Theorem 161 [*] *Let (G,C) be a connected finite network. For any $s \neq t \in V(G)$, we have $C_{eff}(s \leftrightarrow t)k(s,t) = \sum_{x \in V(G)} D(x)$.*

Proof. Let g_s and g_t be the potential distributions of (G,C) when we set $g_s(s) = 0 = g_s(t) - 1$ and when we set $g_t(t) = 0 = g_t(s) - 1$, respectively, through connecting battery to s and t.[†] Define $f_s, f_t \in \mathbb{R}^{V(G)}$ by setting $f_s(x) = E(\tau_s(x))$ and $f_t(x) = E(\tau_t(x))$ for any $x \in V(G)$, respectively.

Note that

$$f_s(s) = 0 \tag{22}$$

and that

$$(\Delta f_s)(x) = -D(x), x \neq s, \tag{23}$$

[*] C.St.J.A. Nash-Williams, Random walks and electric currents in networks, Proc. Cambridge Phil. Soc. 55 (1959), 181–194. P. Tetali, Random walks and effective resistance of networks, J. Theoretical Probability, 1 (1991), 101–109.

[†] Clearly, $g_s + g_t = 1$.

the latter coming from $f_s(x) = 1 + \sum_{y:y\sim x} P_{xy} f_s(y)$ for $x \neq s$. We infer from Exercise 149 and Eq. (22) that $C_{eff}(s \leftrightarrow t) f_s(t) = \langle f_s, -\Delta g_s \rangle$. This, combined with Lemma 159, Eq. (23) and $g_s(s) = 0$, then leads to

$$C_{eff}(s \leftrightarrow t) f_s(t) = \langle \Delta f_s, -g_s \rangle = \langle D, g_s \rangle. \tag{24}$$

Analogously, we have

$$C_{eff}(s \leftrightarrow t) f_t(s) = \langle \Delta f_t, -g_t \rangle = \langle D, g_t \rangle. \tag{25}$$

Putting Eqs. (24) and (25) together, we arrive at $C_{eff}(s \leftrightarrow t)(f_t(s) + f_s(t)) = \langle D, g_s + g_t \rangle = \sum_{x \in V(G)} D(x)$, proving the theorem. ∎

Let G be a graph. For any $u \in V(G)$, let $\mathscr{C}_u(G)$ be the expected number of steps for a random walk starting at u on (G, C), where C takes constant value, to visit all vertices of G after time 0. The cover time $\mathscr{C}(G)$ of G is $\max_{u \in V(G)} \mathscr{C}_u(G)$.

Exercise 162 *Let G be a connected graph with n vertices and m edges. Prove that 1) $\mathscr{C}(G) \leq 2m(n-1)$;* 2) With probability $\geq 1 - \frac{1}{2^k}$, the random walker on the graph G has visited every vertex from time 1 to time $4km(n-1)$.*

*Choose a spanning tree of G. Make use of Theorem 161 and the linearity of expectation.

Theorem 163 (Dirichlet's principle) *The energy dissipation of the network* (G, C) *when two of its vertices* s *and* t *are set at potential difference* 1 *is* $C_{eff}(s \leftrightarrow t) = \min\{\langle df, df\rangle : \ f \in \mathbb{R}^{V(G)}, f(s) = 1, f(t) = 0\}$.

Proof. Let v be the potential distribution which is specified by $v(s) = 1, v(t) = 0$ and $\Delta(v)(x) = 0$ for $x \neq s, t$. Then,

$$\begin{array}{l} \langle dv, dv\rangle = \langle -\Delta(v), v\rangle = -\Delta(v)(s) = (D - C)(v)(s) \\ = D(s)v(s) - \sum_{x:\ x\sim s} C(sx)v(x) = \sum_{x:\ x\sim s} C(sx)(v(s) - v(x)) \\ = \sum_{x:\ x\sim s} i(s, x) = \frac{\sum_{x:\ x\sim s} i(s,x)}{v(s)-v(t)} = C_{eff}(s \leftrightarrow t). \end{array} \tag{26}$$

