PROBABILITY THEORY

MATH 154

Unit 1: What is probability theory?

Introduction

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

1.2. Modern probability theory has been axiomatized as cleanly as the geometric axiomatic approach of Euclid for planar geometry. It plays a role when probing the nature of **space and time**. Fundamental conundrums are why primes show random behavior or why digits obtained by an expansion of fundamental constants pass tests as if they were typical random numbers. For example, if we look at primes p_n of the form $p = x^2 + y^2$ with $0 < y < x$, we see that the angles $X_n = \arctan(y_n/x_n)$ appear pretty random. If we look at the decimal digits X_n of π, all statistical tests indicate that the digit data come from a random process. But even the **normality of** π, the statement that all decimal digits of π appear with the same frequency, is unproven. Probability theory is a play field also in **linear algebra**, where one can looks for example at properties of **random matrices** or in statistical physics, when looking at **random networks**. How are the eigenvalues of random matrices or Laplacians of random networks distributed? In complex analysis, can can study for example the differences between successive roots of the Riemann zeta function using statistical tools.

1.3. Much of probability theory deals with **independent, identically distributed random variables**. This IID assumption is a very strong condition, but it allows to prove **theorems**. Given a sequence of X_k of such random variables, the **law of large numbers** assures that $\frac{1}{n}S_n$ converges to the expectation $\mathrm{E}[X]$. The average of a 1000

random dice events is expected to be close to 3.5. The **law of large numbers** assures that the normalized random variables $\overline{S_n}$ converge in law to the Gaussian distribution. The **law of iterated logarithms** refines this and assures that $\frac{1}{\sqrt{2n\log\log(n)}}S_n$ has its accumulation points in $[-1,1]$. Modern frontiers in the subject try to push these theorems to processes with less assumptions. The Birkhoff ergodic theorem for example generalizes the law of large numbers to ergodic measure preserving transformations. One can for example look at the irrational rotation $x \to x+\alpha = x+\sqrt{2} \bmod 2\pi$ and study $X_n(\theta) = \cos(\theta + n\alpha)$ which are identically distributed random variables. One still has $\frac{1}{n}S_n = \mathrm{E}[X]$ but the growth of S_n is governed by more subtle things. Under smoothness and arithmetic conditions, the sum S_n can even stay bounded.

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

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

We mention this example only to illustrate that we have hardly scratched the surface of the subject.

Oliver Knill, knill@math.harvard.edu, Math 154, Spring, 2025

PROBABILITY THEORY

MATH 154

Unit 2: Classical challenges

The Bertrand paradox

2.1. Bertrand asked in 1889, what the probability is that a random line on the unit disc intersects it with a length $\geq \sqrt{3}$, the length of the inscribed equilateral triangle. Here is the argument for $P = 1/3$: fix a point A on the boundary of the disc we can look at all lines through that point. For a polar angle $0 < \theta < \pi/3$ and $2\pi/3 < \theta < \pi$ the chord is longer than $\sqrt{3}$ and for $\pi/3 < \theta < 2\pi 3$ it is larger. A second answer gives $P = 1/2$: by looking at all points perpendicular to a fixed diameter the chord is longer than $\sqrt{3}$ if the point of intersection lies on the middle half of the diameter. A third answer gives $P = 1/4$: if the midpoint of the chord is in the disc of radius $1/2$, the chord is longer than $\sqrt{3}$. The area of that disk is $1/4$ of the area of the disk.

The Monty-Hall paradox

2.2. 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?
No switching: you choose a door and win with probability $1/3$. The opening of the host does not affect any more your choice. **Switching:** when choosing the door with the car, you loose since you switch. If you choose a door with a goat. The host opens the other door with the goat and you win. There are two such cases, where you win. The probability to win is $2/3$. There are now entire books on the subject [**?**]. The problem is related to **Baysian thinking**.

The Banach Tarski paradox

2.3. We work in the probability space the unit cube Ω in $\mathbb{R}^3$, where the events are the set of all subsets of Ω and where the probability $P[A]$ is the volume of A. We can take unions and intersections of events and keep having events. The axioms of probability theory assure that $P[\Omega] = 1, P[A \cup B] = P[A] + P[B]$ if A and B are disjoint. The volume of events is rotational and translational invariant as long as the turn or translation keeps us in Ω.

Now look at the following theorem: it is possible to write the ball $X = \{x^2+y^2+z^2 \leq 1/9\}$ as a disjoint union of 5 sets $X = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5$ and rotate and translate the sets in Ω to sets B_1, B_2, B_3, B_4, B_5 such that $B_1 \cup B_2 \cup B_3 = \{(x-1/2)^2+y^2+z^2 \leq 1/9\} = X - (1/2, 0, 0)$ and $B_4 \cup B_5 = \{(x+1/2)^2+y^2+z^2 \leq 1/9\} = X + (1/2, 0, 0)$. We have achieved, by cutting, translation and rotation that the ball has doubled in size. As $\mathrm{P}[S] + \mathrm{P}[S] = \mathrm{P}[S]$, this would imply $\mathrm{P}[S] = 0$. But since Ω can be covered by finitely many spheres of radius $1/3$ and each having $\mathrm{P} = 0$, we conclude that $\mathrm{P}[\Omega] = 0$ which is a contradiction. We have not lied. The theorem follows from the Axiom of Choice and is true, the conclusion paradox was correctly derived. But there obviously will have been a problem.

The Petersburg paradox

2.4. Assume you pay an entrance fee c for a game and that you win 2^T, where T is the number of times, the casino flips a coin until "head" appears. For example, if the sequence of coin experiments would give "tail, tail, tail, head", you would win $2^3 - c = 8 - c$, the win minus the entrance fee. For which c is the game fair? We can compute the expectation as $\sum_{k=1}^{\infty} 2^k \mathrm{P}[T = k] = \sum_{k=1}^{\infty} 1 = \infty$. But nobody would agree to pay even an entrance fee $c = 20$. The event $T = 20$ is so improbable that it never occurs in the life-time of a person.

2.5. What would be a reasonable entrance fee in "real life"? Bernoulli proposed to replace the expectation $\mathrm{E}[G]$ of the profit $G = 2^T$ with the expectation of $(\mathrm{E}[u(G)])^2$, where $u(x)$ is a **utility function** like $(x) = \sqrt{x}$. It leads to a fair entrance estimate

$$(\mathrm{E}[\sqrt{G}])^2 = (\sum_{k=1}^{\infty} 2^{k/2} 2^{-k})^2 = \frac{1}{(\sqrt{2}-1)^2} \sim 5.828... \, .$$

On the other hand, given any utility function $u(k)$, one can modify the casino rule. For example, we could just pay $(2^k)^2$ in the case $u(k) = \sqrt{k}$, or pay e^{2^k} for the utility function $u(k) = \log(k)$. Is there a good resolution to the difficulty?

The Martingale paradox

2.6. Here is a bullet proof **martingale strategy** in roulette: bet c dollars on red. If you win, stop, if you lose, bet $2c$ dollars on red. If you win, stop. If you lose, bet $4c$ dollars on red. Keep doubling the bet. Eventually after n steps, red will occur and you will win $2^n c - (c + 2c + \cdots + 2^{n-1}c) = c$ dollars. This example motivates the concept of martingales. Why can this foolproof strategy not used?

The Birthday paradox

2.7. The last example we mention in the notes illustrates that intuition can be misleading. There are 365 days in a year, so that that it appears that we appear to need a larger group of people to expect a Birthday collision. It turns out that already for a group with 23 people, the probability that two have the same birthday is larger than $1/2$. Coincidences happen more frequently. In class, we will look at a "top 10" list of paradoxa including some not mentioned here.

PROBABILITY THEORY

MATH 154

Unit 3: Algebras

3.1. If Ω is a set, a set $\mathcal{A}$ of subsets of Ω is called an **algebra** if it is closed under **intersection** $\cap$ and **symmetric difference** Δ and if $\Omega \in \mathcal{A}$. The algebra of sets behaves like the algebra of integers. We just have to think about $\Delta = +$ as addition and $\cap = \cdot$ as multiplication. Indeed, we can check **commutativity**, **associativity** and **distributivity**. These identities are logical conclusions. We can also visualize them as **Venn diagrams**. Figure 2) shows associativity $(A \cdot B) \cdot C = A \cdot (B \cdot C)$. It just encodes the set of elements in Ω which are in all of the sets. The algebra encodes so basic logical thinking rules that usually are taken for granted. Boolean algebra includes also Boolean logic like the "tertium non datur" $A \cup A^c = \Omega$.

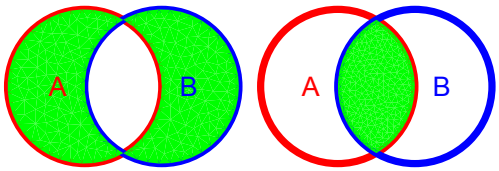


FIGURE 1. Venn diagram of the addition and multiplication in the algebra.

3.2. Any algebra of sets always contains the **empty set** $0 = \emptyset$, which is part of the algebra because it is $\Omega + \Omega$. The empty set is also called **zero** because $0 + A = A$ for every A. The set Ω plays the role of the 1 because $1 \cdot A = A$. The algebra is also known as a **Boolean algebra** because $A + A = 0$ and so $-A = A$. We can form other set operations like the union $A \cup B = AB + A + B = 1 + (1 + A)(1 + B)$ and the set difference $A \setminus B = B + AB$ and the **complement** $A^c = A + 1$. A Boolean algebra is a **commutative ring** with 1. Besides the laws $A + A = 0$ and $A^2 = A$, we have in particular $1 + 1 = 0$.

3.3. A set I is called **countable** if there is a bijection from I to the counting numbers $\mathbb{N} = \{1, 2, 3 \dots\}$. Every countable set of sets can therefore be written as a sequence of sets $\{A_1, A_2, \dots\}$. The **Hilbert hotel** pop-culture picture is the result that $\mathbb{N}$ and $2\mathbb{N} = \{2, 4, 6, 8, 10, \dots\}$ have the same cardinality. We can also count the rationals $\mathbb{Q}$, as seen in class. We can not count the numbers in the interval $[0, 1]$ however as Cantor showed in his famous diagonal argument: just assume to have an enumeration and construct from this a new number that is different from each of the numbers: just change in number k the k'th digit. In probability theory, countability plays a role. Remember the **Vitali set**.

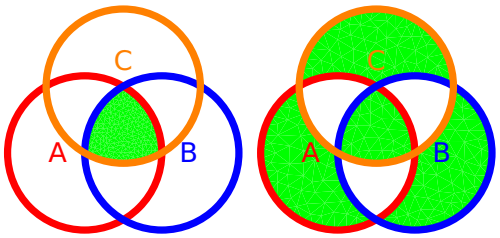


FIGURE 2. The Venn diagrams of multiplicative associativity $A \cdot (B \cdot C) = (A \cdot B) \cdot C$ and additive associativity $A + (B + C) = (A + B) + C$.

3.4. An algebra is called **σ-algebra** if it is closed under the formation of countable unions. A pair $(\Omega, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra on Ω is also called a **measurable space**. Formally, $A_n \in \mathcal{A} \implies \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$. This implies that countable intersections are in $\mathcal{A}$: $\bigcap_n A_n = 1 - (\bigcup_n (1 - A_n))$. An important property is that an arbitrary intersection of σ-algebras is a σ algebra.

3.5. Some examples:
1) For an arbitrary set Ω, $\mathcal{A} = \{\emptyset, \Omega\}$ is a σ-algebra. It is called the **trivial** σ-algebra.
2) If Ω is an arbitrary set, then $\mathcal{A} = 2^\Omega = \{A \subset \Omega\}$ is a σ-algebra. It is the largest σ-algebra one can define on Ω.
3) A finite set of pairwise disjoint sets $A_1, A_2, \ldots, A_n$ of Ω satisfying $\bigcup_j A_j = \Omega$ is called a **finite partition** of Ω. It generates the finite σ algebra $\mathcal{B} = \{A = \bigcup_{j\in J} A_j\}$, where J runs over all subsets of $\{1, \ldots, n\}$. Every finite σ-algebra has this form and its elements $\{A_1, \ldots, A_n\}$ are called the **atoms** of $\mathcal{B}$. **4)** Given an arbitrary set $\mathcal{C}$ of subsets in Ω, we can look at the intersection of all σ algebras which contain $\mathcal{C}$. It is called the σ-algebra **generated** by $\mathcal{C}$.
5) Given a topology $\mathcal{O}$ on Ω, a set of subsets of Ω that contains $\emptyset, \Omega$ and is closed under finite intersections and arbitrary unions. The σ algebra generated by this topology is called the **Borel σ algebra** of the topology.

3.6. Write $A_n \nearrow A$ if $A_n \subset A_{n+1}$ and $\bigcup_n A_n = A$. We say A is a **limit**. $(\Omega, \mathcal{A})$ is called a **π-system**, if $\mathcal{A}$ contains $\emptyset$ and $\mathcal{A}$ is closed under intersections. $(\Omega, \mathcal{A})$ is called a **λ-system** or Dynkin system if $\mathcal{A}$ contains Ω, is is closed under complements and closed under limits. Note that both π systems as well as λ systems do not need to be algebras.

Theorem 1. *$(\Omega, \mathcal{A})$ is a σ-algebra $\Leftrightarrow$ if it is a π-system as well as a λ-system.*

Proof. " $\Rightarrow$ ": Just check that $A \setminus B = A \cup B + B$.
" $\Leftarrow$ ": assume $\mathcal{A}$ is both a π-system and a λ system. Given $A, B \in \mathcal{A}$. By definition we know that $A^c = \Omega \setminus A, B^c = \Omega \setminus B$ is in $\mathcal{A}$. The π-system property implies that $A \cup B = \Omega \setminus (A^c \cap B^c) \in \mathcal{A}$. Since complements can be formed we have $A + B = A \cup B \setminus A \cap B$ in $\mathcal{A}$. Given a sequence $A_n \in \mathcal{A}$. Define $B_n = \bigcup_{k=1}^n A_k \in \mathcal{A}$ and $A = \bigcup_n A_n$. Because $B_n \nearrow A$ we know that A is a limit and that $A \in \mathcal{A}$. This finishes the proof that $\mathcal{A}$ is a σ-algebra. □

PROBABILITY THEORY

MATH 154

Unit 4: Probability Measures

4.1. $(\Omega, \mathcal{A}, \mathrm{P})$ is called a **probability space**, if $\mathcal{A}$ is a σ algebra on Ω and $\mathrm{P} : \mathcal{A} \to [0,1]$ is (i) **non-negative** $\mathrm{P}[A] \geq 0$, (ii) **normalized** $\mathrm{P}[\Omega] = 1$, and (iii) **σ-additive**: $A_n \in \mathcal{A}$ disjoint $\Rightarrow \mathrm{P}[\bigcup_n A_n] = \sum_n \mathrm{P}[A_n]$.

4.2. A function λ from a π-system $\mathcal{I}$ to $[0,1]$ is called **monotone** if $\lambda(A \cap B) \leq \lambda(A)$. If $\lambda(\emptyset) = 0$ and $\lambda(\Omega) = 1$, we also call it a **probability measure on the π-system**. If $\mathcal{I}$ is a π-system, let $\sigma(\mathcal{I})$ denote the smallest σ algebra containing $\mathcal{I}$.

4.3. $A, B \in \mathcal{A}$ are called **independent**, if $\boxed{\mathrm{P}[A \cap B] = \mathrm{P}[A] \cdot \mathrm{P}[B]}$. With the **conditional probability** $\boxed{\mathrm{P}[A|B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}}$ **independence** means $\mathrm{P}[A|B] = \mathrm{P}[A]$.

4.4. Example: A fair dice produces a random number k in $\{1,2,3,4,5,6\}$. Now toss a fair coin k times. You don't see the coins but are told that all coins are head. What is the probability that the dice showed 5? It can be solved using Bayes rule.

4.5. You prove the following ΠΛΣ result in the homework. [1]

Theorem 1 (Sorority)**.** *The smallest λ-system $\mathcal{A}$ containing a π-system $\mathcal{I}$ is $\sigma(\mathcal{I})$.*

4.6. Let $\mathcal{P} = 2^\Omega$ denote the set of all subsets of Ω. A map $\mu : \mathcal{P} \to [0,1]$ is called an **outer measure** if $\mu(\emptyset) = 0$, $A, B \in \mathcal{A}$ with $A \subset B \Rightarrow \mu(A) \leq \mu(B)$ and **σ-subadditivity** holds $A_n \in \mathcal{P} \Rightarrow \mu(\bigcup_n A_n) \leq \sum_n \mu(A_n)$ for all sets. Given an outer measure μ, a set A in $\mathcal{P}$ is a **μ-set**, if $\mu(A \cap G) + \mu(A^c \cap G) = \mu(G)$ for all $G \in \mathcal{P}$. Given an outer measure μ, let $\boxed{\mathcal{A}_\mu}$ be the set of all μ-sets.

Theorem 2. *An outer measure μ defines a σ-algebra $\mathcal{A}_\mu \subset \mathcal{P}$ on which μ is σ-additive.*

Proof. (i) $\mathcal{A}_\mu$ is an algebra. First of all, $\Omega \in \mathcal{A}_\mu$. If $B \in \mathcal{A}_\mu$, then $B^c \in \mathcal{A}_\mu$. Given $B, C \in \mathcal{A}_\mu$. Then $A = B \cap C \in \mathcal{A}_\mu$. Since $C \in \mathcal{A}_\mu$, we get $\mu(C \cap A^c \cap G) + \mu(C^c \cap A^c \cap G) = \mu(A^c \cap G)$. This can be rewritten with $C \cap A^c = C \cap B^c$ and $C^c \cap A^c = C^c$ as $\mu(A^c \cap G) = \mu(C \cap B^c \cap G) + \mu(C^c \cap G)$ Because B is a μ-set, we get using $B \cap C = A$. $\mu(A \cap G) + \mu(B^c \cap C \cap G) = \mu(C \cap G)$. Since C is a μ-set, we have $\mu(C \cap G) + \mu(C^c \cap G) = \mu(G)$. Adding up these three equations shows that $B \cap C$ is a μ-set. If B and C are disjoint in $\mathcal{A}_\mu$ we deduce from the fact that B is a μ-set, $\mu(B \cap (B \cup C) \cap G) + \mu(B^c \cap (B \cup C) \cap G) = \mu((B \cup C) \cap G)$. This can be rewritten as $\mu(B \cap G) + \mu(C \cap G) = \mu((B \cup C) \cap G)$. By induction, $\sum_{k=1}^n \mu(A_k \cap G) = \mu((\bigcup_{k=1}^n A_k) \cap G)$

[1] Never mind that the sorority Pi Lambda Sigma at BU merged with Theta Phi Alpha in 1952.

holds for all $\{A_k\}_{k=1}^n$ and all $G \in \mathcal{A}$.
(ii) Given a disjoint sequence $A_n \in \mathcal{A}_\mu$. We have to show that $A = \bigcup_n A_n \in \mathcal{A}_\mu$ and $\mu(A) = \sum_n \mu(A_n)$. We know that $B_n = \bigcup_{k=1}^n A_k$ is in $\mathcal{A}_\mu$. Because $\mu(G) = \mu(B_n \cap G) + \mu(B_n^c \cap G) \geq \mu(B_n \cap G) + \mu(A^c \cap G) = \sum_{k=1}^n \mu(A_k \cap G) + \mu(A^c \cap G) \geq \mu(A \cap G) + \mu(A^c \cap G)$ and $\mu(G) \leq \mu(A \cap G) + \mu(A^c \cap G)$, we have $\mu(G) = \mu(A \cap G) + \mu(A^c \cap G)$ showing $A \in \mathcal{A}_\mu$. Finally we show that μ is σ-additive on $\mathcal{A}_\mu$: for any $n \geq 1$ we have $\sum_{k=1}^n \mu(A_k) \leq \mu(\bigcup_{k=1}^n A_k \leq \sum_{k=1}^\infty \mu(A_k)$. Take the limit $n \to \infty$. □

Theorem 3 (Carathéodory Extension Theorem)**.** *A probability measure λ on a π-system $\mathcal{I}$ extends uniquely to a probability measure μ on $\sigma(\mathcal{I})$.*

Proof. The function $\mu(A) = \inf\{\sum_{n\in\mathbb{N}} \lambda(A_n) \mid A_n \in \mathcal{I} \text{ with } A \subset \bigcup_n A_n \}$ defines an outer measure on $\mathcal{P}$. We will show that it defines a probability measure on $\mathcal{A}_\mu$. This then extends to the smallest σ-algebra $\mathcal{A}$ containing $\mathcal{I}$.
(i) $\boxed{\mu(\emptyset) = 0 \text{ and } \mu(A) \leq \mu(B) \text{ for } A \subset B}$ follow from the outer measure properties of λ. To see $\boxed{\sigma\text{-sub-additivity of } \mu}$, take a sequence $A_n \in \mathcal{P}$ and fix $\epsilon > 0$. For all $n \in \mathbb{N}$, one can find a sequence $\{B_{n,k}\}_{k\in\mathbb{N}}$ in $\mathcal{I}$ such that $A_n \subset \bigcup_{k\in\mathbb{N}} B_{n,k}$ and $\sum_{k\in\mathbb{N}} \lambda(B_{n,k}) \leq \mu(A_n) + \epsilon 2^{-n}$. Define $A = \bigcup_{n\in\mathbb{N}} A_n \subset \bigcup_{n,k\in\mathbb{N}} B_{n,k}$, so that $\mu(A) \leq \sum_{n,k} \lambda(B_{n,k}) \leq \sum_n \mu(A_n) + \epsilon$. Since ϵ was arbitrary, the σ-subadditivity of μ is proven.
(ii) $\boxed{\lambda = \mu \text{ on } \mathcal{I}}$ Given $A \in \mathcal{I}$. Clearly $\lambda(A) \leq \mu(A)$. Suppose that $A \subset \bigcup_n A_n$, with $A_n \in \mathcal{R}$. Define a sequence $\{B_n\}_{n\in\mathbb{N}}$ of disjoint sets in $\mathcal{R}$ inductively. That is $B_1 = A_1$, $B_n = A_n \cap (\bigcup_{k<n} A_k)^c$ such that $B_n \subset A_n$ and $\bigcup_n B_n = \bigcup_n A_n \supset A$. From the σ-additivity of μ on $\mathcal{I}$ follows $\mu(A) \leq \mu(\bigcup_n A_n) = \mu(\bigcup_n B_n) = \sum_n \mu(B_n)$. Because the choice of A_n is arbitrary, this gives also $\mu(A) \leq \lambda(A)$.
(iii) $\boxed{\mathcal{I} \subset \mathcal{A}_\mu}$. Given $A \in \mathcal{I}$ and $G \in \mathcal{P}$. There exists a sequence $\{B_n\}_{n\in\mathbb{N}}$ in $\mathcal{I}$ such that $G \subset \bigcup_n B_n$ and $\sum_n \mu(B_n) \leq \lambda(G) + \epsilon$. By definition $\sum_n \mu(B_n) = \sum_n \mu(A \cap B_n) + \sum_n \mu(A^c \cap B_n) \geq \mu(A \cap G) + \mu(A^c \cap G)$ because $A \cap G \subset \bigcup_n A \cap B_n$ and $A^c \cap G \subset \bigcup_n A^c \cap B_n$. Since ϵ is arbitrary, we get $\mu(G) \geq \mu(A \cap G) + \lambda(A^c \cap G)$. On the other hand, since μ is sub-additive, we have also $\mu(G) \leq \mu(A \cap G) + \mu(A^c \cap G)$ and A is a μ-set.
(iv) By (i), μ is an outer measure on $\mathcal{P}$. Since by step (iii), $\mathcal{I} \subset \mathcal{A}_\mu$, we know that $\mathcal{A} = \sigma(\mathcal{I}) \subset \mathcal{A}_\mu$, so that μ on $\mathcal{A}$ is defined by restricting μ from $\mathcal{A}_\mu$ to $\mathcal{A} = \sigma(\mathcal{I})$.
(v) $\boxed{\text{Uniqueness.}}$ If two probability measures μ, ν on $\sigma(\mathcal{I})$ satisfy $\mu(A) = \nu(A)$ for $A \in \mathcal{I}$, then $\boxed{\mu = \nu}$: the set $\mathcal{D} = \{A \in \mathcal{A} \mid \mu(A) = \nu(A) \}$ is a λ system: $\Omega \in \mathcal{D}$. Given $A, B \in \mathcal{D}, A \subset B$. Then $\mu(B \setminus A) = \mu(B) - \mu(A) = \nu(B) - \nu(A) = \nu(B \setminus A)$ so that $B \setminus A \in \mathcal{D}$. Given $A_n \in \mathcal{D}$ with $A_n \nearrow A$, then the σ additivity gives $\mu(A) = \limsup_n \mu(A_n) = \limsup_n \nu(A_n) = \nu(A)$, so that $A \in \mathcal{D}$. Since $\mathcal{D}$ is a λ system containing the π-system $\mathcal{I}$, we know that (ask sorority) that $\sigma(\mathcal{I}) =\subset \mathcal{D}$ which means that $\mu = \nu$ on $\sigma(\mathcal{I})$. □

4.7. Examples:
1) $\mu([a, b)) = b - a$ on the π-system $\mathcal{I} = \{[a, b) \subset [0, 1]\}$ extends to $\sigma(I)$.
2) If $(\Omega_1, \mathcal{A}_1, \mathrm{P}_1), (\Omega_2, \mathcal{A}_2, \mathrm{P}_2)$ are probability spaces then $\lambda(A \times B) = \mathrm{P}_1[A]\mathrm{P}_2[B]$ extends from the π system $\mathcal{I}$ of "rectangles" $A \times B$ to the product. Product spaces are a source for independence as $A \times \Omega_2$ and $\Omega_1 \times B$ are always independent.

PROBABILITY THEORY

MATH 154

Unit 5: Random variables

5.1. Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space. A map $X : \Omega \to \mathbb{R}$ is called a **random variable** if all sets $\{X \in [a,b)\}$ are in $\mathcal{A}$ for all $a \leq b$. Random variables are functions which define measurable events. Since we can add and multiply random variables, we get an **algebra** $\mathcal{L}$ of all random variables. It is a vector space over the real numbers $\mathbb{R}$. One could also look at complex-valued random variables or vector-valued random variables like **random matrices**. The Borel σ algebra on the real line generated by half open intervals is denoted by $\mathcal{B}$.

Theorem 1 (Law)**.** *Every random variable $X \in \mathcal{L}$ defines a probability measure $\mu([a,b)) = \mathrm{P}[\{X \in [a,b)\}]$ on the real line $(\mathbb{R}, \mathcal{B})$.*

Proof. This is a direct consequence of the Carathéodory extension theorem: the function $\mu([a,b)) = \mathrm{P}[\{X \in [a,b)\}]$ is a probability measure on the π system of half open intervals. We also have $\mu(\emptyset) = 0$ and $\mu(\mathbb{R}) = 1$. □

5.2. A **step function** is a random variable of the form $X = \sum_{i=1}^n \alpha_i \cdot 1_{A_i}$ with $\alpha_i \in \mathbb{R}$ and where $A_i \in \mathcal{A}$ are disjoint sets. Denote by $\mathcal{S}$ the algebra of step functions. For $X \in \mathcal{S}$, define the **integral**

$$\mathrm{E}[X] := \int_\Omega X \, dP = \sum_{i=1}^n \alpha_i \mathrm{P}[A_i] \; .$$

5.3. Define $\mathcal{L}^1 \subset \mathcal{L}$ as the set of random variables X, for which $\sup_{Y \in \mathcal{S}, Y \leq |X|} \int_\Omega Y \, dP$ is finite. For $X \in \mathcal{L}^1$, define the **integral** or **expectation**

$$\mathrm{E}[X] := \int X \, dP = \sup_{Y \in \mathcal{S}, Y \leq X^+} \int Y \, dP - \sup_{Y \in \mathcal{S}, Y \leq X^-} \int Y \, dP \; ,$$

where $X^+ = X \vee 0 = \max(X, 0)$ and $X^- = -X \vee 0 = \max(-X, 0)$. The vector space $\mathcal{L}^1$ is the space of **integrable random variables**. Similarly, for $p \geq 1$, write $\mathcal{L}^p = \{X \in \mathcal{L}, \mathrm{E}[|X|^p] < \infty\}$.

5.4. For $X \in \mathcal{L}^2$, the **variance** is defined as $\mathrm{Var}[X] := \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$. The non-negative number $\sigma[X] = \mathrm{Var}[X]^{1/2}$ is called the **standard deviation** of X.

5.5. The expectation is also called "average" or "mean". The standard deviation measures how much we can expect the variable to deviate from the mean. Random variables are **centered**, if $\mathrm{E}[X] = 0$. For such X, we can think of σX as a **length**.

5.6. Example: Let $\mathcal{B}$ denote the Lebesgue σ-algebra on $\Omega = [0,1]$. The m'th power random variable $X(x) = x^m$ on $(\Omega, \mathcal{B}, \mathrm{P})$ has expectation $\mathrm{E}[X] = \int_0^1 x^m \, dx = \frac{1}{m+1}$, variance $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \frac{1}{2m+1} - \frac{1}{(m+1)^2} = \frac{m^2}{(1+m)^2(1+2m)}$ and so the standard deviation $\sigma[X] = \frac{m}{(1+m)\sqrt{(1+2m)}}$.

5.7. If X is a random variable, then $\mathrm{E}[X^m]$ is called the m'th **moment** of X. The m'th **central moment** of X is defined as $\mathrm{E}[(X - \mathrm{E}[X])^m]$. The **moment generating function** (MGF) of X is defined as $M_X(t) = \mathrm{E}[e^{tX}]$. It is a tool for a fast simultaneous computation of all the moments. The function

$$\kappa_X(t) = \log(M_X(t))$$

is called the **cumulant generating function**.

5.8. Example: For $X(x) = x$ on $\Omega = [0,1]$, we compute

$$M_X(t) = \mathrm{E}[e^{tX}] = \mathrm{E}[\sum_{m=0}^{\infty} \frac{t^m X^m}{m!}] = \sum_{m=0}^{\infty} t^m \frac{\mathrm{E}[X^m]}{m!} = \sum_{m=0}^{\infty} \frac{t^m}{m!(m+1)} = (e^t - 1)/t \; .$$

5.9. Example. Let $\Omega = \mathbb{R}$. For given constants $m \in \mathbb{R}, \sigma > 0$, define the probability measure $\mathrm{P}[[a,b)] = \int_a^b f(x) \, dx$ with

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \; .$$

Caratheodory assures that this extends to the Lebesgue σ-algebra. This is a probability measure because after a change of variables $y = (x-m)/(\sqrt{2}\sigma)$, the integral $\int_{-\infty}^{\infty} f(x) \, dx$ becomes $\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} \, dy = 1$. The random variable $X(x) = x$ on $(\Omega, \mathcal{A}, \mathrm{P})$ is an example of a random variable with **Gaussian distribution**. We also say it is a **Gaussian random variable** or a random variable with **normal distribution**. Lets justify the constants m and σ: the expectation of X is $\mathrm{E}[X] = \int X \, dP = \int_{-\infty}^{\infty} x f(x) \, dx = m$. The variance is $\mathrm{E}[(X-m)^2] = \int_{-\infty}^{\infty} x^2 f(x) \, dx = \sigma^2$ so that the constant σ is indeed the standard deviation. The moment generating function of X is $M_X(t) = e^{mt + \sigma^2 t^2/2}$. The cumulant generating function is therefore $\kappa_X(t) = mt + \sigma^2 t^2/2$.

5.10. Here comes in some calculus

Theorem 2. *If X has the MGF $M_X(t)$ and $M_X(t)$ has a convergent Taylor series at $t = 0$, then $\mathrm{E}[X^n] = \frac{d^n}{dt^n} M_X(t)|_{t=0}$.*

Proof. Apply the Taylor series to $e^{tX} = \sum_{n=0}^{\infty} t^n X^n / n!$. Taking the n'th derivative on on both sides $X^n e^{tX}_{t=0} = X^n$ and taking the expectation shows that $E[X^n]$ is n'th derivative of $M_X(t)$. □

PROBABILITY THEORY

MATH 154

Unit 6: Distribution Functions

6.1. The **cumulative distribution function** of a random variable X is defined as

$$F_X(s) = \mu((-\infty, s]) = \mathrm{P}[X \leq s] \ .$$

It is often abbreviated as CDF. If $F_X(s)$ is differentiable, it defines the **probability density function** $f_X(s) = F_X'(s)$ abbreviated PDF.

6.2. The function is monotone increasing and satisfies $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$. It does not have to be smooth. For example, if $X(x) = (-1)^x$ is the random variable on $(\Omega = \{1,2,3,4,5,6\}, \mathcal{A} = 2^\Omega, \mathrm{P}(\{x\}) = 1/6)$ then F_X jumps at the values $-1, 1$ and is constant elsewhere. The law of X is the probability measure which is supported on $\{-1, 1\}$ and has weights $1/2$ on both points.

6.3. The distribution function F is useful: To get random variables with a distribution function F, just take a random variable Y with uniform distribution on $[0, 1]$. Now, $X = F^{-1}(Y)$ has the distribution function F because $\{X \in [a, b] \} = \{F^{-1}(Y) \in [a, b] \} = \{Y \in F([a, b]) \} = \{Y \in [F(a), F(b)]\} = F(b) - F(a)$. We need only random number generators that produces uniformly distributed random variables.

6.4. Example. Assume we want to generate a random variable with Cauchy distribution with PDF $f(x) = F'(x) = (\frac{1}{\pi}/(1 + x^2)$. Integrating gives $F(x) = \frac{1}{2} + \arctan(x)/\pi$ and $F^{-1}(x) = \tan(\pi x - \pi/2) = \cot(\pi x)$. We can therefore compute Cauchy distributed random variables by evaluating $\cot(\pi x)$ with uniformly distributed random variables in $[0, 1]$.

6.5. A measure μ on $(\mathbb{R}, \mathcal{B})$ is **absolutely continuous** (ac) with respect to the **Lebesgue measure** $\lambda = dx$ if $\lambda(A) = 0 \Rightarrow \mu(A) = 0, \forall A \in \mathcal{B}$. If $\mu = f d\lambda$ meaning that there exists a non-negative measurable function f satisfying $F(x) = \mu([-\infty, x]) = \int_{-\infty}^x f(x)\ dx$. A measure is **pure point** (pp) if there exists a finite or countable set of real numbers x_n and a sequence of positive numbers $p_n, \sum_n p_n = 1$ with $F(x) = \mu([-\infty, x]) = \sum_{n, x_n \leq x} p_n$. Finally, μ is called **singular continuous** (sc) if μ is continuous ($\mu(\{x\}) = 0$ for all x and $\mu(S) = 1$ for some set S of zero Lebesgue measure. [1]

[1] Many textbooks simply use **continuous** for (ac).

6.6. Nomenclature for μ goes over to CDF's F or random variables X. A (pp) measure μ could be dense on some intervals. It is supported on **atoms**, points x with $\mu(\{x\}) > 0$. For (ac) measures, there is a density function f meaning $\mu = f\ dx$ and $\int_{-\infty}^{\infty} f\ dx = 1$. The existence of this **Radon-Nykodym derivative** is to take the supremum over all non-negative functions with the property $\int_A f d\lambda \leq \mu(A)$ and $\int_{\mathbb{R}} f d\lambda = 1$. A (sc) measure μ has no atoms and satisfies $\mu(S) = 1$ for some $S \in \mathcal{B}$ with $\lambda(S) = 0$.

6.7. These three classes are fundamental:

Theorem 1 (Lebesgue decomposition theorem)**.** *Every probability measure μ on $(\mathbb{R}, \mathcal{B})$ can be decomposed uniquely into $\mu = \mu_{pp} + \mu_{ac} + \mu_{sc}$, where μ_{pp} is pure point, μ_{sc} is singular continuous and μ_{ac} is absolutely continuous.*

Proof. Denote by λ the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ for which $\lambda([a, b]) = b - a$. We first show that any measure μ can be decomposed as $\mu = \mu_{ac} + \mu_s$, where μ_{ac} is absolutely continuous with respect to λ and μ_s is singular. The decomposition is unique: $\mu = \mu_{ac}^{(1)} + \mu_s^{(1)} = \mu_{ac}^{(2)} + \mu_s^{(2)}$ implies that $\mu_{ac}^{(1)} - \mu_{ac}^{(2)} = \mu_s^{(2)} - \mu_s^{(1)}$ is both absolutely continuous and singular with respect to μ which is only possible, if they are zero. To get the existence of the decomposition, define $c = \sup_{A \in \mathcal{A}, \lambda(A)=0} \mu(A)$. If $c = 0$, then μ is absolutely continuous and we are done. If $c > 0$, take an increasing sequence $A_n \in \mathcal{B}$ with $\mu(A_n) \to c$. Define $A = \bigcup_{n \geq 1} A_n$ and μ_s as $\mu_s(B) = \mu(A \cap B)$. To split the singular part μ_s into a singular continuous and pure point part, we again have uniqueness because $\mu_s = \mu_{sc}^{(1)} + \mu_{sc}^{(1)} = \mu_{pp}^{(2)} + \mu_{pp}^{(2)}$ implies that $\nu = \mu_{sc}^{(1)} - \mu_{sc}^{(2)} = \mu_{pp}^{(2)} - \mu_{pp}^{(1)}$ are both singular continuous and pure point which implies that $\nu = 0$. To get existence, define the finite or countable set $A = \{\omega \mid \mu(\omega) > 0\ \}$ and define $\mu_{pp}(B) = \mu(A \cap B)$. □

6.8. Examples of absolutely continuous distributions

- The **normal distribution** $N(m, \sigma^2)$ on $\mathbb{R}$ has PDF $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$.
- The **Cauchy distribution** on $\mathbb{R}$ has $f(x) = \frac{1}{\pi} \frac{b}{b^2 + (x-m)^2}$.
- The **exponential distribution** on $\mathbb{R}^+$ has $f(x) = \lambda e^{-\lambda x}$.

6.9. Example of a singular continuous distribution

- The Cantor distribution on $[0, 1]$ supported on the Cantor middle third. Its CDF is called the **Cantor staircase**.

6.10. Examples of pure point distributions:

- The **binomial distribution** on $\{1, \ldots, n\ \}$ $\mathrm{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$.
- The **Poisson distribution** on $\mathbb{N}$ has $\mathrm{P}[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}$
- The **geometric distribution** on $\mathbb{N} = \{0, 1, 2, 3, \ldots\ \}$ has $\mathrm{P}[X = k] = p(1-p)^k$.

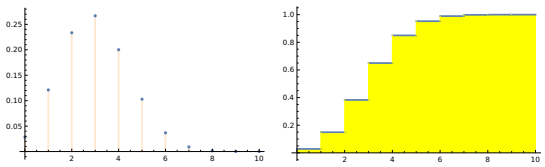


FIGURE 1. The law and CDF of a Binomial distribution.

PROBABILITY THEORY

MATH 154

Unit 7: Independence

7.1. Two events $A, B \in \mathcal{A}$ in a probability space $(\Omega, \mathcal{A}, \mathrm{P})$, are **independent** if $\mathrm{P}[A \cap B] = \mathrm{P}[A]\mathrm{P}[B]$. An arbitrary set of events $\{A_i\}_{i \in I}$ is called **independent**, if for any finite subset J of them, $\mathrm{P}[\bigcap_{j \in J} A_j] = \prod_{j \in J} \mathrm{P}[A_j]$.

7.2. A finite set of subsets $A_1, A_2, \ldots, A_n$ of Ω which are pairwise disjoint and whose union is Ω is called a **finite partition** of Ω. It generates the σ-algebra: $\mathcal{A} = \{A = \bigcup_{j \in J} A_j \}$, where J runs over all subsets of $\{1, \ldots, n\}$. This σ-algebra has 2^n elements. Every finite σ-algebra $\mathcal{A}$ is of this form: just look at the **atoms**, the smallest nonempty elements $\{A_1, \ldots, A_n\}$. They form a disjoint set that cover Ω.

7.3. Two π-systems $\mathcal{I}, \mathcal{J} \subset \mathcal{A}$ are called **independent**, if for all $A \in \mathcal{I}$ and $B \in \mathcal{J}$, $\mathrm{P}[A \cap B] = \mathrm{P}[A] \cdot \mathrm{P}[B]$. Similarly, two σ-algebras $\mathcal{A}, \mathcal{B}$ are called **independent**, if for any pair $A \in \mathcal{A}, B \in \mathcal{B}$, the events A, B are independent. An arbitrary family $\prod_{j \in I} \mathcal{A}_j$ of σ algebras is independent if any finite set $A_j \in \mathcal{A}_j$ of events are independent.

7.4. Examples:
1) On $(\Omega = \{1, 2, 3, 4 \}, 2^\Omega, \mathrm{P}[A] = |A|/|\Omega|)$, the two σ-algebras $\mathcal{A} = \{\emptyset, \{1, 3 \}, \{2, 4 \}, \Omega \}$ and $\mathcal{B} = \{\emptyset, \{1, 2 \}, \{3, 4 \}, \Omega \}$ are independent.
2) For independent sets A, B in a probability space, the sub σ-algebras $\mathcal{A} = \{\emptyset, A, A^c, \Omega\}$ and $\mathcal{B} = \{\emptyset, B, B^c, \Omega\}$ are independent.
3) If $(\Omega_j, \mathcal{A}_j, \mathrm{P}[A_j])$ with $j \in I$ are probability spaces then each factor is independent in the product probability space. of sequences $\Omega = x = \{(x_j)_{j \in I}, x_j \in \Omega_j\}$, where the σ algebra is generated by the π-system of sets $\prod_{j \in J} A_j$ in which only finitely many are not equal to Ω_j and where the measure P is extended from that π system via Carathéodory as $\mathrm{P}[\prod_{j \in J} A_j] = \prod_{j \in J} \mathrm{P}[A_j]$.

7.5. Given a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. Let $\mathcal{G}, \mathcal{H}$ be two σ-sub-algebras of $\mathcal{A}$ and $\mathcal{I}$ and $\mathcal{J}$ be two π-systems satisfying $\sigma(\mathcal{I}) = \mathcal{G}$, $\sigma(\mathcal{J}) = \mathcal{H}$. Then $\mathcal{G}$ and $\mathcal{H}$ are independent if $\mathcal{I}$ and $\mathcal{J}$ are independent. Proof: (i) Fix $I \in \mathcal{I}$ and define on $(\Omega, \mathcal{H})$ the measures $\mu(H) = \mathrm{P}[I \cap H], \nu(H) = \mathrm{P}[I]\mathrm{P}[H]$ of total probability $\mathrm{P}[I]$. By independence of $\mathcal{I}$ and $\mathcal{J}$, they coincide on $\mathcal{I}$ and by the extension, they agree on $\mathcal{H}$ and we have $\mathrm{P}[I \cap H] = \mathrm{P}[I]\mathrm{P}[H]$ for all $I \in \mathcal{I}$ and $H \in \mathcal{H}$.
(ii) Define for fixed $H \in \mathcal{H}$ the measures $\mu(G) = \mathrm{P}[G \cap H]$ and $\nu(G) = \mathrm{P}[G]\mathrm{P}[H]$ of total probability $\mathrm{P}[H]$ on $(\Omega, \mathcal{G})$. They agree on $\mathcal{I}$ and so on $\mathcal{G}$ again by extension. We therefore have $\mathrm{P}[G \cap H] = \mathrm{P}[G]\mathrm{P}[H]$ for all $G \in \mathcal{G}$ and all $H \in \mathcal{H}$.

7.6. A random variable X **generates a σ subalgebra** $\sigma(X)$ of $\mathcal{A}$. It is defined as the smallest σ-algebra that contains all events $A = \{X \in [a,b)\}$. Write $\sigma(X) = X^{-1}(\mathcal{B})$ because $\sigma(X) = \{X^{-1}(B) \mid B \in \mathcal{B}\}$, where $\mathcal{B}$ is the Borel σ algebra on $[0,1]$.