On the other hand, for any $f \in \mathbb{R}^{V(G)}$ satisfying $f(s) = 1$ and $f(t) = 0$, we put $g = f - v$ and can check that

$$\begin{array}{lll} \langle df, df\rangle & = \langle dv, dv\rangle + \langle dg, dg\rangle + 2\langle dv, dg\rangle & \\ & = \langle dv, dv\rangle + \langle dg, dg\rangle - 2\langle \Delta f, g\rangle & \text{By Lemma 159} \\ & \geq \langle dv, dv\rangle & \text{The supports of } \Delta f \text{ and } g \text{ are disjoint} \\ & = C_{eff}(s \leftrightarrow t). & \text{By Eq. (26)} \end{array} \tag{27}$$

∎

Theorem 164 *Denote by $\mathscr{F}(s,t)$ the set of unit flows from s to t in (G,C).* $R_{eff}(s \leftrightarrow t) = \min\{\sum_{a\in A(G)} \frac{f(a)^2}{C(a)} : \ f \in \mathscr{F}(s,t)\} = \min\{\langle\sqrt{C}^{-1}f, \sqrt{C}^{-1}f\rangle : \ f \in \mathscr{F}(s,t)\}$.

Proof. Let v be the potential distribution which is specified by $v(s) = R_{eff}(s \leftrightarrow t), v(t) = 0$ and $\Delta v(x) = 0$ for any $x \in V(G) \setminus \{s,t\}$. Note that there is a current $i \in \mathscr{F}(s,t)$ flowing from s to t under this potential distribution. In view of Eq. (26), we have $R_{eff}(s \leftrightarrow t) = R_{eff}(s \leftrightarrow t)^2 C_{eff}(s \leftrightarrow t) = \langle dv, dv\rangle$. By now, an application of Lemma 140 gives $R_{eff}(s \leftrightarrow t) = \langle\sqrt{C}^{-1}i, \sqrt{C}^{-1}i\rangle$.

For f varying in $\mathscr{F}(s,t)$, $\frac{d}{dt}f$ takes values in all flows of G. Thus, if f is the element among $\mathscr{F}(s,t)$ for which $\langle\sqrt{C}^{-1}f, \sqrt{C}^{-1}f\rangle$ attains the minimum *, we should have $0 = \frac{d}{dt}\langle\sqrt{C}^{-1}f, \sqrt{C}^{-1}f\rangle = 2\langle\frac{d}{dt}\sqrt{C}^{-1}f, \sqrt{C}^{-1}f\rangle = 2\langle\frac{d}{dt}f, C^{-1}f\rangle$. This tells us that $C^{-1}f$ is orthogonal to the flow space and hence Kirchhoff's Cycle Law (Exercise 158) is valid and we come to the conclusion that f is nothing but the current distribution i. ∎

Theorem 163 together with Theorem 164 gives the so-called Thomson's principle, saying that the currents and potentials are distributed in such a way as to minimize the total energy in the network.

*Check that this minimum does exist!

If the specific resistance of any portion of the conductor be changed, that of the remainder being unchanged, the resistance of the whole conductor will be increased if that of the portion is increased, and diminished if that of the portion is diminished. This principle may be regarded as self-evident... – James Clerk Maxwell, Treatise on Electricity and Magnetism, 3rd Edition, 1891.

Either of Theorem 163 and Theorem 164 applies to prove the following result.

Theorem 165 (Rayleigh's Monotonicity Principle) *If the resistance of a wire is increased then the effective resistance between two vertices does not decrease.*

The method of applying shorting and cutting to get lower and upper bounds for the resistance of a resistive medium was introduced by Lord Rayleigh. We will refer to Rayleigh's technique collectively as Rayleigh's short-cut method. This does not do Rayleigh justice, for Rayleigh's method is a whole bag of tricks that goes beyond mere shorting and cutting – but who can resist a pun? – P.G. Doyle, J.L. Snell, Random walks and electric networks, Math. Assoc. Amer., 1984.

The pair of dual operations, Shorting and Cutting, is important in various kinds of mathematics. Its basic role will be clear if you try to understand pieces of Matroid Theory.

Exercise 166 *Prove Theorem 163 and Theorem 164 by appealing to Theorem 165.* [*]

Exercise 167 *Let $\mathbb{Z}^2$ the the two dimensional integer lattice and let C map each edge of G to* 1. *Show that $R_{eff}((0,0) \leftrightarrow \infty) = \infty$ and hence $\mathbb{Z}^2$ is recurrent.*[†]

Lemma 168 [‡] *Let (G,C) be a denumerable connected network. Random walk on G is transient if and only if there is a unit flow on G of finite energy from some vertex to ∞.*

Proof. Take $x \in V(G)$ and let G_r be the graph obtained from G by merging all vertices of G whose distance to x is larger than r into one vertex z_r.

[*]L. Onsager, Reciprocal relations in irreversible processes I., Phys. Rev. 37 (1931), 405–426.

[†]Let S_n, $n \geq 1$, be those $8n$ vertices which form a symmetric square with corners $(\pm 1, \pm 1)$. Let C' be the function which maps any edge connecting vertices in the same S_n to 0 and all other edges to 1. Note that there are $8n+4$ edges between S_n and S_{n+1} and thus $R_{eff}((\mathbb{Z}^2, C'), (0,0) \leftrightarrow \infty) = \sum_{n=0}^{\infty} \frac{1}{8n+4} = \infty$.

[‡]T. Lyons, A simple criterion for transience of a reversible Markov chain, Ann. Probab. 11 (1983), 393–402.

Suppose there is $i \in \mathscr{F}(x, \infty)$ possessing finite energy $\mathfrak{E} = \langle \sqrt{C}^{-1} i, \sqrt{C}^{-1} i \rangle$. Then i restricted to $A(G_r)$ becomes a member of $\mathscr{F}(G_r; x, z_r)$, say i_r. It follows from Theorem 164 that $R_{eff}(x \leftrightarrow z_r) \leq \langle \sqrt{C}^{-1} i_r, \sqrt{C}^{-1} i_r \rangle \leq \mathfrak{E}$ and follows from Theorem 165 that $R_{eff}(G_r, x \leftrightarrow z_r) \leq R_{eff}(G_{r'}, x \leftrightarrow z_{r'})$ when $r < r'$. This then establishes that $R_{eff}(x \leftrightarrow \infty) = \lim_{r \to \infty} R_{eff}(G_r, x \leftrightarrow z_r) \leq \mathfrak{E}$ and hence (G, C) is transient by virtue of Theorem 135 and Remark 152.

Conversely, if (G, C) is transient, we know from Remark 152 and Theorem 164 that there is a constant $\mathfrak{E}$ and for any $r \geq 1$ there is $i_r \in \mathscr{F}(G_r; x, z_r)$ such that $\langle \sqrt{C}^{-1} i, \sqrt{C}^{-1} i \rangle = R_{eff}(G_r, x \leftrightarrow z_r) \leq \mathfrak{E}$. Note that each i_r corresponds to an element f_r of $\mathbb{R}^{A(G)}$ which takes value $i_r(a)$ for those arcs a with $a^+, a^- \in V(G_r) \setminus \{z_r\}$ and takes value zero elsewhere. Utilizing the Cantor diagonal argument, we can choose a subsequence (r_n) such that f_{r_n} converge pointwise to an element $i \in \mathbb{R}^{A(G)}$. It is not hard to check that $i \in \mathscr{F}(x, \infty)$ and has finite energy. ∎

Exercise 169 *A connected network (G, C) is recurrent if and only if for any $\epsilon > 0$, we can find $f \in \mathbb{R}^{V(G)}$ such that $f(x) = 1$, $\lim_{dist(x,y)\to\infty} f(y) = 0$ and $\langle df, df \rangle < \epsilon$. Here $dist(x, y)$ means the distance between x and y on the graph G.*

We use a global method to prove Theorem 163 (decomposing the edge subspace into a direct sum of two subspaces) and a local method to prove Theorem 164 (looking at the extremal point).

Exercise 170 *Use the local method to prove Theorem 163 and the global method to prove Theorem 164.*[*]

[*]Make use of Exercise 160.

The 1996 Gödel Prize for outstanding journal articles in the area of theoretical computer science was awarded on May 23, 1996 jointly to Mark Jerrum and Alistair Sinclair for their papers "Approximate counting, unform generation and rapidly mixing Markov chains," Information and Computation 82 (1989), 93-133, by Sinclair and Jerrum, and "Approximating the permanent," SIAM Journal on Computing 18 (1989), 1149-1178, by Jerrum and Sinclair. The first paper demonstrates a two-way connection between the mixing rate of a Markov chain and a graph-theoretic quantity called conductance, and provided the canonical paths tool for establishing high conductance. The second paper then brilliantly applies this method to prove the rapidly mixing property of a certain random walk on the set of all perfect matchings of a dense graph, thereby providing a polynomial time algorithm for approximating the permanent. Together, these two papers have helped trigger a Markov Chain "renaissance" in the 1990's, and one can conservatively count seventy major papers employing Markov chains in the style and methodology they initiated, covering areas as diverse as matching algorithms, geometric algorithms, mathematical programming, statistics, physics-inspired applications, and dynamical systems. In terms of their innovation and far reaching impact, the Jerrum-Sinclair papers are worthy of a prize named after Kurt Gödel. – `http://sigact.acm.org/prizes/godel/`