7.7. Examples:
1) A constant map $X(x) = c$ defines the **trivial algebra** $\mathcal{A} = \{\emptyset, \Omega\}$.
2) The projection map $X(x,y) = x$ from the square $(\Omega = [0,1] \times [0,1], \sigma(\mathcal{B} \times \mathcal{B}), \lambda \times \lambda)$ to the real line $\mathbb{R}$ defines the algebra $\mathcal{B} = \{A \times [0,1]\}$, where A is in the Borel σ-algebra $\mathcal{B}$ of the interval $[0,1]$.
3) The map X from $\mathbb{Z}_6 = \{0,1,2,3,4,5\}$ to $\{0,1\} \subset \mathbb{R}$ defined by $X(x) = x \bmod 2$ has the value $X(x) = 0$ if x is even and $X(x) = 1$ if x is odd. The σ-algebra generated by X is $\mathcal{A} = \{\emptyset, \{1,3,5\}, \{0,2,4\}, \Omega\}$.

7.8. Two random variables X, Y are called **independent**, if they generate independent σ-algebras. It is enough to check that the events $A = \{X \in (a,b]\}$ and $B = \{Y \in (c,d]\}$ are independent for all intervals $(a,b]$ and $(c,d]$. Independent random variables as two aspects of the laboratory Ω which do not influence each other. Each event $A = \{a < X(\omega) \leq b\}$ is independent of the event $B = \{c < Y(\omega) \leq d\}$.

7.9. Examples:
1) Throwing a dice 3 times is modeled with a laboratory Ω has $6^3 = 216$ elements, where each experiment is a random vector $x = (x_1, x_2, x_3)$. Now, $X_j(x) = x_j \in \{1,2,3,4,5,6\}$ are independent random variables.
2) In full generality, the random variables $X_j(x) = x_j$ on a product probability space $(\Omega = \prod_j \Omega_j, \mathcal{A} = \prod_j \mathcal{A}_j, \mathrm{P} = \prod_j \mathrm{P}_j)$ are independent.

7.10. If a σ-algebra $\mathcal{F} \subset \mathcal{A}$ is independent to itself, then $\mathrm{P}[A \cap A] = \mathrm{P}[A] = \mathrm{P}[A]^2$ so that for every $A \in \mathcal{F}$, $\mathrm{P}[A] \in \{0,1\}$. Such a σ-algebra is called **P-trivial**.
The trivial algebra $\mathcal{F} = \{\emptyset, \Omega\}$ is P-trivial in any probability space $(\Omega, \mathcal{A}, \mathrm{P}P)$. (See HW). Independence implies zero **covariance** $\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = 0$ and zero correlation $\mathrm{Cor}[X,Y] = \mathrm{Cov}[X,Y]/(\sigma[X]\sigma[Y])$.

Theorem 1. *If X, Y are independent $\mathcal{L}^2$ random variables then* $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Proof. $\mathcal{L}^2$ assures that $E[XY] \leq \sqrt{E[X^2]|E[Y^2]}$ exists (Cauchy-Schwarz). By approximation it is enough to check for step functions $X = \sum_{i=1}^n a_i 1_{A_i}$, $Y = \sum_{j=1}^m b_j 1_{B_j}$ which generate finite independent σ algebras meaning every pair A_i, B_j is independent for i, j. Because $\mathrm{E}[X] = \sum_{i=1}^n a_i \mathrm{P}[A_i]$ and $\mathrm{E}[Y] = \sum_{j=1}^m b_j \mathrm{P}[B_j]$, we have $\mathrm{E}[XY] = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathrm{P}[A_i \cap B_j] = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathrm{P}[A_i]\mathrm{P}[B_j] = (\sum_{i=1}^n a_i \mathrm{P}[A_i])(\sum_{j=1}^m b_j \mathrm{P}[B_j])$. $= \mathrm{E}[X]\mathrm{E}[Y]$. □

7.11. The moment generating function $M_X(t) = \mathrm{E}[e^{Xt}]$ is defined if $X \in \mathcal{L}^\infty$ of essentially bounded random variables. In that case, $e^{Xt} = \sum_{k=0}^\infty X^k t^k / k!$ is in $\mathcal{L}^\infty$. In the HW you show: If X, Y are $\mathcal{L}^\infty$ independent random variables, then X^n, Y^m are independent and e^{tX}, e^{tY} are independent.

Theorem 2. *X, Y are independent in $\mathcal{L}^\infty$ then $M_{X+Y}(t) = M_X(t) M_Y(t)$.*

Proof. $M_{X+Y}(t) = \mathrm{E}[e^{t(X+Y)}] = \mathrm{E}[e^{t(X+Y)}] = \mathrm{E}[e^{tX} e^{tY}] = \mathrm{E}[e^{tX}]\mathrm{E}[e^{tY}] = M_X(t) M_Y(t)$. □

PROBABILITY THEORY

MATH 154

Unit 8: Characteristic functions

8.1. Given $X \in \mathcal{L}$, its **characteristic function** is a complex-valued function on $\mathbb{R}$ defined as $\phi_X(t) = \mathrm{E}[e^{itX}]$. Compare this with the **moment generating function** $M_X(t) = \mathrm{E}[e^{tX}]$. It is important to note that the characteristic function is better behaved because it always exists. The moment generating function needed boundedness. For a Cauchy distributed random variable for example, the moment generating function does not exist as not even $\mathrm{E}[X^2]$ exists. The characteristic function however does exist as $e^{itX} = \cos(tX) + i\sin(tX)$ is a bounded complex-valued random variable.

8.2. If $F = F_X$ is the distribution function of X and $\mu = \mu_X$ is its law, the characteristic function of X is also known as the Fourier-Stieltjes transform because $\phi_X(t) = \int_{\mathbb{R}} e^{itx}\, dF(x) = \int_{\mathbb{R}} e^{itx}\, \mu(dx)$. If F has a derivative f, the PDF, then ϕ_X is called the **Fourier transform** of the density function f_X: $\phi_X(t) = \int_{\mathbb{R}} e^{itx}\, f_X(x)\, dx$.

8.3. Example: For a random variable with density $f_X(x) = x^m/(m+1)$ on $\Omega = [0,1]$ the characteristic function is

$$\phi_X(t) = \int_0^1 e^{itx} x^m\, dx/(m+1) = \frac{m!(1 - e^{it} e_m(-it))}{(-it)^{1+m}(m+1)} ,$$

where $e_n(x) = \sum_{k=0}^n x^k/(k!)$ is the n'th **partial exponential function**.

8.4.

Theorem 1 (Lévy)**.** *The characteristic function ϕ_X determines the distribution of X.*

There are explicit formulas. If a, b are points of continuity of F, then

$$F_X(b) - F_X(a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t)\, dt .$$

If one or both of the end points have mass, then

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t)\, dt = \mu[(a,b)] + \frac{1}{2}\mu[\{a\}] + \frac{1}{2}\mu[\{b\}] .$$

Proof. Because a distribution function F has only countably many points of discontinuities, it is enough to determine $F(b) - F(a)$ in terms of ϕ if a and b are continuity points of F. The verification of the **Lévy formula** is then a computation. For continuous distributions with density $F'_X = f_X$ is the inverse formula for the Fourier transform: $f_X(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ita}\phi_X(t)\, dt$ so that $F_X(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita}}{-it}\phi_X(t)\, dt$. This proves the inversion formula if a and b are points of continuity.

The general formula needs only to be verified when μ is a point measure at the boundary of the interval. By linearity, one can assume μ is located on a single point b with $p = \mathrm{P}[X = b] > 0$. The Fourier transform of the Dirac measure $p\delta_b$ is $\phi_X(t) = pe^{itb}$. The claim reduces to

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it}\, pe^{itb}\, dt = \frac{p}{2}$$

which is equivalent to the claim $\lim_{R\to\infty} \int_{-R}^{R} \frac{e^{itc}-1}{it}\, dt = \pi$ for $c > 0$. Because the imaginary part is zero for every R by symmetry, only

$$\lim_{R\to\infty} \int_{-R}^{R} \frac{\sin(tc)}{t}\, dt = \pi$$

remains. The verification of this integral is a prototype computation in residue calculus. □

We say that a sequence X_n of random variables converges weakly to X if and only if its characteristic functions converge point wise: $\phi_{X_n}(x) \to \phi_X$. Here is a table of characteristic functions (CF) $\phi_X(t) = \mathrm{E}[e^{itX}]$ and moment generating functions (MGF) $M_X(t) = \mathrm{E}[e^{tX}]$ for some familiar random variables:

Distribution	Parameter	CF	MGF
Normal	$m \in \mathbb{R}$, $\sigma^2 > 0$	$e^{mit-\sigma^2t^2/2}$	$e^{mt+\sigma^2t^2/2}$
$N(0,1)$		$e^{-t^2/2}$	$e^{t^2/2}$
Uniform	$[-a,a]$	$\sin(at)/(at)$	$\sinh(at)/(at)$
Exponential	$\lambda > 0$	$\lambda/(\lambda - it)$	$\lambda/(\lambda - t)$
Binomial	$n \geq 1$, $p \in [0,1]$	$(1-p+pe^{it})^n$	$(1-p+pe^t)^n$
Poisson	$\lambda > 0$, λ	$e^{\lambda(e^{it}-1)}$	$e^{\lambda(e^t-1)}$
Geometric	$p \in (0,1)$	$\frac{p}{(1-(1-p)e^{it}}$	$\frac{p}{(1-(1-p)e^t}$
Cauchy	$m \in \mathbb{R}$, $b > 0$	$e^{imt-\lvert t\rvert}$	$e^{mt-\lvert t\rvert}$

Characteristic functions become especially useful, if one deals with independent random variables. Their characteristic functions multiply:

Theorem 2. *Given independent random variables X, Y, then $\phi_X(t)\phi_Y(t) = \phi_{X+Y}(t)$.*

Proof. Since X_j are independent, we get for any set of complex valued continuous functions g_j, for which $\mathrm{E}[g_j(X_j)]$ exists:

$$\mathrm{E}[\prod_{j=1}^{n} g_j(X_j)] = \prod_{j=1}^{n} \mathrm{E}[g_j(X_j)]\ .$$

Proof: This follows almost immediately from the definition of independence since one can check it first for functions $g_j = 1_{A_j}$, where A_j are $\sigma(X_j$ measurable functions for which $g_j(X_j)g_k(X_k) = 1_{A_j \cap A_k}$ and

$$\mathrm{E}[g_j(X_j)g_k(X_k)] = \mathrm{P}[A_j]\mathrm{P}[A_k] = \mathrm{E}[g_j(X_j)]\mathrm{E}[g_k(X_k)]\ ,$$

then for step functions by linearity and then by approximation for arbitrary continuous functions.
If we put $g_j(x) = \exp(ix)$, the proposition is proved. □

PROBABILITY THEORY

MATH 154

Unit 9: Tail algebras

9.1. Given a family $\{\mathcal{A}_i\}_{i\in I}$ of σ-subalgebras of $\mathcal{A}$. For any nonempty set $J \subset I$, let $\mathcal{A}_J := \bigvee_{j\in J} \mathcal{A}_j$ be the σ-algebra generated by $\bigcup_{j\in J} \mathcal{A}_j$. Define also $\mathcal{A}_\emptyset = \{\emptyset, \Omega\}$. The **tail σ-algebra** $\mathcal{T}$ of $\{\mathcal{A}\}_{i\in I}$ is defined as $\mathcal{T} = \bigcap_{J\subset I, J \text{ finite}} \mathcal{A}_{J^c}$, where $\mathcal{J}^c = \mathcal{I} \setminus \mathcal{J}$. Recall that an algebra $\mathcal{A}$ is **trivial** if $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$ for all $A \in \mathcal{A}$. As you have seen in a homework, there are σ algebras with infinitely many elements that are trivial. It does not have to be the smallest possible σ-algebra $\{\Omega, \emptyset\}$ which is trivial as a sub-algebra of any choice of probability space $(\Omega, \mathcal{A}, \mathrm{P})$.

Theorem 1 (Kolmogorov's $0-1$ law)**.** *If $\{\mathcal{A}_i\}_{i\in I}$ are independent σ-algebras, then the tail σ-algebra $\mathcal{T}$ is trivial.*

Proof. (i) Assume $F, G \subset I$ are disjoint sets. Then $\mathcal{A}_F$ and $\mathcal{A}_G$ are independent.
Proof. Define for $H \subset I$ the π-system $\mathcal{I}_H = \{A \in \mathcal{A} \mid A = \bigcap_{i\in K} A_i, K \subset_f H, A_i \in \mathcal{A}_i\}$. The π-systems $\mathcal{I}_F$ and $\mathcal{I}_G$ are independent and generate the σ-algebras $\mathcal{A}_F$ and $\mathcal{A}_G$.
(ii) Especially: $\mathcal{A}_J$ is independent of $\mathcal{A}_{J^c}$ for every $J \subset I$.
(iii) $\mathcal{T}$ is independent of $\mathcal{A}_I$.
Proof. $\mathcal{T} = \bigcap_{J\subset_f I} \mathcal{A}_{J^c}$ is independent of any $\mathcal{A}_K$ for $K \subset_f I$. It is therefore independent to the π-system $\mathcal{I}_I$ which generates $\mathcal{A}_I$.
(iv) $\mathcal{T}$ is a sub-σ-algebra of $\mathcal{A}_I$. Therefore $\mathcal{T}$ is independent of itself which implies that it is P-trivial. □

9.2. Example: Let X_n be a sequence of independent random variables and let $A = \{\omega \in \Omega \mid \sum_{n=1}^\infty X_n \text{ converges}\}$. Then $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$. Proof. Because $\sum_{n=1}^\infty X_n$ converges if and only if $Y_n = \sum_{k=n}^\infty X_k$ converges, we have $A \in \sigma(A_n, A_{n+1} \dots)$. Therefore, A is in $\mathcal{T}$, the tail σ- algebra defined by the independent σ-algebras $\mathcal{A}_n = \sigma(X_n)$. If for example, if X_n takes values $\pm 1/n$, each with probability $1/2$, then $\mathrm{P}[A] = 0$. If X_n takes values $\pm 1/n^2$ each with probability $1/2$, then $\mathrm{P}[A] = 1$. The decision whether $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$ is related to the convergence or divergence of a series. This will be discussed later again in the context of limit theorems.

9.3. Example: Let $\{A_n\}_{n\in\mathbb{N}}$ be a sequence of subsets of Ω. The set $A_\infty := \limsup_{n\to\infty} A_n = \bigcap_{m=1}^\infty \bigcup_{n\geq m} A_n$ consists of the set $\{\omega \in \Omega\}$ such that $\omega \in A_n$ for infinitely many $n \in \mathbb{N}$. The set A_∞ is contained in the tail σ-algebra of $\mathcal{A}_n = \{\emptyset, A_n, A_n^c, \Omega\}$. It follows from Kolmogorov's $0-1$ law that $\mathrm{P}[A_\infty] \in \{0, 1\}$ if $A_n \in \mathcal{A}$ and $\{A_n\}$ are P-independent.

Theorem 2 (Borel-Cantelli Lemma)**.** *Take any sequence $A_n \in \mathcal{A}$.*
a) $\sum_{n\in\mathbb{N}} \mathrm{P}[A_n] < \infty \Rightarrow \mathrm{P}[A_\infty] = 0$ always holds.
b) $\sum_{n\in\mathbb{N}} \mathrm{P}[A_n] = \infty \Rightarrow \mathrm{P}[A_\infty] = 1$, if A_n are independent.

Proof. a) $\mathrm{P}[A_\infty] = \lim_{n\to\infty} \mathrm{P}[\bigcup_{k\geq n} A_k] \leq \lim_{n\to\infty} \sum_{k\geq n} \mathrm{P}[A_k] = 0$.
b) For every integer $n \in \mathbb{N}$,

$$\mathrm{P}[\bigcap_{k\geq n} A_k^c] = \prod_{k\geq n} \mathrm{P}[A_k^c] = \prod_{k\geq n} (1 - \mathrm{P}[A_k]) \leq \prod_{k\geq n} e^{-\mathrm{P}[A_k]} = e^{-\sum_{k\geq n} \mathrm{P}[A_k]} .$$

From

$$\mathrm{P}[A_\infty^c] = \mathrm{P}[\bigcup_{n\in\mathbb{N}} \bigcap_{k\geq n} A_k^c] \leq \sum_{n\in\mathbb{N}} \mathrm{P}[\bigcap_{k\geq n} A_k^c] = 0$$

follows $\mathrm{P}[A_\infty^c] = 0$. □

9.4. The following example illustrates that independence is necessary in the part b) of the Borel-Cantelli lemma: take the probability space $([0,1], \mathcal{B}, \mathrm{P})$, where $\mathrm{P} = \lambda$ is the Lebesgue measure on the Borel σ-algebra $\mathcal{B}$ of $[0,1]$. For $A_n = [0, 1/n]$ we get $A_\infty = \emptyset$ and so $\mathrm{P}[A_\infty] = 0$. But because $\mathrm{P}[A_n] = 1/n$ we have $\sum_{n=1}^\infty \mathrm{P}[A_n] = \sum_{n=1}^\infty \frac{1}{n} = \infty$ because the **harmonic series** $\sum_{n=1}^\infty 1/n$ diverges: $\sum_{n=1}^R \frac{1}{n} \geq \int_1^R \frac{1}{x}\, dx = \log(R)$.

9.5. Writing a novel amounts to enter a sequence of N symbols into a computer. For "Hamlet", Shakespeare had to enter $N = 180'000$ characters. Pop-culture [1] imagines a monkey typing randomly for an indefinite time, producing a random text. Call A_n the event that Monkey types Hamlet on the interval $[nN, , ...nN + N]$. These sets A_n are all independent and have probability 26^{-N}. Since $\sum_n A_n = \infty$ Borel-Cantelli assures that the even appears infinitely often Reality produces constraints like that monkeys like humans live less than $4 * 10^9$ seconds but mathematicians do not care about such things. Their ideas live for ever!

9.6. A nice application of Borel-Cantelli are **percolation problems**. If we take an infinite connected network = graph and delete each bond=edge randomly with probability p, then there will be a threshold p_c such that for $p > p_c$ the network has no infinite cluster and for $p < p_c$ there is an infinite cluster. The event "there is an infinite cluster" is in the tail σ algebra of a set of σ algebras $\mathcal{A}_e = \{0, 1, N_e, N_e^c\}$, where N_e are the set of networks for which bond e is active and N_e^c the set of networks for which edge e is broken. The index set $J = E$ enumerates all the edges of the network and the tail σ-algebra $\mathcal{T}$ **consists of all events that do not change if a finite part of the network is altered.**

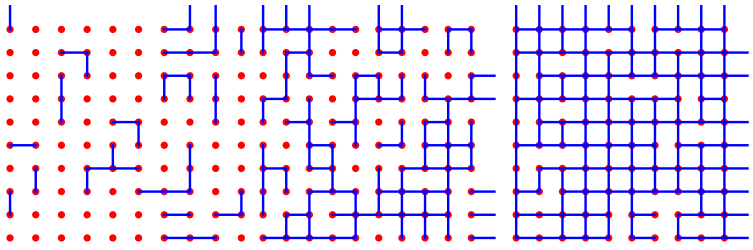


FIGURE 1. The bond percolation threshold in dimension 2 is known to be $p = 1/2$. We see random lattice networks with $p = 0.2, p = 0.5, p = 0.8$.

[1] First appearing in Feller's book from 1950

PROBABILITY THEORY

MATH 154

Unit 10: Problem Seminar

10.1. Finite probability taps into **combinatorics**. Here are some examples:

$n!$	There are $5! = 120$ possible ways to redistribute 5 coats to 5 people.
$\frac{n!}{n_1!\cdots n_k!}$	With $\{A, A, B, B, B, B, E, U, U, U\}$, form $10!/(2!4!1!3!)$ ten letter words.
$\frac{n!}{(n-k)!}$	10 people can sit $10!/6! = 10 * 9 * 8 * 7$ possible ways on 4 chairs.
n^k	There are 6^{10} possible ways to throw 10 dices.
$\frac{n!}{(n-k)!k!}$	There are $52!/(5!47!)$ hands of 5 cards a deck of 52.

Problem 1: a) You pick 7 cards at random from a deck of 81 cards of the game "set". What is the probability that all of them are red (27 are red)?
b) You throw a dice 7 times. What is the probability that all 7 numbers are even?
c) Your gym lock consist of 3 different numbers from 1 to 40. Having forgotten the number, you try a random combinations. What is the probability to open it?
d) How many bijective functions $X \to X$ are there on $X = \{1, 2, 3, 4, 5, 6, 7\}$?
e) How many functions $X \to Y$ are there from $X = \{1, 2, 3, 4, 5, 6, 7\}$ to $\{1, 2\}$?

10.2. The **laboratory** Ω is a set of experiments. The **σ-algebra $\mathcal{A}$** consists of events. A σ-algebra is a Boolean algebra which allows to perform **countably** many operations.

$A \cdot B = A \cap B = \{\omega \in \Omega \mid \omega \in A \text{ and } \omega \in B\}$	"Both events A and B happen"
$A \cup B = \{\omega \in \Omega \mid \omega \in A \text{ or } \omega \in B\}$	"Either A or B happens"
$A + B = A\Delta B = \{\omega \in \Omega \mid \omega \in A \text{ xor } \omega \in B\}$	"One of the events A or B happens"
$A \setminus B = \{\omega \in \Omega \mid \omega \in A \text{ but not } \omega \in B\}$	"A but not B happens"
$A^c = \{\omega \in \Omega \mid \omega \notin A\}$	"A does not happen"
$\bigcap_n A_n = \{\omega \in \Omega \mid \omega \in A_n, \text{ for all } n\}$	"All events A_n happen"
$\bigcup_n A_n = \{\omega \in \Omega \mid \omega \in A_n, \text{ for at least one } n\}$	"At least one event A_n happens"

Problem 2: We use the notation $A \cdot B = A \cap B$ and $A + B = A\Delta B$ and $1 = \Omega$ and $0 = \emptyset$ in the Boolean algebra $\mathcal{P} = 2^\Omega$ of all subsets of Ω.
a) Draw the Venn diagram picture proving that $A(B - C) = AB - AC$.
b) Simplify $(5A + 2)(3A^2 + A - 1)$.
c) Write $A^n = A \cdot A \cdot A \cdots A$ for the n'th power. Simplify $(A - 1)(A + A^2 + A^3)$.
d) Why is $(1 + A)^3 = 1 + A$?
e) Show that $A \cup B = A + B + AB$.