Enjoy Great Theoretical Ideas In Computer Science at

`http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15251-s05/Site/Materials/Lectures/`

Especially, look into the lecture on the great idea of 'Random Walk' here:

`http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15251-s04/Site/Materials/Lectures/Lecture24/`

A drunk man will find his way home, but a drunk bird may get lost forever. – Shizuo Kakutani (August 28, 1911 – August 17, 2004) See `http://improbable.com/pages/airchives/paperair/volume11/v11i2/AIR-11-2-kakutani.pdf` *for lost theorems of Kakutani.*

The Drunken Old Man's thoughts are not set on wine, but on the pleasures to be found among the mountains and rivers. – OUYANG Xiu (1007 – 1072)

End of Lesson Twenty Two 22/12/06

In the early sixties György Pólya gave a talk in Budapest where he told the story of this Theorem. He studied in the teens at the ETH (Federal Polytechnical School) in Zürich, where he had a roommate. It happened once that the roommate was visited by his fiancée. From politeness Pólya left the room and went for a walk on a nearby mountain. After some time he met the couple. Both the couple and Pólya continued their walks in different directions. However, they met again. When it happened the third time, Pólya had a bad feeling. The couple might think that he is spying on them. Hence he asked himself what is the probability of such meetings if both parties are walking randomly and independently. If this probability is big, then Pólya might claim that he is innocent.

By a simple generalization of the Recurrence Theorem it is easy to see that if the parties wander independently according to the law of the random walk then they meet infinitely often with probability 1. – Pál Révész, Random walk in random and non-random environments (2nd edition), World Scientific, 2005.

We now apply this form of the cutting method to give another proof that simple random walk on the three dimensional lattice is transient. All we need is a flow to infinity with finite dissipation. The flow we are going to describe is not the first flow one could think of. In case you are curious, the flow described here was constructed as a side effect of an unsuccessful attempt to derive the isoperimetric inequality from the "max-flow min-cut" theorem. The idea is to find a flow in the positive orthant having the property that the same amount flows through all points at the same distance from $\mathbf{0}$*. – P.G. Doyle, J.L. Snell, Random walks and electric networks, Math. Assoc. Amer., 1984.*

Theorem 171 (George Pólya 1921) * *$\mathbb{Z}^d$ is transient if and only if $d \geq 3$.*

Proof. By Remark 152, Theorem 165 and Exercise 167, it suffices to treat the case of $d = 3$. For this purpose, according to Lemma 168, we need to construct a unit flow from $(0,0,0)$ to ∞ having finite energy.

Let e_1, e_2, e_3 be the three unit vectors of $\mathbb{Z}^3$ and represent each vertex v as $v = \sum_{i=1}^3 v_i e_i, v_i \in \mathbb{Z}$. For $t \geq 0$, let $\mathfrak{S}_t = \{v \in \mathbb{Z}^3 : \sum_{i=1}^3 v_i = \sum_{i=1}^3 |v_i| = t\}$ and let $\mathfrak{B}_t$ be the set of arcs going from a vertex in $\mathfrak{S}_t$ to a vertex in $\mathfrak{S}_{t+1}$.

*Recall Exercise 124.

Let $\mathfrak{B}_t(i) = \{a \in \mathfrak{B}_t : \ a_i^- = a_i^+ - 1\}$, $i = 1, 2, 3$. For any $a \in \mathfrak{B}_t(i)$, put $i(a) = \frac{2a_i^+}{(t+1)(t+2)(t+3)}$ and set $i(a) = 0$ for $a \in A(\mathbb{Z}^3) \setminus (\cup_{t=0}^{\infty} \mathfrak{B}_t)$.

- $B^\top i(0) = \frac{2\times 3}{6} = 1$.
- $B^\top i(v) = \frac{2(v_1+1+v_2+1+v_3+1)}{(t+1)(t+2)(t+3)} - \frac{2(v_1+v_2+v_3)}{t(t+1)(t+2)} = \frac{2}{(t+1)(t+2)} - \frac{2}{(t+1)(t+2)} = 0$, if $v \in \mathfrak{S}_t, t \geq 1$.
- $B^\top i(v) = 0$, if $v \notin \cup_{t=0}^{\infty} \mathfrak{S}_t$.
- $\mathfrak{S}_t$ contains $\binom{t+2}{2}$ vertices, each issuing three arcs to $\mathfrak{S}_{t+1}$ with a flow no greater than $\frac{2}{(t+2)(t+3)}$. Henceforth, letting R be the constant resistance of each edge of the network, the total energy dissipation is $\leq \sum_{t=0}^{\infty} |\mathfrak{S}_t| (\frac{2}{(t+2)(t+3)})^2 R \leq \sum_{t=0}^{\infty} \frac{2R}{(t+2)(t+3)} = R$.

∎

Please read several other beautiful proofs of Theorem 171 in: P.G. Doyle, J.L. Snell, Random walks and electric networks, Math. Assoc. Amer., 1984.

The analog of Pólya's theorem in this connection is that wind instruments are possible in our three-dimensional world, but are not possible in Flatland. – P.G. Doyle, J.L. Snell, Random walks and electric networks, Math. Assoc. Amer., 1984.

I am too good for philosophy and not good enough for physics. Mathematics is in between. – George Pólya

We had a course by George Pólya on probability, with many nice and amusing applications, including his favorite topic Random Walks, so beautifully related to discrete harmonic functions in a graph. We liked it, but we were unhappy with the so-called definition of the probability concept. Pólya did not care much, intuition and application were more important to him. – Beno Eckmann, Probability and Cohomology, in: Mathematical Miniatures, Swiss Federal Institute (ETH), `http://www.fim.math.ethz.ch/preprints/2005`

Gerald L. Alexanderson (Ed.), The Random Walks of George Pólya, Cambridge University Press, 2000.

Exercise 172 (Laszlo Babai) [*] *Consider a 'phased' random walk on $\mathbb{Z}^d$, where at step n, we move with the probability $1-2p$ of staying put and each with the probability p of moving one step along the two ways in the direction of dimension n mod d. Show that the probability of returning to the origin at the nth step is $O(n^{-2/d})$. Is it possible to find a simple proof of Pólya's Theorem (Theorem 171) by approximating the random walk on $\mathbb{Z}^d$ considered there by this phrased random walk?*

[*]M. Waliji, Monkeys and walks, available at `http://www.math.uchicago.edu/~may/VIGRE/VIGREREU2006.html`.

On Looking for a Hermit and not Finding Him – JIA Dao (779–843)
I questioned a boy under the pine trees.
"My Master went herb-gathering" he says,
"He is still somewhere on the mountain-side,
So deep in the clouds I can't tell where."

http://www.resurgence.org/2005/badiner228.htm

A NOTE LEFT FOR AN ABSENT ECLUSE – JIA Dao (779–843)
When I questioned your pupil, under a pine-tree,
"My teacher," he answered, " went for herbs,
But toward which corner of the mountain,
How can I tell, through all these clouds ?"

http://etext.virginia.edu/chinese/tangeng.html

http://www.print-sozai.sakura.ne.jp/07inoshishi/samtext/01/03.htm

End of The Last Lesson 26/12/06

Slides Graveyard

Lemma 121 is a special case of the following.

Theorem 173 (Kolmogorov's Zero-one Law)

A set $x_1, \ldots, x_k$ of positive integers is said to have distinct sums if all sums $\sum_{i \in S} x_i$, $S \subseteq \{1, \ldots, k\}$, are distinct. Let $f(n)$ denote the maximum k for which there exists a set $\{x_1, \ldots, x_k\} \subseteq \{1, \ldots, n\}$ with distinct sums.

Theorem 174 * $f(n) \leq \log_2 n + \frac{1}{2}\log_2 \log_2 n + O(1)$.

Proof. Fix $\{x_1, \ldots, x_k\} \subseteq \{1, \ldots, n\}$ with distinct sums. Let $\epsilon_1, \ldots, \epsilon_k$.

Proof. ∎

*Noga Alon, Joel H. Spencer, The Probabilistic Method, Wiley, 1992.