10.3. If B has positive probability, then $\mathrm{P}[A|B] = \mathrm{P}[A \cap B]/\mathrm{P}[B]$ is called the **conditional probability** of A under the condition that event B takes place.

Problem 3: a) If the probability that a student is sick at a given day is 1 percent and the probability that a student has an exam at a given day is 5 percent. Suppose that 6 percent of the students with exams are ill. What is the probability that an ill student has an exam on a given day?
b) Suppose that A, B are subsets of a sample space with a probability function P. We know that $\mathrm{P}[A] = 4/5$ and $\mathrm{P}[B] = 3/5$. Explain why $\mathrm{P}[B|A]$ is at least $1/2$.

10.4. The linear space $\mathcal{L}^2$ has an **inner product** $\mathcal{X} \cdot \mathcal{Y} = \mathrm{E}[XY]$ and so a **length** $|X| = \sqrt{X \cdot X}$. The **standard deviation** of X is the length of **centered random** variable $X - \mathrm{E}[X]$. The **correlation** $-1 \leq \mathrm{Cov}[X, Y]/(\sigma[X]\sigma[Y]) \leq 1$ is $\cos(\alpha)$ and defines an **angle** α between $X - \mathrm{E}[X]$ and $Y - \mathrm{E}[Y]$. If X takes finitely many values (which means $X \in \mathcal{S}$), then $\mathrm{E}[X^n] = \sum_{x \in X(\Omega)} x^n \mathrm{P}[X = x]$. For $X \in \mathcal{L}^n$ with a PDF f, then $\mathrm{E}[X^n] = \int_\infty^\infty x^n f(x)\, dx$. In the box, c, λ are constants.

$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$	$\mathrm{E}[\lambda X] = \lambda \mathrm{E}[X]$
$X \leq Y \Rightarrow \mathrm{E}[X] \leq \mathrm{E}[Y]$	$\mathrm{E}[X^2] = 0 \Leftrightarrow X = 0$
$\mathrm{E}[X] = c$ if $X(\omega) = c$	$\mathrm{E}[X - \mathrm{E}[X]] = 0$
$\mathrm{Var}[X] \geq 0$	$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$
$\mathrm{Var}[\lambda X] = \lambda^2 \mathrm{Var}[X]$	$\mathrm{Cov}[X, X] = \mathrm{Var}[X]$

Problem 4: Let $([-\pi, \pi], \mathcal{B}, \mathrm{P})$ be the Lebesgue probability space, where $\mathrm{P}[[a, b)] = (b-a)/(2\pi)$ on the π system of all half open intervals on $\Omega = [-\pi, \pi]$.
a) Which theorem assures that the measure P exists?
b) Let $X(x) = \sin(3x), Y(x) = \cos(3x)$. Compute the $\mathrm{E}[X], \mathrm{E}[Y], \sigma[X], \sigma[Y]$.
c) What is the correlation $\mathrm{Cor}[X, Y] = \mathrm{Cov}[X, Y]/\sigma(X)\sigma(Y)$? Are X, Y independent?

10.5. Assume X has a probability density $F' = f$ then $\mathrm{E}[X^n] = \int x^n f(x)\, dx$ and $\mathrm{E}[e^{tX}] = \int e^{tx} f(x)\, dx$, $\mathrm{E}[e^{itX}] = \int e^{itx} f(x)\, dx$. Now form $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$ etc.

Problem 5: The PDF $f(x) = \frac{2}{\pi\sqrt{1-x^2}}$ is supported on $[-1, 1]$.
a) Compute the cumulative distribution function $F(x) = \int_{-\infty}^x f(t)\, dt = \int_{-1}^x f(t)\, dt$.
b) Write down the integral for the moment generating function $M_X(t)$.
c) Express the variance in terms of $M_X'(0)$ and $M_X''(0)$.
d) Relate $M_X(t)$ and $M_Y(t)$ and $M_{X+Y}(t)$ for independent random variables! State the same law for the characteristic function $\phi_X(t), \phi_Y(t)$.

10.6. Assume X is a random variable taking a finite or countable number of values $P[X = x_k] = p_k$. Then $\mathrm{E}[X^n] = \sum_k x_k^n p_k$, $\mathrm{E}[e^{tX}] = \sum_k e^{tx_k} p_k$ and $\mathrm{E}[e^{itX}] = \sum_k e^{itx_k} p_k$.

Problem 6: Assume X is a random variable that takes the value 3 with probability $1/3$ and the value 6 with probability $2/3$.
a) Find the expectation $m = \mathrm{E}[X]$ and $n = \mathrm{E}[X^2]$.
b) Are X, X^2 independent? Explain.
c) Write down the characteristic function.

PROBABILITY THEORY

MATH 154

Unit 11: Jensen-Hölder-Minkowski

11.1. A continuous function $h : \mathbb{R} \to \mathbb{R}$ is called **convex**, if there exists for all $x_0 \in \mathbb{R}$ a linear map $l(x) = ax + b$ such that $l(x_0) = h(x_0)$ and for all $x \in \mathbb{R}$ the inequality $l(x) \leq h(x)$ holds.

11.2. Examples:
a) A linear function $h(x) = ax + b$ is convex, b) $h(x) = x^2$ is convex, c) $h(x) = e^x$ is convex, d) $h(x) = -x^2$ is not convex, e) $h(x) = -\log(x)$ is convex on $(0, \infty)$.

Theorem 1 (Jensen inequality)**.** *For $X \in \mathcal{L}^1$ and h convex,* $\mathrm{E}[h(X)] \geq h(\mathrm{E}[X])$.

Proof. Fix $x_0 = \mathrm{E}[X]$. By definition of convexity, there is a linear function $l(x)$, producing a lower bound for h at x_0 meaning $l(x) \leq h(x)$. By the linearity and monotonicity of expectation, we get $h(\mathrm{E}[X]) = l(\mathrm{E}[X]) = \mathrm{E}[l(X)] \leq \mathrm{E}[h(X)]$. If $h(X)$ should not be in $\mathcal{L}$, the statement still holds with the understanding $\mathrm{E}[h(X)] = \infty$. □

FIGURE 1. A convex function

11.3. For $h(x) = x^2$, this gives $\mathrm{E}[X^2] \geq E[X]^2$ which is equivalent to $\mathrm{Var}[X] = \mathrm{E}[X^2] - E[X]^2 \geq 0$. That the variance is non-negative was already clear from $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$.

11.4. For $h(x) = -\log(x)$ and a two point probability space $\Omega = \{0, 1\}$ with a random variable X satisfying $\mathrm{P}[X = \{0\}] = x, \mathrm{P}[X = \{1\}] = y$ and $\mathrm{P}[A] = |A|/2$, we get the **geometric-arithmetic inequality** $\sqrt{xy} \leq (x + y)/2$.

11.5. Given $p \leq q$. Define $h(x) = |x|^{q/p}$. It is convex. Jensen's inequality gives $\mathrm{E}[|X|^q] = \mathrm{E}[h(|X|^p)] \geq h(\mathrm{E}[|X|^p]) = \mathrm{E}[|X|^p]^{q/p}$. This implies that $||X||_q := \mathrm{E}[|X|^q]^{1/q} \geq \mathrm{E}[|X|^p]^{1/p} = ||X||_p$ for $p \leq q$ and so

Theorem 2 (L^p stratification)**.**

$$\mathcal{L}^\infty \subset \mathcal{L}^q \subset \mathcal{L}^p \subset \mathcal{L}^1$$

for $p \leq q$.

The smallest space is $\mathcal{L}^\infty$ which is the space of all bounded random variables.

11.6. The Jensen inequality in the case $\Omega = \{u, v\}$, $\mathrm{P}[\{u\}] = \mathrm{P}[\{v\}] = 1/2$ and with $X(u) = a, X(v) = b$. The function h in this picture is a quadratic function of the form $h(x) = (x - s)^2 + t$.

Theorem 3 (Hölder inequality)**.** *Given* $p, q \in [1, \infty]$ *with* $p^{-1} + q^{-1} = 1$ *and* $X \in \mathcal{L}^p$ *and* $Y \in \mathcal{L}^q$*. Then* $XY \in \mathcal{L}^1$ *and* $||XY||_1 \leq ||X||_p ||Y||_q$.

Proof. The random variables X, Y are defined over a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. The identity $p^{-1} + q^{-1} = 1$ is equivalent to $q + p = pq$ or $q(p - 1) = p$.
Without loss of generality we can restrict us to $X, Y \geq 0$ because replacing X with $|X|$ and Y with $|Y|$ does not change anything. We can also assume $||X||_p > 0$ because otherwise $X = 0$, where both sides are zero. We can write therefore X instead of $|X|$ and assume X is not zero. The key idea of the proof is to introduce a new probability measure

$$\mathrm{Q} = \frac{X^p \mathrm{P}}{\mathrm{E}[X^p]} .$$

If $\mathrm{P}[A] = \int_A 1 d\mathrm{P}(x)$ then $\mathrm{Q}[A] = [\int_A X^p(x) d\mathrm{P}(x)]/\mathrm{E}[X^p]$ so that $\mathrm{Q}[\Omega] = \mathrm{E}[X^p]/\mathrm{E}[X^p] = 1$ and Q is a probability measure. Let E_Q denote the expectation with respect to this new measure. Define the new random variable $U = 1_{\{X>0\}} Y/X^{p-1}$. Jensen's inequality applied to the convex function $h(x) = x^q$ gives

$$\mathrm{E}_\mathrm{Q}[U]^q \leq \mathrm{E}_\mathrm{Q}[U^q] . \tag{1}$$

Using $\mathrm{E}_\mathrm{Q}[U] = \mathrm{E}_\mathrm{Q}[\frac{Y}{X^{p-1}}] = \frac{\mathrm{E}[XY]}{\mathrm{E}[X^p]}$ and $\mathrm{E}_\mathrm{Q}[U^q] = \mathrm{E}_\mathrm{Q}[\frac{Y^q}{X^{q(p-1)}}] = \mathrm{E}_\mathrm{Q}[\frac{Y^q}{X^p}] = \frac{\mathrm{E}[Y^q]}{\mathrm{E}[X^p]}$, Equation (1) can be rewritten as $\frac{\mathrm{E}[XY]^q}{\mathrm{E}[X^p]^q} \leq \frac{\mathrm{E}[Y^q]}{\mathrm{E}[X^p]}$ which implies

$$\mathrm{E}[XY] \leq \mathrm{E}[Y^q]^{1/q} \mathrm{E}[X^p]^{1-1/q} = \mathrm{E}[Y^q]^{1/q} \mathrm{E}[X^p]^{1/p} .$$

This is equivalent to $||XY||_1 \leq ||X||_p ||Y||_q$. □

11.7. A special case of the Hölder's inequality is the **Cauchy-Schwarz** inequality

$$||XY||_1 \leq ||X||_2 \cdot ||Y||_2 .$$

The semi-norm property of $\mathcal{L}^p$ follows from the following fact:

Theorem 4 (Minkowski inequality (1896))**.** *Given* $p \in [1, \infty]$ *and* $X, Y \in \mathcal{L}^p$*. Then*

$$||X + Y||_p \leq ||X||_p + ||Y||_p .$$

Proof. We use Hölder's inequality from below to get

$$\mathrm{E}[|X + Y|^p] \leq \mathrm{E}[|X||X + Y|^{p-1}] + \mathrm{E}[|Y||X + Y|^{p-1}] \leq ||X||_p C + ||Y||_p C ,$$

where $C = |||X + Y|^{p-1}||_q = \mathrm{E}[|X + Y|^p]^{1/q}$ which leads to the claim. □

PROBABILITY THEORY

MATH 154

Unit 12: Chebyshev-Markov-Chernoff-Gibbs

Chebyshev

12.1. We use the short-hand notation $\mathrm{P}[X \geq c]$ for $\mathrm{P}[\{\omega \in \Omega \mid X(\omega) \geq c\}]$.

Theorem 1 (Chebyshev-Markov inequality)**.** *Let h be a monotone function on $\mathbb{R}$ with $h \geq 0$. For every $c > 0$, and $h(X) \in \mathcal{L}^1$ we have*

$$h(c) \cdot \mathrm{P}[X \geq c] \leq \mathrm{E}[h(X)] \ .$$

Proof. Integrate the inequality $h(c)1_{X \geq c} \leq h(X)$ and use the monotonicity and linearity of the expectation. □

12.2. $h(x) = |x|$ leads to $\mathrm{P}[|X| \geq c] \leq ||X||_1/c$ which implies for example the statement

$$\mathrm{E}[|X|] = 0 \Rightarrow \mathrm{P}[X = 0] = 1 \ .$$

12.3. For $X \in \mathcal{L}^\infty$ we had used the moment generating function $M_X(t) = \mathrm{E}[e^{Xt}]$. For every t we have

$$e^{tc}\mathrm{P}[X \geq c] \leq M_X(t) \ .$$

This gives the

Theorem 2 (Chernoff bound)**.**

$$\mathrm{P}[X \geq c] \leq \inf_{t \geq 0} e^{-tc} M_X(t) \ .$$

12.4. An important case of the Chebyshev-Markov inequality is the **Chebyshev inequality**:

Theorem 3 (Chebyshev inequality)**.** *If $X \in \mathcal{L}^2$, then*

$$\mathrm{P}[|X - \mathrm{E}[X]| \geq c] \leq \frac{\mathrm{Var}[X]}{c^2} \ .$$

Proof. Take $h(x) = x^2$ and apply the Chebyshev-Markov inequality to the random variable $Y = X - \mathrm{E}[X]$ which is in $\mathcal{L}^1$ because $X \in \mathcal{L}^2$. □

Entropy

12.5. If $(\Omega, \mathcal{A}, \mathrm{P})$ is a probability space and $X \in \mathcal{S}$, $X = \sum_{k=1}^n x_k 1_{A_k}$ is a random variable taking finitely many values x_k, its **entropy** is defined as $\sum_k \log(\frac{1}{p_k})p_k$, where $p_k = P[X = x_k]$ and $\log(\frac{1}{p_k})p_k = 0$ if $p_k = 0$. It has been introduced by Ludwig Boltzmann in statistical mechanics ("$S = k \log[W]$") and Claude Shannon in information theory.

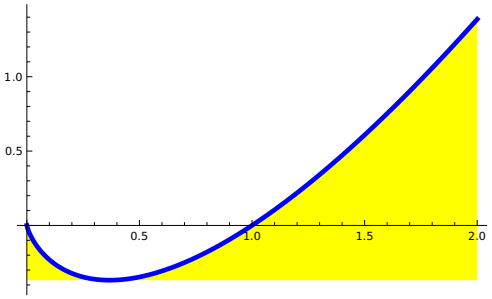


FIGURE 1. The convex function $x \log(x)$.

12.6. The entropy of a finite sub-algebra $\mathcal{B}$ of $\mathcal{A}$ is defined as $\sum_{A \in \mathcal{B}} \mathrm{P}[A] \log(\frac{1}{\mathrm{P}[A]})$. This is well defined, if we understand that $\mathrm{P}[A] \log(\frac{1}{\mathrm{P}[A]}) = 0$ for $\mathrm{P}[A] = 0$. It is a common assumption to extend the convex function $h(x) = x \log(x)$ to $x = 0$ by setting it to $h(0) = 0$.

12.7. A random variable $X \in \mathcal{S}$ defines a finite probability distribution. We just look at the sequence p_k of probabilities. The **relative entropy** of two such distributions is defined as

$$D[p, q] = \sum_k p_k \log(\frac{p_k}{q_k}) \ ,$$

We can rewrite this as $S(p, q) - S(p)$, where

$$S(p, q) = \sum_k p_k \log(\frac{1}{q_k})$$

is the **cross entropy**.

12.8. The relative entropy is also known as the **Kullback-Leibler divergence**. In nice cases, this goes over to more general random variables like continuous distributions where entropy is $S(f) = \int f(x) \log(\frac{1}{f(x)}) \, dx$ and $D[f, g] = \int f(x) \log(\frac{f(x)}{g(x)}) \, dx$. [1]

Theorem 4 (Gibbs inquality)**.** *The relative entropy is non-negative:* $D[p, q] \geq 0$.

Proof. Since $-\log$ is convex, we get from the Jensen inequality $D(p, q) = -\sum_k p_k \log(\frac{q_k}{p_k}) \geq -\log(\sum_k p_k \frac{q_k}{p_k}) = \log(\sum_k q_k) = \log(1) = 0$. □

12.9. For $p_k = 1/n$, the uniform distribution, then entropy is $\log(n)$. It is an multivariable calculus verification that $0 \leq S(p) \leq \log(n)$, where $S(p) = 0$ in the deterministic case where only one $p_k = 1$ and the others are zero. The uniform distribution has maximal entropy.

12.10. Remark. Here is an experimental observation for which I do not know the answer. Pick a large n and random p_k a random permutation $\pi \in S_n$ and $q_k = p_{\pi(k)}$, then the Kullback-Leibler divergence $D(p, q)/n \sim 1$.

OLIVER KNILL, KNILL@MATH.HARVARD.EDU, MATH 154, SPRING, 2025

[1] Entropy and relative entropy is not defined for some perfectly nice bounded smooth distributions!

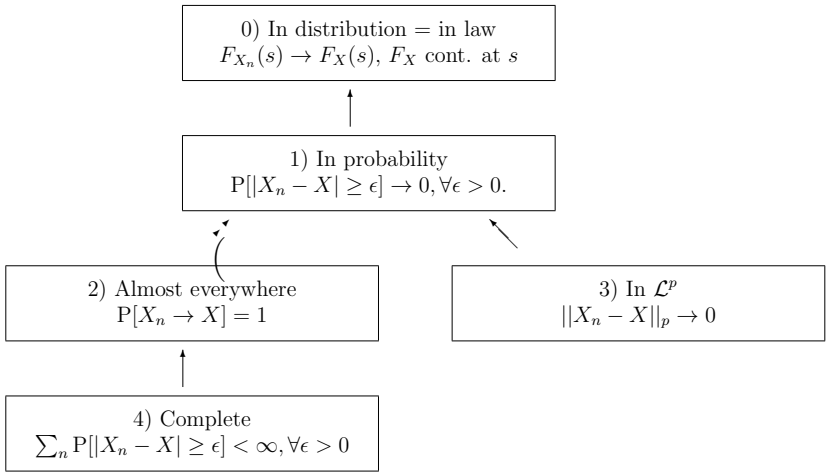
PROBABILITY THEORY

MATH 154

Unit 13: Stochastic convergence

13.1. Given a sequence of random variables X_n on $(\Omega, \mathcal{A}, \mathrm{P})$. We say X_n converges **in probability** to X, if $\mathrm{P}[|X_n - X| \geq \epsilon] \to 0$ for all $\epsilon > 0$. We say X_n converges **almost everywhere** or **almost surely** to X if $\mathrm{P}[X_n \to X] = 1$. We say X_n **converges in $\mathcal{L}^p$** to X, if $||X_n - X||_p \to 0$ for $n \to \infty$. Finally, X_n converges **completely** if $\sum_n \mathrm{P}[|X_n - X| \geq \epsilon] < \infty$ for all $\epsilon > 0$.

13.2. A sequence X_n of random variables X_n on $(\Omega_n, \mathcal{A}_n, \mathrm{P}_n)$ converges **in distribution**, if $F_{X_n}(s) \to F_X(s)$ at all continuity points s of F_X. We say X_n converges **in law** to X, if the laws μ_n of X_n converge weakly to the law μ of X meaning that for every continuous function f on $\mathbb{R}$ of compact support, one has $\int f(x)\, d\mu_n(x) \to \int f(x)\, d\mu(x)$.



Proof. $\boxed{2) \Rightarrow 1)}$: $\{X_n \to X\} = \bigcap_k \bigcup_m \bigcap_{n \geq m} \{|X_n - X| \leq 1/k\}$ is equivalent to $1 = \mathrm{P}[\bigcup_m \bigcap_{n \geq m} \{|X_n - X| \leq \frac{1}{k}\}] = \lim_{m \to \infty} \mathrm{P}[\bigcap_{n \geq m} \{|X_n - X| \leq \frac{1}{k}\}]$ for all k and so $0 = \lim_{n \to \infty} \mathrm{P}[\bigcup_{n \geq m} \{|X_n - X| \geq \frac{1}{k}\}]$ for all k. Therefore $\mathrm{P}[|X_m - X| \geq \epsilon] \leq \mathrm{P}[\bigcup_{n \geq m} \{|X_n - X| \geq \epsilon\}] \to 0$ for all $\epsilon > 0$. $\boxed{4) \Rightarrow 2)}$: The first Borel-Cantelli lemma implies that for all $\epsilon > 0$ $\mathrm{P}[\{|X_n - X| \geq \epsilon, \text{ infinitely often}\}] = 0$. We get so for $\epsilon_k \to 0$ the relation $\mathrm{P}[B_k] = \mathrm{P}[\bigcup_n \{|X_n - X| \geq \epsilon_k, \text{ infinitely often}\}] \leq \sum_n \mathrm{P}[\{|X_n - X| \geq \epsilon_k, \text{ infinitely often}\}] = 0$ and $\bigcup_k B_k$ has measure zero. Now $\mathrm{P}[X_n \to X] = 1 - \mathrm{P}[\bigcup_k B_k] = 1 - 0 = 1$. $\boxed{3) \Rightarrow 1)}$: Chebychev-Markov implies $\mathrm{P}[|X_n - X| \geq \epsilon] \leq \frac{\mathrm{E}[|X_n - X|^p]}{\epsilon^p}$. $\boxed{1) \Rightarrow 0)}$: $\mathrm{P}[X_n \leq c] \leq \mathrm{P}[X \leq c + \epsilon] + \mathrm{P}[|X_n - X| > \epsilon]$. □

Theorem 1. *Given $X_n \in \mathcal{L}^\infty$ with $||X_n||_\infty \leq K$ for all n, then $X_n \to X$ in probability if and only if $X_n \to X$ in $\mathcal{L}^1$.*

Proof. (i) For $k \in \mathbb{N}$, $\mathrm{P}[|X| > K + \frac{1}{k}] \leq \mathrm{P}[|X - X_n| > \frac{1}{k}] \to 0, n \to \infty$ so that $\mathrm{P}[|X| > K + \frac{1}{k}] = 0$. Therefore $\mathrm{P}[|X| > K] = \mathrm{P}[\bigcup_k \{|X| > K + \frac{1}{k}\}] = 0$. (ii) Given $\epsilon > 0$. Choose m such that $\mathrm{P}[|X_n - X| > \frac{\epsilon}{3}] < \frac{\epsilon}{3K}$ for all $n > m$. Use the notation $\mathrm{E}[X; A] = \mathrm{E}[X \cdot 1_A]$. By (i) we have $\mathrm{E}[|X_n - X|] = \mathrm{E}[(|X_n - X|; |X_n - X| > \frac{\epsilon}{3}] + \mathrm{E}[(|X_n - X|; |X_n - X| \leq \frac{\epsilon}{3}] \leq 2K\mathrm{P}[|X_n - X| > \frac{\epsilon}{3}] + \frac{\epsilon}{3} \leq \epsilon$. □

13.3. A family $\mathcal{C} \subset \mathcal{L}^1$ of random variables is called **uniformly integrable**, if

$$\lim_{R \to \infty} \sup_{X \in \mathcal{C}} \mathrm{E}[X; |X| > R] = 0$$

for all $X \in \mathcal{C}$ still using notation $\mathrm{E}[X; A] = \mathrm{E}[X 1_A]$.

Theorem 2. *Given $X \in \mathcal{L}^1$ and $\epsilon > 0$. There exists $K \geq 0$ with $\mathrm{E}[|X|; |X| > K] < \epsilon$.*

Proof. Assume we are given $\epsilon > 0$. If $X \in \mathcal{L}^1$, we can find $\delta > 0$ such that if $\mathrm{P}[A] < \delta$, then $\mathrm{E}[|X|; A] < \epsilon$. Since $K\mathrm{P}[|X| > K] \leq \mathrm{E}[|X|]$, we can choose K such that $\mathrm{P}[|X| > K] < \delta$. Therefore $\mathrm{E}[|X|; |X| > K] < \epsilon$. □

13.4. The next proposition gives a necessary and sufficient condition for $\mathcal{L}^1$ convergence.

Theorem 3. *Given $X_n \in \mathcal{L}^1$ and $X \in \mathcal{L}^1$. The following is equivalent:*
a) X_n converges in probability to X and $\{X_n\}_{n \in \mathbb{N}}$ is uniformly integrable.
b) X_n converges in $\mathcal{L}^1$ to X.

Proof. $a) \Rightarrow b)$. For any random variable X and $K \geq 0$ define the bounded variable $X^{(K)} = X \cdot 1_{\{-K \leq X \leq K\}} + K \cdot 1_{\{X > K\}} - K \cdot 1_{\{X < -K\}}$. By the uniform integrability condition and the previous theorem applied to $X^{(K)}$ and X, we can choose K such that for all n, $\mathrm{E}[|X_n^{(K)} - X_n|] < \frac{\epsilon}{3}$, $\mathrm{E}[|X^{(K)} - X|] < \frac{\epsilon}{3}$. Since $|X_n^{(K)} - X^{(K)}| \leq |X_n - X|$, we have $X_n^{(K)} \to X^{(K)}$ in probability. By the last theorem we know $X_n^{(K)} \to X^{(K)}$ in $\mathcal{L}^1$ so that for $n > m$ $\mathrm{E}[|X_n^{(K)} - X^{(K)}|] \leq \epsilon/3$. Therefore, for $n > m$ also

$$\mathrm{E}[|X_n - X|] \leq \mathrm{E}[|X_n - X_n^{(K)}|] + \mathrm{E}[|X_n^{(K)} - X^{(K)}|] + \mathrm{E}[|X^{(K)} - X|] \leq \epsilon \ .$$

$b) \Rightarrow a)$. We have seen already that $X_n \to X$ in probability if $||X_n - X||_1 \to 0$. We have to show that $X_n \to X$ in $\mathcal{L}^1$ implies that X_n is uniformly integrable.
Given $\epsilon > 0$. There exists m such that $\mathrm{E}[|X_n - X|] < \epsilon/2$ for $n > m$. By the absolutely continuity property, we can choose $\delta > 0$ such that $\mathrm{P}[A] < \delta$ implies

$$\mathrm{E}[|X_n|; A] < \epsilon, 1 \leq n \leq m, \mathrm{E}[|X|; A] < \epsilon/2 \ .$$

Because X_n is bounded in $\mathcal{L}^1$, we can choose K such that $K^{-1} \sup_n \mathrm{E}[|X_n|] < \delta$ which implies $\mathrm{P}[|X_n| > K] < \delta$. For $n \geq m$, we have therefore, using the notation $\mathrm{E}[X; A] = \mathrm{E}[X \cdot 1_A]$

$$\mathrm{E}[|X_n|; |X_n| > K] \leq \mathrm{E}[|X|; |X_n| > K] + \mathrm{E}[|X - X_n|] < \epsilon \ .$$

□

PROBABILITY THEORY

MATH 154

Unit 14: Weak law of large numbers

14.1. A sequence of random variables X_i is also called a **stochastic process**. We often deal with sums $S_n = X_1 + X_2 + \cdots + X_n$ and especially the time averages S_n/n. For example, if X_i is the outcome of a dice, then S_n/n is the average of all the dice outcomes. We of course know what this average should be. Experience shows that it is the average of the distribution which is $m = (1+2+3+4+5+6)/6 = 21/6 = 3.5$.

14.2. The weak law of large numbers holds for pairwise uncorrelated random variables. This is a remarkably weak assumption.

Theorem 1. *Assume $X_i \in \mathcal{L}^2$ are pairwise uncorrelated, have a common mean $\mathrm{E}[X_i] = m$ and $M = \sup_n \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[X_i] < \infty$. Then $\frac{S_n}{n} \to m$ in probability.*

Proof. Since $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\cdot\mathrm{Cov}[X,Y]$ and X_n are pairwise uncorrelated, we get $\mathrm{Var}[X_n + X_m] = \mathrm{Var}[X_n] + \mathrm{Var}[X_m]$ and by induction $\mathrm{Var}[S_n] = \sum_{i=1}^n \mathrm{Var}[X_n]$. Using linearity, we obtain $\mathrm{E}[S_n/n] = m$ and

$$\mathrm{Var}[\frac{S_n}{n}] = \mathrm{E}[\frac{S_n^2}{n^2}] - \frac{\mathrm{E}[S_n]^2}{n^2} = \frac{\mathrm{Var}[S_n]}{n^2} = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_n] \; .$$

The right hand side converges to zero for $n \to \infty$. With Chebychev's inequality we obtain

$$\mathrm{P}[|\frac{S_n}{n} - m| \geq \epsilon] \leq \frac{\mathrm{Var}[\frac{S_n}{n}]}{\epsilon^2} \leq \frac{M}{n\epsilon^2} \; .$$

□

14.3. As an application in analysis, this leads to a constructive proof of a **theorem of Weierstrass** which states that polynomials are dense in the space $C[0,1]$ of all continuous functions on the interval $[0,1]$.

Theorem 2. *For every $f \in C[0,1]$, the* **Bernstein polynomials**

$$B_n(x) = \sum_{k=1}^n f(\frac{k}{n}) \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{n-k}$$

converge uniformly to f. If $f(x) \geq 0$, then also $B_n(x) \geq 0$.

Proof. For $x \in [0,1]$, let X_n be a sequence of independent $\{0,1\}$- valued random variables with mean value x. In other words, we take the probability space $(\{0,1\}^{\mathbb{N}}, \mathcal{A}, \mathrm{P})$

defined by $\mathrm{P}[\omega_n = 1] = x$. Since $\mathrm{P}[S_n = k] = \binom{n}{k} x^k (1-p)^{n-k}$, we can write $B_n(x) = \mathrm{E}[f(\frac{S_n}{n})]$. We estimate with $||f|| = \max_{0 \leq x \leq 1} |f(x)|$

$$\begin{aligned} |B_n(x) - f(x)| &= |\mathrm{E}[f(\frac{S_n}{n})] - f(x)| \leq \mathrm{E}[|f(\frac{S_n}{n}) - f(x)|] \\ &\leq 2||f|| \cdot \mathrm{P}[|\frac{S_n}{n} - x| \geq \delta] \\ &+ \sup_{|x-y| \leq \delta} |f(x) - f(y)| \cdot \mathrm{P}[|\frac{S_n}{n} - x| < \delta] \\ &\leq 2||f|| \cdot \mathrm{P}[|\frac{S_n}{n} - x| \geq \delta] + \sup_{|x-y| \leq \delta} |f(x) - f(y)| \; . \end{aligned}$$

The second term in the last line is called the **continuity module** of f. It converges to zero for $\delta \to 0$. By the Chebychev inequality and the proof of the weak law of large numbers, the first term can be estimated from above by

$$2||f|| \frac{\mathrm{Var}[X_i]}{n\delta^2} \; ,$$

a bound which goes to zero for $n \to \infty$ because the variance satisfies $\mathrm{Var}[X_i] = x(1-x) \leq 1/4$. □

14.4. In the weak law of large numbers, we only assumed the random variables to be uncorrelated. Under the stronger condition of independence and the moment assumption $X^4 \in \mathcal{L}^1$, the convergence can be accelerated:

Theorem 3. *Assume $X_i \in \mathcal{L}^4$ have common expectation $\mathrm{E}[X_i] = m$ and satisfy $M = \sup_n ||X||_4 < \infty$. If X_i are independent, then $S_n/n \to m$ in probability. Even $\sum_{n=1}^{\infty} \mathrm{P}[|\frac{S_n}{n} - m| \geq \epsilon]$ converges for all $\epsilon > 0$.*

Proof. We can assume without loss of generality that $m = 0$. Because the X_i are independent, we get

$$\mathrm{E}[S_n^4] = \sum_{i_1, i_2, i_3, i_4 = 1}^{n} \mathrm{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}] \; .$$

Again by independence, a summand $\mathrm{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}]$ is zero if an index $i = i_k$ occurs alone, it is $\mathrm{E}[X_i^4]$ if all indices are the same and $\mathrm{E}[X_i^2]\mathrm{E}[X_j^2]$, if there are two pairwise equal indices. Since by Jensen's inequality $\mathrm{E}[X_i^2]^2 \leq \mathrm{E}[X_i^4] \leq M$ we get

$$\mathrm{E}[S_n^4] \leq nM + n(n-1)M \; .$$

Use now the Chebyshev-Markov inequality with $h(x) = x^4$ to get

$$\begin{aligned} \mathrm{P}[|\frac{S_n}{n}| \geq \epsilon] &\leq \frac{\mathrm{E}[(S_n/n)^4]}{\epsilon^4} \\ &\leq M \frac{n + n^2}{\epsilon^4 n^4} \leq 2M \frac{1}{\epsilon^4 n^2} \; . \end{aligned}$$

□

PROBABILITY THEORY

MATH 154

Unit 15: Strong law of large numbers

15.1. Laws of large numbers make statements about the stochastic convergence of sums $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n}$ of random variables X_n. The weak law dealt with convergence in probability. The strong laws of large numbers make analog statements about almost everywhere convergence.

15.2. The first version of the strong law does not assume the random variables to have the same distribution. They are assumed to have the same expectation and have to be bounded in $\mathcal{L}^4$. We only need decorrelation.

Theorem 1. *Assume X_n are pairwise decorrelated random variables in $\mathcal{L}^4$ with common expectation $\mathrm{E}[X_n] = m$ and for which $M = \sup_n ||X_n||_4^4 < \infty$. Then $S_n/n \to m$ almost everywhere.*

Proof. In the proof of the weak law of large numbers, we derived from these assumptions $\mathrm{P}[|\frac{S_n}{n} - m| \geq \epsilon] \leq 2M\frac{1}{\epsilon^4 n^2}$. This means that $S_n/n \to m$ converges completely. But complete convergence implies almost everywhere convergence. □

15.3. The strong law for IID $\mathcal{L}^1$ random variables was first proven by Kolmogorov in 1930 under the assumption of independence. Much later, in 1981, it has been observed that the weaker notion of **pairwise independence** is sufficient.

Theorem 2 (Etemadi)**.** *Assume $X_n \in \mathcal{L}^1$ are pairwise independent and identically distributed random variables with mean m. Then $S_n/n \to m$ almost everywhere.*

Proof. (0) We can assume without loss of generality that $\boxed{X_n \geq 0}$ because we can split $X_n = X_n^+ + X_n^-$ into its positive $X_n^+ = X_n \vee 0 = \max(X_n, 0)$ and negative part $X^- = -X \vee 0 = \max(-X, 0)$ and because knowing the result for $X_n^\pm$ implies it for X_n. Define $f_R(t) = t \cdot 1_{[-R,R]}$, the random variables $X_n^{(R)} = f_R(X_n)$ and $Y_n = X_n^{(n)}$ as well as $S_n = \frac{1}{n}\sum_{i=1}^n X_i$, $T_n = \frac{1}{n}\sum_{i=1}^n Y_i$.
(i) $\boxed{T_n - \mathrm{E}[T_n] \to 0 \text{ implies } S_n - \mathrm{E}[S_n] \to 0}$.
Proof. Since $\mathrm{E}[Y_n] \to m$, we have $\mathrm{E}[T_n] \to m$. Because of (0) $\sum_{n\geq 1} \mathrm{P}[Y_n \neq X_n] \leq \sum_{n\geq 1} \mathrm{P}[X_n \geq n] = \sum_{n\geq 1} \mathrm{P}[X_1 \geq n] = \sum_{n\geq 1}\sum_{k\geq n} \mathrm{P}[X_n \in [k, k+1]]$
$= \sum_{k\geq 1} k \cdot \mathrm{P}[X_1 \in [k, k+1]] \leq \mathrm{E}[X_1] = m < \infty$, we get by the first Borel-Cantelli lemma that $\mathrm{P}[Y_n \neq X_n, \text{ infinitely often}] = 0$. This means $T_n - S_n \to 0$ almost everywhere, proving $\mathrm{E}[S_n] \to m$ if $\mathrm{E}[T_n] \to m$.
(ii) $\boxed{\text{Complete convergence along subsequence.}}$ Fix a real number $\alpha > 1$ and define an exponentially growing subsequence $k_n = [\alpha^n]$ which is the **integer part** of α^n. Denote

by μ the law of the random variables X_n. For every $\epsilon > 0$, we get (using Chebyshev's inequality using pairwise independence) for $k_n = [\alpha^n]$ some constants C which can vary in each step: $\sum_{n=1}^{\infty} \mathrm{P}[|T_{k_n} - \mathrm{E}[T_{k_n}]| \geq \epsilon] \leq \sum_{n=1}^{\infty} \frac{\mathrm{Var}[T_{k_n}]}{\epsilon^2} = \sum_{n=1}^{\infty} \frac{1}{\epsilon^2 k_n^2} \sum_{m=1}^{k_n} \mathrm{Var}[Y_m]$ $= \frac{1}{\epsilon^2} \sum_{m=1}^{\infty} \mathrm{Var}[Y_m] \sum_{n,k_n \geq m} \frac{1}{k_n^2} \leq^{(1)} \frac{1}{\epsilon^2} \sum_{m=1}^{\infty} \mathrm{Var}[Y_m] \frac{C}{m^2} \leq C \sum_{m=1}^{\infty} \frac{1}{m^2} \mathrm{E}[Y_m^2]$. In (1) we used that with $k_n = [\alpha^n]$ one has $\sum_{n:k_n \geq m} k_n^{-2} \leq C \cdot m^{-2}$. In the last step we used that $\mathrm{Var}[Y_m] = \mathrm{E}[Y_m^2] - \mathrm{E}[Y_m]^2 \leq \mathrm{E}[Y_m^2]$. After catching breath, we continue:

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathrm{P}[|T_{k_n} - \mathrm{E}[T_{k_n}]| \geq \epsilon] &\leq C \sum_{m=1}^{\infty} \frac{1}{m^2} \mathrm{E}[Y_m^2] \\
&\leq C \sum_{m=1}^{\infty} \frac{1}{m^2} \sum_{l=0}^{m-1} \int_l^{l+1} x^2 \, d\mu(x) \\
&= C \sum_{l=0}^{\infty} \sum_{m=l+1}^{\infty} \frac{1}{m^2} \int_l^{l+1} x^2 \, d\mu(x) \\
&\leq C \sum_{l=0}^{\infty} \sum_{m=l+1}^{\infty} \frac{(l+1)}{m^2} \int_l^{l+1} x \, d\mu(x) \\
&\leq^{(2)} C \sum_{l=0}^{\infty} \int_l^{l+1} x \, d\mu(x) \\
&\leq C \cdot \mathrm{E}[X_1] < \infty \, .
\end{aligned}$$

In (2) we used that $\sum_{m=l+1}^{n} m^{-2} \leq C \cdot (l+1)^{-1}$.
We have now proved complete convergence. As before, this implies the almost everywhere convergence of $T_{k_n} - \mathrm{E}[T_{k_n}] \to 0$.
(iii) General case. Convergence has been verified along a subsequence k_n. Because we assumed $X_n \geq 0$, the sequence $U_n = \sum_{i=1}^{n} Y_n = nT_n$ is monotonically increasing. For $n \in [k_m, k_{m+1}]$, we get therefore $\frac{k_m}{k_{m+1}} \frac{U_{k_m}}{k_m} = \frac{U_{k_m}}{k_{m+1}} \leq \frac{U_n}{n} \leq \frac{U_{k_{m+1}}}{k_m} = \frac{k_{m+1}}{k_m} \frac{U_{k_{m+1}}}{k_{m+1}}$ and from $\lim_{n\to\infty} T_{k_m} = \mathrm{E}[X_1]$ almost everywhere, the statement $\frac{1}{\alpha}\mathrm{E}[X_1] \leq \liminf_n T_n \leq \limsup_n T_n \leq \alpha \mathrm{E}[X_1]$ follows. □

15.4. The strong law of large numbers can be interpreted as a statement about the growth of the sequence $\sum_{k=1}^{n} X_n$. For $\mathrm{E}[X_1] = 0$, the convergence $\frac{1}{n} \sum_{k=1}^{n} X_n \to 0$ means that for all $\epsilon > 0$ there exists m such that for $n > m$

$$|\sum_{k=1}^{n} X_n| \leq \epsilon n \, .$$

This means that the trajectory $\sum_{k=1}^{n} X_n$ is finally contained in any arbitrary narrow cone. In other words, it grows slower than linear. The exact description for the growth of $\sum_{k=1}^{n} X_n$ is given by the **law of the iterated logarithm of Khinchin** which says that a sequence of IID random variables X_n with $\mathrm{E}[X_n] = m$ and $\sigma(X_n) = \sigma \neq 0$ satisfies with $\Lambda_n = \sqrt{2\sigma^2 n \log\log n}$:

$$\limsup_{n\to\infty} \frac{S_n}{\Lambda_n} = +1, \liminf_{n\to\infty} \frac{S_n}{\Lambda_n} = -1 \, .$$

PROBABILITY THEORY

MATH 154

Unit 16: Ergodic theorem

16.1. A map $T : \Omega \to \Omega$ on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ is **measurable** if $T^{-1}(A) \in \mathcal{A}$ for every $A \in \mathcal{A}$. A single random variable X defines a sequence $X_n(\omega) = X(T^n(\omega))$ of random variables where $T^n(\omega) = T(T(\dots T(\omega)))$ is the n'th iterate. T is called **measure preserving**, if $\mathrm{P}[T^{-1}(A)] = \mathrm{P}[A]$ for all $A \in \mathcal{A}$. It is **ergodic** if $T(A) = A$ implies $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$. The map T is called **invertible**, if there exists a measurable, measure preserving inverse T^{-1} of T. Then, T is called an **automorphism**. Ergodicity of T is equivalent to the statement that the linear map $U(f)(x) = f(T(x))$ has a one-dimensional eigenspace, consisting of constant functions. Given a random variable X define $X_k = X(T^k)$ and $S_n = \sum_{k=0}^{n-1} X_k$. Eberhard Hopf showed:

Theorem 1 (Maximal ergodic theorem)**.** *Given $X \in \mathcal{L}^1$ and a measure preserving transformation T, the event $A = \{\sup_n S_n > 0\}$ satisfies $\mathrm{E}[X; A] = \mathrm{E}[1_A X] \geq 0$.*