Let $p_1, \ldots, p_n$ be a collection of points in the plane. For simplicity, we assume that no two points are vertically aligned and that no three points are colinear. Given $1 \leq i < j \leq n$, put $n_{i,j}$ to be the number of points lying strictly below the line passing through p_i and p_j, and let f_k denote the number of pairs i, j such that $n_{i,j} = k$.

Theorem 175 [*] *$f_k = O(kn)$.*

Proof. ∎

[*] K.L. Clarkson, P.W. Shor, Applications of random sampling in computational geometry, II, Disc. Comput. Geometry 4 (1989), 387–421.

Basic research is like shooting an arrow into the air and, where it lands, painting a target. − Homer Burton Adkins (1892-1949, American organic chemist)

New opinions are always suspected, and usually opposed, without any other reason but because they are not already common. − John Locke

It is important for him who wants to discover not to confine himself to one chapter of science, but to keep in touch with various others. − Jacques Hadamard

law of large numbers and information theory, data compression: $X_1, X_2, \ldots,$ i.i.d., $\lim_{n\to\infty} -\frac{\log p(X_1,\ldots,X_n)}{n} = H(X)$

Finally, and this is mentioned here only in response to a query by Doob, I chose to present the brutal Theorem 5.3.2 in the original form given by Kolmogorov because I want to expose the student to hardships in mathematics. − Kai Lai Chung, A Course in Probability Theory, (Sec. Ed.) Academic Press, 1974.

Central limit theorem

Borel, Probability and Certainty, Chap. 3.

The probability of male birth

The sex of twins

Predicting the result of poll

Mebane, Walter R., Jr. Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections, Prepared for delivery at the Election Fraud Conference, Salt Lake City, Utah, September 29–30, 2006. available at `http://macht.arts.cornell.edu/wrm1/fraud06.pdf`

Markov chain Monte Carlo (MCMC): The method originates in physics, where the earliest uses go back to the 1950's. It later enjoyed huge booms in other areas, especially in image analysis in the 1980's, and in the increasingly important area of statistics known as Bayesian statistics in the 1990's.*

Zipf Distribution

Probabilistically thinking helps you do very complex double counting and extract important invariants, without using such language it will sometimes gets very messy and makes you lost. Probabilistic method is a powerful tool and it is a great loss of you if you only use it in your probability final exam.

*Olle Häggström, Finite Markov Chains and Algorithmic Applications, Cambridge University Press, 2003.

Lov$\acute{a}$sz's sieve: Suppose that an experiment can fail if any one of n bad events occurs. We want to know if there is a non-zero probability that the experiment will succeed. The Lovasz local lemma guarantees that an experiment will succeed with nonzero probability when the events are "almost independent". There exists a satisfying truth assignment for any instance of k-SAT for $k \geq 10$ in which each variable is contained in at most $2^{\frac{k}{2}}$ clauses.

Capacity of binary symmetric channel, encoding, decoding,

Phase transition

Some Elementary Results around the Wigner Semicircle Law

Often you cannot analytically compute the results you want from the model, either because the mathematics at your command are inadequate, or because you simply cannot formulate the problem in a sufficiently mathematical form to handle it. Then, of course, you have only simulations and experiments to guide you. When experiment and simulation differ, that can be annoying but can also be a possibility for finding new things that the experimentalist never thought about. The difference should not be swept under the rug and explained away as a sampling fluctuation; it is an opportunity to learn new things. – R.W. Hamming, The Art of Probability, Addison-Wesley, 1991.

Consider the hypercube $H := \{+1, -1\}^n$.

Define a bipartition $\{A, B\}$ of H into two disjoint subsets A and B a convex split if the convex hull $[A]$ of A does not intersect that of B.

(C1) A bipartition $\{A, B\}$ of H into two disjoint subsets A and B is convex iff $a, a' \in A$ and $b, b' \in B$ implies the existence of some index $i \in \{1, 2, \ldots, n\}$ with $\{a_i, a'_i\} \cap \{b_i, b'_i\} = \emptyset$.

In passing, I firmly believe that research should be offset by a certain amount of teaching, if only as a change from the agony of research. The trouble, however, I freely admit, is that in practice you get either no teaching, or else far too much. – The Mathematician's Art of Work – in Béla Bollobás (ed.), Littlewood's Miscellany, Cambridge, Cambridge University Press, 1986.