Proof. Define $Z_n = \max_{0 \leq k \leq n} S_k$ and the sets $A_n = \{Z_n > 0\} \subset A_{n+1}$. Then $A = \bigcup_n A_n$. Clearly $Z_n \in \mathcal{L}^1$. For $0 \leq k \leq n$, we have $Z_n \geq S_k$ and so $Z_n(T) \geq S_k(T)$ and hence $Z_n(T) + X \geq S_{k+1}$. By taking the maxima on both sides over $0 \leq k \leq n$, we get $Z_n(T) + X \geq \max_{1 \leq k \leq n+1} S_k$. On $A_n = \{Z_n > 0\}$, we can extend this to $Z_n(T) + X \geq \max_{1 \leq k \leq n+1} S_k \geq \max_{0 \leq k \leq n+1} S_k = Z_{n+1} \geq Z_n$ so that on A_n we have $X \geq Z_n - Z_n(T)$. Integration over the set A_n gives $\mathrm{E}[X; A_n] \geq \mathrm{E}[Z_n; A_n] - \mathrm{E}[Z_n(T); A_n]$ Using (1) this inequality, the fact (2) that $Z_n = 0$ on $\Omega \setminus A_n$, the (3) inequality $Z_n(T) \geq S_n(T) \geq 0$ on A_n and finally that T is measure preserving (4), leads to

$$\begin{aligned} \mathrm{E}[X; A_n] &\geq_{(1)} \mathrm{E}[Z_n; A_n] - \mathrm{E}[Z_n(T); A_n] \\ &=_{(2)} \mathrm{E}[Z_n] - \mathrm{E}[Z_n(T); A_n] \\ &\geq_{(3)} \mathrm{E}[Z_n - Z_n(T)] =_{(4)} 0 \end{aligned}$$

for every n and so to $\mathrm{E}[X; A] \geq 0$. □

Theorem 2 (Birkhoff, 1931)**.** *For $X \in \mathcal{L}^1$ and ergodic T, one has $\frac{S_n}{n} \to^{a.e.} \mathrm{E}[X]$.*

Proof. Define $\overline{X} = \limsup_{n \to \infty} \frac{S_n}{n}$, $\underline{X} = \liminf_{n \to \infty} \frac{S_n}{n}$. We get $\overline{X} = \overline{X(T)}$ and $\underline{X} = \underline{X}(T)$ because

$$\frac{n+1}{n} \frac{S_{n+1}}{(n+1)} - \frac{S_n(T)}{n} = \frac{X}{n} .$$

(i) $\overline{X} = \underline{X}$. Define for $\beta < \alpha \in \mathbb{R}$ the set $A_{\alpha,\beta} = \{\underline{X} < \beta < \alpha < \overline{X}\}$. It is T-invariant because $\overline{X}, \underline{X}$ are T-invariant as mentioned at the beginning of the proof. Because $\{\underline{X} < \overline{X}\} = \bigcup_{\beta < \alpha, \alpha, \beta \in \mathbb{A}} A_{\alpha,\beta}$, it is enough to show that $\mathrm{P}[A_{\alpha,\beta}] = 0$ for

rational $\beta < \alpha$. The rest of the proof establishes this. In order to use the maximal ergodic theorem, we also define

$$\begin{aligned} B_{\alpha,\beta} &= \{\sup_n(S_n - n\alpha) > 0\ , \sup_n(S_n - n\beta) < 0\} \\ &= \{\sup_n(\frac{S_n}{n} - \alpha) > 0, \sup_n(\frac{S_n}{n} - \beta) < 0\ \} \\ &\supset \{\limsup_n(\frac{S_n}{n} - \alpha) > 0, \limsup_n(\frac{S_n}{n} - \beta) < 0\ \} \\ &= \{\overline{X} - \alpha > 0, \underline{X} - \beta < 0\ \} = A_{\alpha,\beta}\ . \end{aligned}$$

Because $A_{\alpha,\beta} \subset B_{\alpha,\beta}$ and $A_{\alpha,\beta}$ is T-invariant, we get from the maximal ergodic theorem $\mathrm{E}[\overline{X} - \alpha; A_{\alpha,\beta}] \geq 0$ and so

$$\mathrm{E}[\overline{X}; A_{\alpha,\beta}] \geq \alpha \cdot \mathrm{P}[A_{\alpha,\beta}]\ .$$

Because $A_{\alpha,\beta}$ is T-invariant, we we get from (i) restricted to the system T on $A_{\alpha,\beta}$ that $\mathrm{E}[\overline{X}; A_{\alpha,\beta}] = \mathrm{E}[X; A_{\alpha,\beta}]$ and so

$$\mathrm{E}[X; A_{\alpha,\beta}] \geq \alpha \cdot \mathrm{P}[A_{\alpha,\beta}]\ . \tag{1}$$

Replacing X, α, β with $-X, -\beta, -\alpha$ and using $\overline{-X} = -\underline{X}$ shows in exactly the same way that

$$\mathrm{E}[X; A_{\alpha,\beta}] \leq \beta \cdot \mathrm{P}[A_{\alpha,\beta}]\ . \tag{2}$$

The two equations (1),(2) imply that

$$\beta \mathrm{P}[A_{\alpha,\beta}] \geq \alpha \mathrm{P}[A_{\alpha,\beta}]$$

which together with $\beta < \alpha$ only leave us to conclude $\mathrm{P}[A_{\alpha,\beta}] = 0$.

(ii) $\overline{X} \in \mathcal{L}^1$ We have $|S_n/n| \leq |X|$, and by (i) that S_n/n converges point-wise to $\overline{X} = \underline{X}$ and $X \in \mathcal{L}^1$. The Lebesgue's dominated convergence theorem assures $\overline{X} \in \mathcal{L}^1$.

(iii) $\mathrm{E}[X] = \mathrm{E}[\overline{X}]$. Define the T-invariant sets $B_{k,n} = \{\overline{X} \in [\frac{k}{n}, \frac{k+1}{n})\}$ for $k \in \mathbb{Z}, n \geq 1$. Define for $\epsilon > 0$ the random variable $Y = X - \frac{k}{n} + \epsilon$ and call $\tilde{S}_n$ the sums where X is replaced by Y. We know that for n large enough $\sup_n \tilde{S}_n \geq 0$ on $B_{k,n}$. When applying the maximal ergodic theorem applied to the random variable Y on $B_{k,n}$. we get $\mathrm{E}[Y; B_{k,n}] \geq 0$. Because $\epsilon > 0$ was arbitrary, $\mathrm{E}[X; B_{k,n}] \geq \frac{k}{n}\mathrm{P}[B_{k,n}]$. With this inequality,

$$\mathrm{E}[\overline{X}, B_{k,n}] \leq \frac{k+1}{n}\mathrm{P}[B_{k,n}] \leq \frac{1}{n}\mathrm{P}[B_{k,n}] + \frac{k}{n}\mathrm{P}[B_{k,n}] \leq \frac{1}{n}\mathrm{P}[B_{k,n}] + \mathrm{E}[X; B_{k,n}]\ .$$

Summing over k gives

$$\mathrm{E}[\overline{X}] \leq \frac{1}{n} + \mathrm{E}[X]$$

and because n was arbitrary, $\mathrm{E}[\overline{X}] \leq \mathrm{E}[X]$. Doing the same with $-X$ we end with

$$\mathrm{E}[-\overline{X}] = \mathrm{E}[\underline{-X}] \leq \mathrm{E}[\overline{-X}] \leq \mathrm{E}[-X]\ .$$

□

PROBABILITY THEORY

MATH 154

Unit 17: Recurrence

17.1. Fix a probability space $(\Omega, \mathcal{A}, \mathrm{P})$. A map $T : \Omega \to \Omega$ is **measurable** if $T^{-1}(A) \in \mathcal{A}$ for every $A \in \mathcal{A}$. It is **measure preserving** if $\mathrm{P}[T^{-1}(A)] = \mathrm{P}[A]$, $\forall A \in \mathcal{A}$. If T is invertible and T^{-1} is measure preserving too, T is an **automorphism** of $(\Omega, \mathcal{A}, \mathrm{P})$. Automorphisms form a group. For example, on a finite probability space with $\mathrm{P}[A] = |A|/|\Omega|$, the automorphisms are permutations and the ergodic ones are cyclic.

17.2. Probability spaces with measure preserving maps as **morphisms** form a nice **category**. Poincaré proved in 1890:

Theorem 1 (Poincaré recurrence)**.** *Given an automorphism T and $A \in \mathcal{A}$ with $\mathrm{P}[A] > 0$, there exists n such that $\mathrm{P}[T^n(A) \cap A] > 0$.*

Proof. If not, then $A_n = T^n(A)$ is a distinct set of events. Let $n > 1/\mathrm{P}[A]$ be an integer. Use finite additivity to see $1 = \mathrm{P}[\Omega] \geq \mathrm{P}[\bigcup_{k=1}^n A_k] = \sum_{k=1}^n \mathrm{P}[A_k] = n\mathrm{P}[A] > 1$ which is a contradiction. □

17.3. Recall that T is called **ergodic** if every fixed point $T(A) = A$ has probability 0 or 1. In short, this means $\mathrm{P}[A + T(A)] = 0 \Rightarrow \mathrm{P}[A]^2 = \mathrm{P}[A]$. If we take a single random variable X, it defines a sequence $X_n(\omega) = X(T^n(\omega))$ of random variables, where $T^n(\omega) = T(T(\ldots T(\omega)))$ is the n'th iterate. If T is ergodic and $X = 1_A$ then $\mathrm{E}[S_n]/n \to \mathrm{E}[X] = \mathrm{P}[A]$. The ergodic theorem tells us that the frequency of the number of times that we hit A is the same than the probability of A. The catch phrase is that **"space average agrees with time average"**.

17.4. Let $\Omega = \{|z| = 1\} \subset \mathbb{C}$ be the unit circle in the complex plane equipped with the probability measure $\mathrm{P}[\mathrm{Arg}(z) \in [a, b]] = (b - a)/(2\pi)$ for $0 < a < b < 2\pi$ and the Borel σ-algebra $\mathcal{A}$. If $w = e^{2\pi i\alpha}$ is a complex number of length 1, then the rotation $T(z) = wz$ defines a measure preserving transformation on $(\Omega, \mathcal{A}, \mathrm{P})$. It is invertible with inverse $T^{-1}(z) = z/w$. This system is called the **Kronecker system**. It can be written additively as $\theta \to \theta + \alpha \bmod 2\pi$.

Theorem 2. *If α is irrational, then the Kronecker system is ergodic.*

Proof. With $z = e^{2\pi ix}$, one can write a random variable $X \in \mathcal{L}^2$ on Ω as a Fourier series $f(z) = \sum_{n=-\infty}^{\infty} a_n z^n$ with $a_n = \mathrm{E}[z^n X]$. We can write $f = f_0 + f_+ + f_-$, where $f_+ = \sum_{n=1}^{\infty} a_n z^n$ is analytic in $|z| < 1$ and $f_- = \sum_{n=1}^{\infty} a_n z^{-n}$ is analytic in $|z| > 1$ and f_0 is constant. By doing the same decomposition for $f(T(z)) = \sum_{n=-\infty}^{\infty} a_n w^n z^n$, we see that $f_+ = \sum_{n=1}^{\infty} a_n z^n = \sum_{n=1}^{\infty} a_n w^n z^n$. But these are the Taylor expansions of

$f_+ = f_+(T)$ and so $a_n = a_n w^n$. Because $w^n \neq 1$ for irrational α, we deduce that $a_n = 0$ for $n \geq 1$. Similarly, one derives $a_n = 0$ for $n \leq -1$. Therefore $f(z) = a_0$, meaning f is constant. □

17.5. It follows that for every number $x \in [0,1]$ and every irrational α and every $\epsilon > 0$, there exists n such that $|n\alpha - x| < \epsilon$.

17.6. The transformation $T(z) = z^2$ on the same probability space as in the previous example is also measure preserving. Note that $\mathrm{P}[T(A)] = 2\mathrm{P}[A]$ but $\mathrm{P}[T^{-1}(A)] = \mathrm{P}[A]$ for all $A \in \mathcal{A}$. The map is measure preserving, but it is **not invertible**.

Theorem 3. *The squaring transformation $T(z) = z^2$ on the circle is ergodic.*

Proof. A Fourier argument shows it again: T preserves again the decomposition of f into three analytic functions $f = f_- + f_0 + f_+$ so that $f(T(z)) = \sum_{n=-\infty}^{\infty} a_n z^{2n} = \sum_{n=-\infty}^{\infty} a_n z^n$ implies $\sum_{n=1}^{\infty} a_n z^{2n} = \sum_{n=1}^{\infty} a_n z^n$. Comparing Taylor coefficients of this identity for analytic functions shows $a_n = 0$ for odd n because the left hand side has zero Taylor coefficients for odd powers of z. But because for even $n = 2^l k$ with odd k, we have $a_n = a_{2^l k} = a_{2^{l-1} k} = \cdots = a_k = 0$, all coefficients $a_k = 0$ for $k \geq 1$. Similarly, one sees $a_k = 0$ for $k \leq -1$. □

17.7. The single angle random variable $X(x) = \arg(x)$ on Ω, produces a sequence of random variables $X_n(x) = X(T^n(x))$. The squaring system is conjugated to the shift $S(x)_n = x_{n+1}$ on the product probability space $(\{0,1\}^{\mathbb{N}}, \mathcal{B}, \mathrm{P})$. The conjugating map is $\phi(x) = e^{2\pi x_i / 2^i} \in \Omega$. We have $\phi(S(x)) = T(\phi(x))$.

17.8. If A is an event with $\mathrm{P}[A] > 0$ and T is a measure preserving automorphism, we can define a new transformation $T_A(x) = T^{n_A(x)}(x)$, where $n_A(x)$ is the **return time**, the smallest $n > 0$ with $T^n(x) \in A$. By Poincaré recurrence, the random variable $n_A(x)$ is finite for almost all $x \in A$. We can look at the **conditional probability space** $(A, \mathcal{A} \cap A, \mathrm{P}/\mathrm{P}[A])$ and the **induced dynamical system** T_A.

Theorem 4. *T_A is an automorphism of $(A, \mathcal{A} \cap A, \mathrm{P}/\mathrm{P}[A])$. It is ergodic if T is ergodic.*

Proof. (i) T_A is measure preserving. Decompose $A = \bigcup_{k=1}^{\infty} A_k$ with $A_k = \{n_A = k\}$. Now $T_A(x) = T^k(x)$ for $x \in A_k$. Given $B \in \mathcal{A} \cap A$ define $B_k = B \cap A_k$ so that $B = \bigcup_k B_k$. $\mathrm{P}[T_A^{-1}(B)] = \mathrm{P}[T_A^{-1}(\bigcup_k B_k)] = \mathrm{P}[\bigcup_k T_A^{-1}(B_k)] = \mathrm{P}[\bigcup_k T^{-k} B_k] = \sum_k \mathrm{P}[T^{-k}(B_k)] = \sum_k \mathrm{P}[B_k] = \mathrm{P}[\bigcup_k B_k] = \mathrm{P}[B]$. (ii) If T is ergodic then T_A is ergodic. Proof: use contradiction. If $T_A(B) = B$ has $\mathrm{P}[B] < \mathrm{P}[A]$ then $C = \bigcup_k T^{-k}(B)$ is T invariant with $\mathrm{P}[C] < 1$ and T is not ergodic. □

17.9. Note that it is possible that T_A is ergodic but T is not ergodic. Kakutani noticed that if $\bigcup_k T^k A = \Omega$, then T_A is ergodic if and only if T is ergodic.

17.10. If T is a measure preserving transformation on a probability space, we can look at the longer incidences $A \cap T^{-n}(A) \cap \cdots \cap T^{-kn}(A)$. Fürstenberg showed in 1977:

Theorem 5 (Multiple recurrence theorem)**.** *For any $A \in \mathcal{A}$ with $\mathrm{P}[A] > 0$, there exists n such that $\mathrm{P}[A \cap T^{-n}(A) \cap \cdots \cap T^{-kn}(A) \cap A] > 0$.*

It implies the **van der Waerden theorem** telling that any r coloring of the integers contains a color which contains arbitrary large arithmetic progressions.

PROBABILITY THEORY

MATH 154

Unit 18: Mixing

18.1. T is **ergodic** if all events $A \in \mathcal{A}$ fixed by T have probability 0 or 1. This can be rephrased as **Césaro decay of correlations** of all random variables $X = 1_{T^{-k}(A)}, Y_k = 1_B$ because $\mathrm{E}[X_k] = \mathrm{P}[T^{-k}(A)], \mathrm{E}[Y] = \mathrm{P}[B]$ and $\mathrm{Cov}[X_k, Y] = \mathrm{P}[T^{-k}(A) \cap B]$ and:

Theorem 1. *T ergodic* $\Leftrightarrow \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} (\mathrm{P}[T^{-k}(A) \cap B] - \mathrm{P}[A]\mathrm{P}[B]) = 0 \ \forall A, B \in \mathcal{A}$.

Proof. (i) The ergodic theorem gives for $X \in \mathcal{L}^2$ the limit $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} X(T^{-k}(x)) \to \mathrm{E}[X]$. Given $Y \in \mathcal{L}^2$, we have $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} X(T^{-k}(x))Y(x) \to \mathrm{E}[X]Y(x)$. Now take expectations $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} E[X(T^{-k})Y] \to E[X]\mathrm{E}[Y]$. If we take $X = 1_A, Y = 1_B$ we get the statement.
(ii) Use contraposition. Assume T is not ergodic. Take any B with $T(B) = B$ with $0 < \mathrm{P}[B] < 1$ and chose $A = B$. We want to show that B is trivial. The assumed identity gives $\mathrm{P}[A] = \frac{1}{n} \sum_{k=0}^{n-1} \mathrm{P}[A] = \frac{1}{n} \sum_{k=0}^{n-1} \mathrm{P}[T^{-k}(A) \cap B] \to \mathrm{P}[A]\mathrm{P}[B] = \mathrm{P}[A]^2$. But this implies either $\mathrm{P}[A] = 1$ or $\mathrm{P}[A] = 0$. □

18.2. T is called **weakly mixing** if $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} |\mathrm{P}[A \cap T^{-k}(B)] - \mathrm{P}[A]\mathrm{P}[B]| = 0$. This looks similar than the ergodic property but there are absolute values! We have **absolute Césaro convergence**. A Kronecker system for example is ergodic but not weakly mixing. Also, no permutation on a finite probability space is weakly mixing but it is ergodic if there is a single cycle. The previous theorem shows that if T is weakly mixing, then T is ergodic. Putting the absolute values outside makes it smaller in general by the triangle inequality.

18.3. Define $T \times T(x, y) = (T(x), T(y))$ on the product probability space. If T is not ergodic, then $T^{-1}(A) = A$ then $T^{-1}(A \times \Omega) = (T^{-1}(A) \times \Omega) = (A, \Omega)$ so that $T \times T$ is not ergodic. We see that ergodicity of $T \times T$ is stronger than ergodicity. What does it mean?

Theorem 2. *T is weakly mixing if and only if $T \times T$ is ergodic.*

Proof. We will show this in class. It needs a bit of real analysis (see HW 8). A second proof is spectral theoretic using that U_T has no eigenvalues. There is not enough space here on two pages. We will show the even stronger statement: T is weakly mixing if and only if $T \times T$ is weakly mixing. □

18.4. Note that if T is ergodic, then the power $T^2(x) = T(T(x))$ is not necessarily ergodic. A simple example is the measure preserving transformation $T(x) = x + 1 \bmod 6$ on the finite probability space $\Omega = \mathbb{Z}_6 = \{0, 1, \ldots, 5\}$ with $\mathcal{A} = 2^\Omega$ and uniform probability measure $\mathrm{P}[A] = |A|/|\Omega|$. The transformation T is ergodic, but T^2 leaves the set $A = \{0, 2, 4\}$ invariant and $\mathrm{P}[A] = 1/2$ is not in $\{0, 1\}$.

18.5. A measure preserving is called **mixing** if $\mathrm{P}[T^{-n}(A) \cap B] \to \mathrm{P}[A]\mathrm{P}[B]$. For mixing transformations, the events $T^n(A)$ and B become more and more independent. In particular, $T^n(A)$ and A become more and more independent like a colored A of a dough Ω gets mixed by kneading and folding. In probability theory, mixing means that the random variables $X_n = 1_{T^{-n}(A)}$ and $Y = 1_B$ become more and more decorrelated. In particular $\mathrm{Cov}[X_n, X_0] \to 0$. Obviously, if T is mixing, then T is weakly mixing.

18.6. The linear operator $U_T(X) = X(T^{-1})$ on $\mathcal{L}^2$ is called the **Koopman operator** associated with T. The Hilbert space $\mathcal{L}^2$ has the inner product $\langle X, Y\rangle = \mathrm{E}[\overline{X}Y]$. The unitary property $\langle UX, UY\rangle = \langle X, Y\rangle$ follows from the fact that T preserves probabilities. Note that U always has the trivial eigenvalue 1 with constant eigenfunction $X = c$. This eigenvalue is not interesting. The **spectrum** of the dynamical system is defined as the spectrum of U on the orthogonal complement of the constant random variables (which are functions of zero expectation). Every random variable X defines a **spectral measure** $\mu = \mu_X$ on the complex unit circle defined by $\hat{\mu}_n = \langle X, U^n X\rangle - E[X]^2 = \mathrm{E}[XX(T^{-n})] - E[X]E[X(T^{-n})] = \mathrm{Cov}[X, X(T^{-n})]$.

Theorem 3. *T is weakly mixing if and only if U_T has continuous spectrum.*

18.7. Weakly mixing implies $(1/n)\sum_{k=0}^{n-1} \mathrm{P}[A(T^{-k}) \cap A] - \mathrm{P}[A]^2 \to 0$ implying that the Fourier transform of the spectral measure 1_A goes to zero for every 1_A. The Wiener theorem gives the reverse: if $\hat{\mu}_k \to 0$ in a Cesaro sense, then μ has no point spectrum.

Theorem 4 (Wiener Theorem)**.** *If μ is a measure on the circle $\mathbb{T}$ with Fourier coefficients $\hat{\mu}_k$, then for every $x \in \mathbb{T}$, one has $\mu(\{x\}) = \lim_{n\to\infty} \frac{1}{2n+1}\sum_{k=-n}^n \hat{\mu}_k e^{ikx}$.*

Proof. The **Dirichlet kernel** $D_n(t) = \sum_{k=-n}^n e^{ikt} = \frac{\sin((k+1/2)t)}{\sin(t/2)}$ satisfies $D_n \star f(x) = S_n(f)(x) = \sum_{k=-n}^n \hat{f}(k)e^{ikx}$. The functions $f_n(t) := \frac{1}{2n+1} D_n(t-x) = \frac{1}{2n+1}\sum_{k=-n}^n e^{-ikx}e^{ikt}$ are bounded by 1 and go to zero uniformly outside any neighborhood of $t = x$. From $\lim_{\epsilon\to 0}\int_{x-\epsilon}^{x+\epsilon} |d(\mu - \mu(\{x\})\delta_x)| = 0$ follows $\lim_{n\to\infty}\langle f_n, \mu - \mu(\{x\})\rangle = 0$. But we also have $\langle f_n, \mu - \mu(\{x\})\rangle = \langle f_n, \mu\rangle - \langle f_n, \mu(\{x\})\rangle = \frac{1}{2n+1}\sum_{k=-n}^n \hat{\mu}_k e^{ikx} - \mu(\{x\})$. □

18.8. If U_T has absolutely continuous spectrum, then by the Riemann-Lebesgue lemma, $\hat{\mu}_n \to 0$ so that T is mixing.

18.9. A measure preserving transformation is called **Bernoulli** if it is isomorphic to the shift of a product probability space $\prod_n(\Omega_n, \mathcal{A}_n, \mathrm{P}_n)$ where each $(\Omega_n, \mathcal{A}_n, \mathrm{P}_n)$ is the same finite probability space with $(\Omega = \{1, \ldots, m\}, \mathcal{A} = 2^\Omega, \mathrm{P}[\{j\}] = p_j)$.

Theorem 5. *A Bernoulli transformation has absolutely continuous spectrum and so is mixing.*

18.10. We have the following **"chaos levels"**

$$\{\text{Bernoulli}\} \subset \{\text{Mixing}\} \subset \{\text{Weakly mixing}\} \subset \{\text{Ergodic}\}\ .$$

PROBABILITY THEORY

MATH 154

Unit 19: Central limit theorem

19.1. Any non-constant random variable $X \in \mathcal{L}^2$ can be **normalized** to $X^* = \frac{(X-\mathrm{E}[X])}{\sigma(X)}$. This normalized variable has zero **mean** $\mathrm{E}[X^*] = 0$ and **variance** $\sigma(X^*) = \sqrt{\mathrm{Var}[X^*]} = 1$. We can not normalize every random variable. A Cauchy distributed random variable for example has no finite variance and so can not be scaled to have variance 1. But we can normalize any non-constant random variable in $\mathcal{L}^2$.

Theorem 1 (Central limit theorem)**.** *Given $X_n \in \mathcal{L}^2$ which are IID with mean 0 and finite variance $\sigma^2 > 0$. Then $S_n/(\sigma\sqrt{n}) \to N(0,1)$ in distribution.*

19.2. The CLT can be encoded more briefly as $S_n^* \to^d N(0,1)$. Lets look first at some quantities for the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

which belongs to the **normal distribution** $N(0,\sigma^2)$. We we have $\mathrm{E}[|X|^p] = 2\int_0^\infty x^p f(x)\, dx$ which is after a substitution $u = x^2/(2\sigma^2)$ equal to

$$\frac{1}{\sqrt{\pi}} 2^{p/2}\sigma^p \int_0^\infty u^{\frac{1}{2}(p+1)-1} e^{-u}\, du \ .$$

The integral to the right is by definition equal to $\frac{2^{p/2}}{\sqrt{\pi}}\Gamma(\frac{1}{2}(p+1))$. We can also compute the characteristic function. $\phi_X(t) = e^{-t^2\sigma^2/2}$. To the proof:

Proof. By the Levy criterion for weak convergence and noting that $e^{-t^2/2}$ has no atoms, we have to show that for all $t \in \mathbb{R}$

$$\mathrm{E}[e^{it\frac{S_n}{\sigma\sqrt{n}}}] \to e^{-t^2/2} \ .$$

Denote by ϕ_{X_n} the characteristic function of X_n. As this is independent of n, we just write ϕ. Since by assumption $\mathrm{E}[X_n] = 0$ and $\mathrm{E}[X_n^2] = \sigma^2$, we can use, using the Taylor formula with remainder term:

$$\phi(t) = 1 - \frac{\sigma^2}{2}t^2 + o(t^2) \ .$$

This works despite that $\phi(t)$ does not necessarily have a full Taylor expansion. Lets use the **Landau notation** $o(f)$ for a term which could be replaced by a function g

satisfying $g(t)/f(t) \to 0$.

$$\begin{aligned} \mathrm{E}[e^{it\frac{S_n}{\sigma\sqrt{n}}}] &= \phi(\frac{t}{\sigma\sqrt{n}})^n \\ &= (1 - \frac{1}{2}\frac{t^2}{n} + o(\frac{1}{n}))^n \\ &= e^{-t^2/2} + o(1) \ . \end{aligned}$$

□

19.3. Remark: There is a different proof that only needs independence and allows for different distributions for each X_i. It needs some assumptions however like $M = \sup_i ||X_i||_3 < \infty and \delta = \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[X_i] > 0$. One can then show that

$$\lim_{n\to\infty} \mathrm{P}[S_n^* \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2}\, dy, \ \ \forall x \in \mathbb{R} \ .$$

A $N(0,\sigma^2)$ distributed random variable X satisfies $\mathrm{E}[|X|^p] = \frac{1}{\sqrt{\pi}} 2^{p/2}\sigma^p \Gamma(\frac{1}{2}(p+1)$ and so $\mathrm{E}[|X|^3] = \sqrt{\frac{8}{\pi}}\sigma^3$. But the proof is more technical.

19.4. Let $\mathcal{P}$ denote the space of probability measure μ on $(\mathbb{R}, \mathcal{B})$ which have the properties that $\int_{\mathbb{R}} x^2\, d\mu(x) = 1, \int_{\mathbb{R}} x\, d\mu(x) = 0$. Define the map on $\mathcal{P}$ as follows:

$$T\mu(A) = \int_{\mathbb{R}}\int_{\mathbb{R}} 1_A(\frac{x+y}{\sqrt{2}}) d\mu(x)\, d\mu(y) \ .$$

Technically, we can realize this map by taking for a given μ a random variable X with that law, then build a new random variable Y with the same distribution that is independent, then look at the law of $(X+Y)/\sqrt{2}$.

Theorem 2 (Renormalisation fixed point)**.** *The only fixed point of T on $\mathcal{P}$ is the law $N(0,1)$ of the standard normal distribution. It is an attractive fixed point in the sense that $T^n\mu \to N(0,1)$ starting with any initial condition μ.*

Proof. If μ is the law of a random variables X, Y with $\mathrm{Var}[X] = \mathrm{Var}[Y] = 1$ and $\mathrm{E}[X] = \mathrm{E}[Y] = 0$. Then $T(\mu)$ is the law of the normalized random variable $(X+Y)/\sqrt{2}$ because the independent random variables X, Y can be realized on the probability space $(\mathbb{R}^2, \mathcal{B}, \mu\times\mu)$ as coordinate functions $X((x,y)) = x, Y((x,y)) = y$. Then $T(\mu)$ is obviously the law of $(X+Y)/\sqrt{2}$. Now use that $T^n(X) = (S_{2^n})^*$ converges in distribution to $N(0,1)$. □

19.5. An other cool fact is that we can see that for normalized random variables with continuous PDF f that has finite **differentiable entropy** $S(X) = -\int_{\mathbb{R}} f(x)\log(f(x))\, dx$, the entropy increases when applying T. The Gibbs inequality from lecture 12 $D[p,q] \geq 0$ gives after a short computation. It uses that the entropy of $N(0,1)$ is $\log(\sqrt{2\pi e}) = 1.41894....$ [1]

Theorem 3. *The normal distribution is the distribution of maximal entropy among all distributions of finite differentiable entropy in $\mathcal{P}$.*

Oliver Knill, knill@math.harvard.edu, Math 154, Spring, 2025

[1] Pretty cool: the log of the geometric mean of 2π and e.

PROBABILITY THEORY

MATH 154

Unit 20: De Moivre-Laplace and Poisson

20.1. Assume X_i are IID random variables in $\mathcal{L}^2$ with mean m and standard deviation σ. What is the probability that the average S_n/n is within $\epsilon/\sqrt{n}$ to the mean m? An important applications of the central limit is that it allows us to validate data by averaging experiments.

Theorem 1. *The probability that S_n/n deviates more than $t\sigma/\sqrt{n}$ from $\mathrm{E}[X]$ can for large n be estimated by*

$$\frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \ .$$

Proof. Let $m = \mathrm{E}[X]$ denote the mean of X_k and σ the standard deviation. Denote by X a random variable which has the standard normal distribution $N(0,1)$. Write $X_n \sim Y_n$ if $X_n \to^d Y_n$ in distribution. By the central limit theorem

$$\frac{S_n - nm}{\sqrt{n}\sigma} \sim X \ .$$

Dividing both nominator and denominator by n gives $\frac{\sqrt{n}}{\sigma}(\frac{S_n}{n} - m) \sim X$ so that as distributions

$$\frac{S_n}{n} - m \sim X \frac{\sigma}{\sqrt{n}} \ .$$

But this means that we can estimate the deviation as $F_{N(0,1)}(t)$, which is the expression in the theorem. □

20.2. The term $\sigma/\sqrt{n}$ is called the **standard error**. The central limit theorem gives some insight why the standard error is important.

20.3. The case of coin tossing, meaning independent $\{0,1\}$-valued random variables with win probability $p \in (0,1)$ was historically the starting point for the central limit theorem. The sum S_n has a **Binomial distribution** $B(n,p)$ of mean np and variance $np(1-p)$. As we have just seen, the fact that

$$\lim_{n\to\infty} \mathrm{P}[\frac{(S_n - np)}{\sqrt{np(1-p)}} \leq x] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2}\, dy$$

is a consequence of the central limit theorem. It has already been proven by de Moivre in 1730 in the case $p = 1/2$ and for general $p \in (0,1)$ by Laplace in 1812.

Theorem 2 (DeMoivre-Laplace limit theorem)**.** *If S_n have the Binomial distribution $B(n,p)$, then S_n^* converges in distribution to $N(0,1)$.*

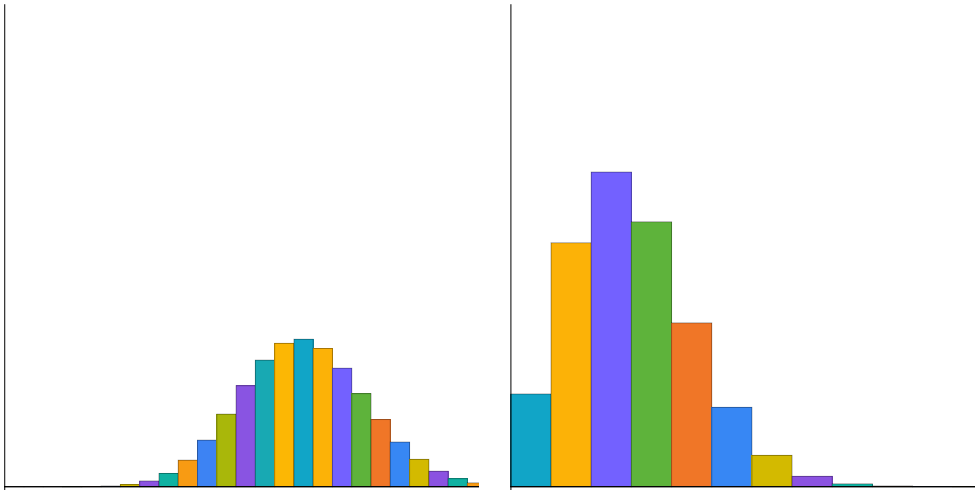


FIGURE 1. Binomial distribution $B(n, 0.3)$ to the left for $n = 50$. The Binomial distribution $B(n, 1/n)$ to the right for $n = 50$. The left will for $n \to \infty$ when rescaled will converge to the normal distribution. The right will converge to the Poisson distribution.

20.4. The **Poisson distribution** P_λ supported on $\mathbb{N} = \{0, 1, 2, , 3 \dots\}$ is defined as

$$P[X = k] = e^{-\lambda} \frac{\lambda^k}{k!} .$$

Random variables with such distribution describe waiting times. Its moment generating function is

$$M_X(t) = \sum_{k=0}^{\infty} P[X = k] e^{tk} = e^{\lambda(1-e^t)} .$$

That the Poisson distribution is natural follows from:

Theorem 3 (Poisson limit theorem)**.** *Let X_n be a $B(n, p_n)$-distributed and suppose $np_n \to \lambda$. Then X_n converges in distribution to a random variable X with Poisson distribution with parameter λ.*

20.5. For the proof we need the already in the proof of the central limit theorem used **compound interest statement** that if $a_n \to 0, b_n a_n \to c$ implies $(1 + a_n)^{b_n} \to e^c$ which is a tiny generalization of the definition $(1 + c/n)^{1/n} = e^c$.

Proof. We have to show that $P[X_n = k] \to P[X = k]$ for each fixed $k \in \mathbb{N}$.

$$\begin{aligned} P[X_n = k] &= \binom{n}{k} p_n^k (1 - p_n)^{n-k} \\ &= \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} p_n^k (1 - p_n)^{n-k} \\ &\sim \frac{1}{k!}(np_n)^k (1 - \frac{np_n}{n})^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda} . \end{aligned}$$

□

PROBABILITY THEORY

MATH 154

Unit 21: Random walks

21.1. A **random walk** or **Markov chain** on an undirected graph (V, E) is defined by a linear map which preserves probability vectors on V. The **adjacency matrix** A of (V, E) defines a **stochastic matrix** $M_{xy} = A_{xy}(x)/d(y)$, where $d(y)$ is the vertex degree of a vertex $y \in V$. The probability measures $M^n p(o)$ with initial probability measure $p(o)$ located on the initial point o. It is the distribution of the **standard random walk**. We return to the Markov picture next class.

21.2. If (V, E) is the standard lattice $\mathbb{Z}^d$, then each vertex degree is $d(x) = 2d$. We can describe a path of a random walk by defining IID random vectors X_i which take values in $I = \{e \in \mathbb{Z}^d \mid |e| = \sum_{i=1}^d |e_i| = 1 \}$ and which have the uniform distribution defined by $\mathrm{P}[X_n = e] = (2d)^{-1}$ for all $e \in I$. The random variable $S_n = \sum_{i=1}^n X_i$ with $S_0 = 0$ describes the position of the walker at time n. The stochastic process S_n is called the **random walk** on the lattice $\mathbb{Z}^d$. The law of S_n is a measure on $\mathbb{Z}^d$ which agrees with $M^n p(0)$.

21.3. Define the sets $A_n = \{S_n = 0 \}$ and the random variables $Y_n = 1_{A_n}$. If the walker has returned to position $0 \in \mathbb{Z}^d$ at time n, then $Y_n = 1$, otherwise $Y_n = 0$. The sum $B_n = \sum_{k=0}^n Y_k$ counts the number of visits of the origin 0 of the walker up to time n and $B = \sum_{k=0}^\infty Y_k$ counts the total number of visits at the origin. The expectation

$$\mathrm{E}[B] = \sum_{n=0}^{\infty} \mathrm{P}[A_n]$$

tells us how many times the walker is expected to return to the origin. We write $\mathrm{E}[B] = \infty$ if the sum diverges. In this case, the walker returns back to the origin infinitely many times.

Theorem 1 (Polya)**.** $\mathrm{E}[B] = \infty$ *for* $d = 1, 2$ *and* $\mathrm{E}[B] < \infty$ *for* $d > 2$.

Proof. Fix $n \in \mathbb{N}$ and define $a^{(n)}(k) = \mathrm{P}[S_n = k]$ for $k \in \mathbb{Z}^d$. Because the walker can reach in time n only a bounded region, the function $a^{(n)} : \mathbb{Z}^d \to \mathbb{R}$ is zero outside a bounded set. We can therefore define its Fourier transform

$$\phi_{S_n}(x) = \sum_{k \in \mathbb{Z}^d} a^{(n)}(k) e^{2\pi i k \cdot x}$$

which is a smooth function on $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. It is the characteristic function of S_n because

$$\mathrm{E}[e^{ixS_n}] = \sum_{k\in\mathbb{Z}^d} \mathrm{P}[S_n = k]e^{ik\cdot x} \ .$$

The characteristic function ϕ_X of X_k is

$$\phi_X(x) = \frac{1}{2d}\sum_{|j|=1} e^{2\pi i x_j} = \frac{1}{d}\sum_{i=1}^{d} \cos(2\pi x_i) \ .$$

Because the S_n is a sum of n independent random variables X_j

$$\phi_{S_n} = \phi_{X_1}(x)\phi_{X_2}(x)\dots\phi_{X_n}(x) = \frac{1}{d^n}(\sum_{i=1}^{d} \cos(2\pi x_i))^n \ .$$

Note that $a_n(0) = \mathrm{P}[S_n = 0] = \int_{\mathbb{T}} \phi_{S_n}(x)\ dx$.
We now show that $\mathrm{E}[B] = \sum_{n\geq 0} \phi_{S_n}(0)$ is finite if and only if $d < 3$. The Fourier inversion formula using the normalized Volume measure dx on $\mathbb{T}^3$ gives

$$\sum_n \mathrm{P}[S_n = 0] = \int_{\mathbb{T}^d} \sum_{n=0}^{\infty} \phi_X^n(x)\ dx = \int_{\mathbb{T}^d} \frac{1}{1-\phi_X(x)}\ dx \ .$$

A Taylor expansion $\phi_X(x) = 1 - \sum_j \frac{x_j^2}{2}(2\pi)^2 + \dots$ shows

$$\frac{1}{2}\cdot\frac{(2\pi)^2}{2d}|x|^2 \leq 1-\phi_X(x) \leq 2\cdot\frac{(2\pi)^2}{2d}|x|^2 \ .$$

The claim of the theorem follows because the integral $\int_{\{|x|<\epsilon\}} \frac{1}{|x|^2}\ dx$ over the ball of radius ϵ in $\mathbb{R}^d$ is finite if and only if $d \geq 3$. □

21.4. We can now decide whether the random walker returns infinitely many times to 0 or not.

Theorem 2. *The walker returns to the origin infinitely often almost surely if $d \leq 2$. For $d \geq 3$, the walker almost surely returns only finitely many times and* $\mathrm{P}[\lim_{n\to\infty} |S_n| = \infty] = 1$.

Proof. If $d > 2$, then $A_\infty = \limsup_n A_n$ is the subset of Ω, for which the particles returns to 0 infinitely many times. Since $\mathrm{E}[B] = \sum_{n=0}^{\infty} \mathrm{P}[A_n]$, the Borel-Cantelli lemma gives $\mathrm{P}[A_\infty] = 0$ for $d > 2$. The particle returns therefore back to 0 only finitely many times and in the same way it visits each lattice point only finitely many times. This means that the particle eventually leaves every bounded set and converges to infinity. If $d \leq 2$, let $p = \mathrm{P}[\bigcup_n A_n]$ be the probability that the random walk returns to 0. Then p^{m-1} is the probability that there are at least m visits in 0 and the probability is $p^{m-1} - p^m = p^{m-1}(1-p)$ that there are exactly m visits. We can write

$$\mathrm{E}[B] = \sum_{m\geq 1} mp^{m-1}(1-p) = \frac{1}{1-p} \ .$$

Because $\mathrm{E}[B] = \infty$, we know that $p = 1$. □

PROBABILITY THEORY

MATH 154

Unit 22: Markov Chains

22.1. A discrete time **Markov process** is a stochastic process X_i such that the outcome of X_{n+1} given the past only depends on X_n for every n. We will next week rephrase this in Martingale language. For now we look at an important simple case, where the random variables X_i take only finitely many values S, the set of **states**. A **Markov chain** is a stochastic process with values in S such that the conditional probability $\mathrm{P}[X_{n+1} = x_{n+1}|X_1 = x_1, \ldots, X_n = x_n]$ is $\mathrm{P}[X_{n+1} = x_{n+1}|X_n = x_n]$. If the probabilities $P[X_{n+1} = a|P[X_n = b]$ are independent of n, we talk about a **time homogeneous Markov** chain.

22.2. It follows from the definition of a Markov process that X_n satisfies the **elementary Markov property**: for $n > k$,

$$\mathrm{P}[X_n \in B \mid X_1, \ldots, X_k] = \mathrm{P}[X_n \in B \mid X_k] \; .$$

This means that the probability distribution of X_n is determined by knowing the probability distribution of X_{n-1}. The future depends only on the present and not on the past. In the time homogeneous case, the stochastic process defines a transformation T on a probability space $(S^{\mathbb{N}}, \mathcal{A} = \mathcal{B}^{\mathbb{N}})$, where $\mathcal{B}$ is the set of all subsets of S. As we will see, there are often measures π on S such that $P = \pi^{\mathbb{N}}$ is invariant. We want to understand such equilibria.

22.3. We now look at a homogeneous Markov chain on a finite state space S with s elements. Probability measures on S are vectors p with entries $p_i \geq 0$ such that $\sum_i p_i = 1$. The Markov chain is now determined by the left stochastic $s \times s$ matrix $M_{ij} = \mathrm{P}[X_{n+1} = j|X_n = x_i]$. [1] The matrix M^T has the eigenvalue 1 with eigenvector $[1, \ldots, 1]$. Therefore, M has also an eigenvalues 1. Its eigenvector is a **stationary measure** describing a stable probability distribution. As Oskar Perron in 1907 and Georg Frobenius in 1908 have shown there is one if M has positive entries:

Theorem 1 (Perron-Frobenius)**.** *If all entries of a left stochastic $n \times n$ matrix A are positive, there is a unique eigenvector to the eigenvalue* 1*.*

Proof. The set $X = \{\sum_i x_i^2 = 1, x_1 \geq 0, \ldots, x_n \geq 0\}$ is closed and bounded. If the entries of A are non-negative, the map $T(v) = Av/|Av|$ maps X to itself. The Brouwer fixed point theorem gives then already fixed point and so an eigenvector to the eigenvalue 1. If A has positive entries then TX is even contained in the interior

[1] Sometimes, right stochastic matrices are used. Matrix multiplication is applied to the right to row vectors.

of X. This fixed point is unique because the map is a **contraction**: There exists $0 < \lambda(x) < 1$ such that $d(Tx, Ty) \leq \lambda d(x, y)$, where d is the geodesic sphere distance. If there was a contraction with a uniform λ we could have used the Banach fixed point theorem. We do not need it: assume $T(x) = x$ and $T(y) = y$ are both fixed points, then the contraction property gives $d(x, y) = d(Tx, Ty) \leq \lambda(x)d(x, y) < d(x, y)$ a contradiction. We have now a unique fixed point $Av = \lambda v$ provided v has non-negative entries. Assume $Aw = w$ and w has some negative entry and $||w|| = 1$. Write $|w|$ for the vector with coordinates $|w_j|$. The computation

$$|w|_i = |w_i| = |\sum_j A_{ij} w_j| \leq \sum_j |A_{ij}||w_j| = \sum_j A_{ij}|w_j| = (A|w|)_i$$

shows that $(A|w|)$ is a vector with norm smaller or equal than 1. For any i with $w_i < 0$ we have an inequality so that $||Aw|| < 1$ contradicting $||w|| = 1$. The only eigenvectors to the eigenvalue 1 must be in X where we had a unique one. □

22.4. T is a true contraction with respect to the **Hilbert metric** on X. One can then use directly the **Banach fixed point theorem**. There is also a connection to ergodic theory. Given an initial measure μ_1 on S, the map M defines measures μ_k on S.

Theorem 2. *A Markov chain defines a measure preserving map T on the product probability space $(\Omega, \mathcal{A}) = (S^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, \prod_k \mu_k)$.*

Proof. The product space $(\Omega, \mathcal{A}) = (S^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$ has the π-system $\mathcal{C}$ consisting of cylinder-sets $\prod_{n\in\mathbb{N}} B_n$ given by a sequence $B_n \in \mathcal{B}$ such that $B_n = S$ except for finitely many n. The $P = P_\mu$ on $(\Omega, \mathcal{C})$ is the product measure. This measure has a unique extension to the σ-algebra $\mathcal{A}$. The shift map T on Ω is measure preserving. □

22.5. If M has positive entries and μ is the stable distribution, then T is the shift on $(S^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, \mu^{\mathbb{N}})$ and the random variables $X_i(x) = x_i$ are independent.

22.6. For a countable state space S one is in a random walk situation The transition matrix M_{ij} then now a bounded linear operator on $l^2(S)$. We have seen in the last lecture that if $S = \mathbb{Z}^d$ and M is a scaled version of the adjacency matrix one could use Fourier theory to understand recurrence. In the infinite case like $S = \mathbb{Z}$, there is no equilibrium measure. The probability distribution of the walker diffuses like a solution of the heat equation. We can still look at $M^n\mu$, where μ is an initial probability measure and study its dynamics. On a translational invariant lattice the walk is also a sum of IID random variables $S_n = X_1 + X_2 + \cdots X_n$, where X_i take finitely many values. Since by the central limit theorem, the variance of S_n grows linearly in time n, the standard deviation grows like $\sqrt{n}$.

22.7. Given a finite stochastic matrix M and a point $x \in S$, the measures $P(x, \cdot)$ are the probability vectors, which are the columns of M. It is also denoted **Markov field**. We have $P^n(x, B) = \sum_{y\in B} P^n(x, y)$. We can see the transition probability functions also as elements in $\mathcal{L}(S, M_1(S))$, by thinking about each column as a probability measure in the set $M_1(S)$ of Borel probability measures on S. This point of view is often taken in economics.

PROBABILITY THEORY

MATH 154

Unit 23: Conditional Expectation

23.1. Conditional probability $\mathrm{P}[A|B]$ for events leads to **conditional expectation** for σ algebras: it is denoted $\mathrm{E}[X|\mathcal{B}]$ if $\mathcal{B} \subset \mathcal{A}$ is a sub-σ-algebra.

Theorem 1 (Kolmogorov)**.** *Given $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathrm{P})$ and a sub σ-algebra $\mathcal{B} \subset \mathcal{A}$. There exists a random variable $Y \in \mathcal{L}^1(\Omega, \mathcal{B}, \mathrm{P})$ denoted $\mathrm{E}[X|\mathcal{B}]$ satisfying $\int_A Y\, dP = \int_A X\, dP$ for all $A \in \mathcal{B}$. We call it $\mathrm{E}[X|\mathcal{B}]$*

Proof. For $X = \sum_i a_i 1_{A_i} \in \mathcal{S}$ define $\mathrm{E}[X; A] = \sum_i a_i 1_{A_i \cap A}/\mathrm{P}[A]$. For $X \in \mathcal{L}^1$ define $\mathrm{E}[X; A]$ as a limit. We also write $\int_A X\, d\mathrm{P}$ for this conditional integration. Define the two measures $\tilde{\mathrm{P}}[A] = \mathrm{P}[A]$ and $\mathrm{P}'[A] = \int_A X\, dP = \mathrm{E}[X; A]$ on the measure space $(\Omega, \mathcal{B})$. Given a set $B \in \mathcal{B}$ with $\tilde{P}[B] = 0$, then $\mathrm{P}'[B] = 0$ so that P' is absolutely continuous with respect to $\tilde{P}$. The **Radon-Nykodym** theorem from real analysis gives a random variable $Y \in \mathcal{L}^1(\mathcal{B})$ with $\mathrm{P}'[A] = \int_A X\, dP = \int_A Y\, dP$. □

23.2. Examples:
a) if $\mathcal{B} = \{\emptyset, \Omega\}$, then $\mathrm{E}[X|\mathcal{B}] = \mathrm{E}[X]$.
b) if $\mathcal{B} = \{\emptyset, \Omega, B, B^c\}$ then $E[X|\mathcal{B}]$ takes the value $\int_B X\, dP/\mathrm{P}[B]$ on B and the value $\int_{B^c} X\, dP/\mathrm{P}[B^c]$ on B^c.
c) if $\mathcal{B} = \mathcal{A}$, then $E[X|\mathcal{B}] = X$.
d) if $\mathcal{B} = \mathcal{A}_X$ is the σ algebra generated by X, then $E[X|\mathcal{B}] = X$.
e) if $X(x, y)$ is a continuous function on the unit square $\Omega = [0, 1]^2$ with $\mathrm{P} = dxdy$ as a probability measure and where $Y(x, y) = x$. In that case, $\mathrm{E}[X|Y]$ is a function of x alone, given by $\mathrm{E}[X|Y](x) = \int_0^1 f(x, y)\, dy$. It is called **a conditional integral**.

23.3. The map $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathrm{P}) \to \mathrm{E}[X, \mathcal{B}] \in \mathcal{L}^1(\Omega, \mathcal{B}, \mathrm{P})$ is a **projection**. To see this geometrically, we work in the Hilbert space $\mathcal{L}^2$:

Theorem 2. *Conditional expectation $X \mapsto \mathrm{E}[X|\mathcal{B}]$ is the projection $\mathcal{L}^2(\mathcal{A}) \to \mathcal{L}^2(\mathcal{B})$.*

Proof. The space $\mathcal{L}^2(\mathcal{B})$ of square integrable $\mathcal{B}$-measurable functions is a linear subspace of $\mathcal{L}^2(\mathcal{A})$. When identifying functions which agree almost everywhere, then $L^2(\mathcal{B})$ is a Hilbert space which is a linear subspace of the Hilbert space $L^2(\mathcal{A})$. For any $X \in \mathcal{L}^2(\mathcal{A})$, there exists a unique projection $p(X) \in \mathcal{L}^2(\mathcal{B})$. The orthogonal complement $\mathcal{L}^2(\mathcal{B})^\perp$ is defined as

$$\mathcal{L}^2(\mathcal{B})^\perp = \{ Z \in \mathcal{L}^2(\mathcal{A}) \mid (Z, Y) := \mathrm{E}[Z \cdot Y] = 0 \text{ for all } Y \in \mathcal{L}^2(\mathcal{B}) \} .$$

By the definition of the conditional expectation, we have for $A \in \mathcal{B}$

$$(X - \mathrm{E}[X|\mathcal{B}], 1_A) = \mathrm{E}[X - \mathrm{E}[X|\mathcal{B}]; A] = 0 \; .$$

Therefore $X - \mathrm{E}[X|\mathcal{B}] \in \mathcal{L}^2(\mathcal{B})^\perp$. Because the map $q(X) = \mathrm{E}[X|\mathcal{B}]$ satisfies $q^2 = q$, it is linear and has the property that $(1-q)(X)$ is perpendicular to $\mathcal{L}^2(\mathcal{B})$, the map q is a projection which must agree with p. □

Theorem 3 (Properties)**.** *For all $X, X_n, Y \in \mathcal{L}^1$:*
(1) Linearity: The map $X \mapsto \mathrm{E}[X|\mathcal{B}]$ is linear.
(2) Positivity: $X \geq 0 \Rightarrow \mathrm{E}[X|\mathcal{B}] \geq 0$.
(3) Tower property: $\mathcal{C} \subset \mathcal{B} \subset \mathcal{A} \Rightarrow \mathrm{E}[E[X|\mathcal{B}]|\mathcal{C}] = \mathrm{E}[X|\mathcal{C}]$.
(4) Cond. Fatou: $|X_n| \leq X$, $\mathrm{E}[\liminf_{n\to\infty} X_n|\mathcal{B}] \leq \liminf_{n\to\infty} \mathrm{E}[X_n|\mathcal{B}]$.
(5) Cond. dominated convergence: $|X_n| \leq X, X_n \to X$ a.e. $\Rightarrow \mathrm{E}[X_n|\mathcal{B}] \to \mathrm{E}[X|\mathcal{B}]$ a.e.
(6) Cond. Jensen: if h is convex, then $\mathrm{E}[h(X)|\mathcal{B}] \geq h(E[X|\mathcal{B}])$.
(7) Especially $||\mathrm{E}[X|\mathcal{B}]||_p \leq ||X||_p$.
(9) Extracting knowledge: For $Z \in \mathcal{L}^\infty(\mathcal{B})$, one has $\mathrm{E}[ZX|\mathcal{B}] = Z\mathrm{E}[X|\mathcal{B}]$.
(9) Independence: if X is independent of $\mathcal{C}$, then $\mathrm{E}[X|\mathcal{C}] = \mathrm{E}[X]$.

Proof. (1) The conditional expectation is a projection by the previous theorem so linear.
(2) If $Y = \mathrm{E}[X|\mathcal{B}]$ would be negative on a set of positive measure, then $A = Y^{-1}((-\infty, -1/n]) \in \mathcal{B}$ would have positive probability for some n. This would lead to the contradiction $0 \leq \mathrm{E}[1_A X] = \mathrm{E}[1_A Y] \leq -n^{-1}m(A) < 0$.
(3) Use that $P'' \ll P' \ll P$ implies $P'' = Y'P' = Y'YP$ and $P'' \ll P$ gives $P'' = ZP$ so that $Z = Y'Y$ almost everywhere.
This is especially useful when applied to the algebra $\mathcal{C}_Y = \{\emptyset, Y, Y^c, \Omega\}$. Because $X \leq Y$ almost everywhere if and only if $\mathrm{E}[X|\mathcal{C}_Y] \leq \mathrm{E}[Y|\mathcal{C}_Y]$ for all $Y \in \mathcal{B}$.
(4)-(5) The conditional versions of the Fatou lemma or the dominated convergence theorem are true, if they are true conditioned with $\mathcal{C}_Y$ for each $Y \in \mathcal{B}$. The tower property reduces these statements to versions with $\mathcal{B} = \mathcal{C}_Y$ which are then on each of the sets Y, Y^c the usual theorems.
(6) Chose a sequence $(a_n, b_n) \in \mathbb{R}^2$ such that $h(x) = \sup_n a_n x + b_n$ for all $x \in \mathbb{R}$. We get from $h(X) \geq a_n X + b_n$ that almost surely $\mathrm{E}[h(X)|\mathcal{G}] \geq a_n \mathrm{E}[X|\mathcal{G}] + b_n$. These inequalities hold therefore simultaneously for all n and we obtain almost surely

$$\mathrm{E}[h(X)|\mathcal{G}] \geq \sup_n (a_n \mathrm{E}[X|\mathcal{G}] + b_n) = h(E[X|\mathcal{G}]) \; .$$

(7) This is a special case of (6) using $h(x) = |x|^p$.
(8) It is enough to condition it to each algebra $\mathcal{C}_Y$ for $Y \in \mathcal{B}$. The tower property reduces these statements to linearity.
(9) By linearity, we can assume $X \geq 0$. For $B \in \mathcal{B}$ and $C \in \mathcal{C}$, the random variables $X1_B$ and 1_C are independent so that $\mathrm{E}[X1_{B\cap C}] = \mathrm{E}[X1_B 1_C] = \mathrm{E}[X1_B]\mathrm{P}[C]$. The random variable $Y = \mathrm{E}[X|\mathcal{B}]$ is $\mathcal{B}$ measurable and because $Y1_B$ is independent of $\mathcal{C}$ we get $\mathrm{E}[(Y1_B)1_C] = \mathrm{E}[Y1_B]\mathrm{P}[C]$ so that $\mathrm{E}[1_{B\cap C}X] = \mathrm{E}[1_{B\cap C}Y]$. The measures on $\sigma(\mathcal{B}, \mathcal{C})$

$$\mu : A \mapsto \mathrm{E}[1_A X], \nu : A \mapsto \mathrm{E}[1_A Y]$$

agree therefore on the π-system of the form $B \cap C$ with $B \in \mathcal{B}$ and $C \in \mathcal{C}$ and consequently everywhere on $\sigma(\mathcal{B}, \mathcal{C})$. □

PROBABILITY THEORY

MATH 154

Unit 24: Martingales

24.1. A sequence $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$ of sub σ-algebras of $\mathcal{A}$ is called a **filtration**, if $\mathcal{A}_0 \subset \mathcal{A}_1 \subset \cdots \subset \mathcal{A}$. Given a filtration $\{\mathcal{A}_n\}_{n\in\mathbb{N}}$, one gets a **filtered space** $(\Omega, \mathcal{A}, \{\mathcal{A}_n\}_{n\in\mathbb{N}}, \mathrm{P})$.

24.2. A **discrete time stochastic process** $X = \{X_n\}_{n\in\mathbb{N}}$ is called **adapted to a filtration** $\{\mathcal{A}_n\}$ if X_n is $\mathcal{A}_n$-measurable for all $n \in \mathbb{N}$.

24.3. A $\mathcal{L}^1$-process which is adapted to a filtration $\{\mathcal{A}_n\}$ is called a **martingale** if

$$\mathrm{E}[X_n|\mathcal{A}_{n-1}] = X_{n-1}$$

for all $n \geq 1$. It is called a **super-martingale** if $\mathrm{E}[X_n|\mathcal{A}_{n-1}] \leq X_{n-1}$ and a **sub-martingale** if

$$\mathrm{E}[X_n|\mathcal{A}_{n-1}] \geq X_{n-1} \ .$$

If we mean either sub-martingale or super-martingale (or martingale) we speak of a **semi-martingale**.

24.4. It immediately follows that for a martingale

$$\mathrm{E}[X_n|\mathcal{A}_m] = X_m$$

if $m < n$ and that $\mathrm{E}[X_n]$ is constant.

Allan Gut mentions in his book that a martingale is an allegory for "life" itself: *the expected state of the future given the past history is equal the present state and on average, nothing happens.* The word "martingale" originally denoted a gambling system strategy in which losing bets are doubled. It is also the name of a part of a horse's harness or a belt on the back of a man's coat.

24.5. If a martingale X_n is given with respect to a filtered space $\mathcal{A}_n = \sigma(Y_0, \ldots, Y_n)$, where Y_n is a given process, then X is called a **martingale with respect** Y.

24.6. If X is a super-martingale, then $-X$ is a sub-martingale and vice versa. A super-martingale, which is also a sub-martingale is a martingale. Since we can change X to $X - X_0$ without destroying any of the martingale properties, we could assume the process is **null at** 0 which means $X_0 = 0$.

24.7. Given a martingale. From the tower property of conditional expectation follows that for $m < n$

$$\mathrm{E}[X_n|\mathcal{A}_m] = \mathrm{E}[E[X_n|\mathcal{A}_{n-1}]|\mathcal{A}_m] = \mathrm{E}[X_{n-1}|\mathcal{A}_m] = \cdots = X_m \ .$$

24.8. Sum of independent random variables
Let $X_i \in \mathcal{L}^1$ be a sequence of independent random variables with mean $\mathrm{E}[X_i] = 0$. Define $S_0 = 0, S_n = \sum_{k=1}^n X_k$ and $\mathcal{A}_n = \sigma(X_1, \ldots, X_n)$ with $\mathcal{A}_0 = \{\emptyset, \Omega\}$. Then S_n is a martingale since S_n is an $\{\mathcal{A}_n\}$-adapted $\mathcal{L}^1$-process and

$$\mathrm{E}[S_n|\mathcal{A}_{n-1}] = \mathrm{E}[S_{n-1}|\mathcal{A}_{n-1}] + \mathrm{E}[X_n|\mathcal{A}_{n-1}] = S_{n-1} + \mathrm{E}[X_n] = S_{n-1} \; .$$

We have used linearity and the independence property of the conditional expectation.

24.9. Example a) Conditional expectation
Given a random variable $X \in \mathcal{L}^1$ on a filtered space $(\Omega, \mathcal{A}, \{\mathcal{A}_n\}_{n\in\mathbb{N}}, \mathrm{P})$. Then $X_n = \mathrm{E}[X|\mathcal{A}_n]$ is a martingale.
Especially: given a sequence Y_n of random variables. Then $\mathcal{A}_n = \sigma(Y_0, \ldots, Y_n)$ is a filtered space and $X_n = \mathrm{E}[X|Y_0, \ldots, Y_n]$ is a martingale. Proof: by the tower property

$$\begin{aligned} \mathrm{E}[X_n|\mathcal{A}_{n-1}] &= \mathrm{E}[X_n|Y_0, \ldots, Y_{n-1}] \\ &= \mathrm{E}[\mathrm{E}[X|Y_0, \ldots, Y_n]|Y_0, \ldots, Y_{n-1}] \\ &= \mathrm{E}[X|Y_0, \ldots, Y_{n-1}] = X_{n-1} \; . \end{aligned}$$

verifying the martingale property $\mathrm{E}[X_n|\mathcal{A}_{n-1}] = X_{n-1}$.
We say X is a **martingale with respect to** Y. Note that because X_n is by definition $\sigma(Y_0, \ldots, Y_n)$-measurable, there exist Borel measurable functions $h_n : \mathbb{R}^{n+1} \to \mathbb{R}$ such that $X_n = h_n(Y_0, \ldots, Y_{n-1})$.

24.10. Example b) Product of positive variables
Given a sequence Y_n of independent random variables $Y_n \geq 0$ satisfying with $\mathrm{E}[Y_n] = 1$. Define $X_0 = 1$ and $X_n = \prod_{i=0}^n Y_i$ and $\mathcal{A}_n = \sigma(Y_1, \ldots, Y_n)$. Then X_n is a martingale. This is an exercise. Note that the martingale property does not follow directly by taking logarithms.

24.11. Example c) Product of matrix-valued random variables
Given a sequence of independent random variables Z_n with values in the group $GL(N, \mathbb{R})$ of invertible $N \times N$ matrices and let $\mathcal{A}_n = \sigma(Z_1, \ldots, Z_n)$. Assume $\mathrm{E}[\log ||Z_n||] \leq 0$, if $||Z_n||$ denotes the norm of the matrix (the square root of the maximal eigenvalue of $Z_n \cdot Z_n^*$, where Z_n^* is the adjoint). Define the real-valued random variables $X_n = \log ||Z_1 \cdot Z_2 \cdots Z_n||$, where $\cdot$ denotes matrix multiplication. Because $X_n \leq \log ||Z_n|| + X_{n-1}$, we get

$$\begin{aligned} \mathrm{E}[X_n|\mathcal{A}_{n-1}] &\leq \mathrm{E}[\log ||Z_n|| \mid \mathcal{A}_{n-1}] + \mathrm{E}[X_{n-1}|\mathcal{A}_{n-1}] \\ &= \mathrm{E}[\log ||Z_n||] + X_{n-1} \leq X_{n-1} \end{aligned}$$

so that X_n is a super-martingale. In ergodic theory, such a matrix-valued process X_n is called **sub-additive**.

24.12. Example d) If Z_n is a sequence of matrix-valued random variables, we can also look at the sequence of random variables $Y_n = ||Z_1 \cdot Z_2 \cdots Z_n||$. If $\mathrm{E}[||Z_n||] = 1$, then Y_n is a super-martingale.

PROBABILITY THEORY

MATH 154

Homework 1

Probability

Problem 1.1: a) You pick a random point (x, y) in the square $[-1, 1] \times [-1, 1]$. What is the probability that $x^2 + y^2 \leq 1$?
b) You pick a random point (x, y, z) in the unit cube $[-1, 1]^3$. What is the probability that $x^2 + y^2 + z^2 \leq 1$?
c) What is the probability that $x_1^2 + x_2^2 + ... + x_{1000}^2 \leq 1$ if the point $x = (x_1, \ldots, x_{1000})$ is chosen randomly in the 1000-dimensional unit cube $[-1, 1]^{1000}$.

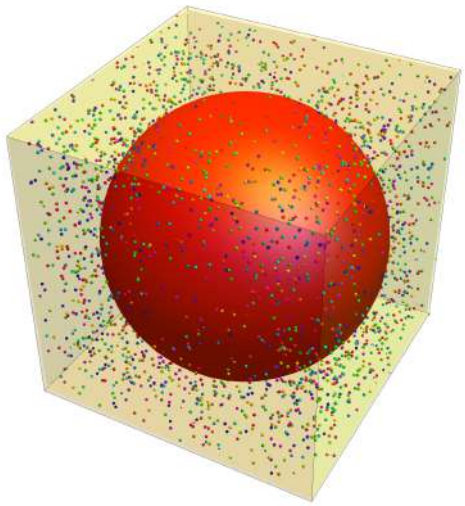


Figure 1. What is the probability to hit the sphere?

Problem 1.2: The card game "set" contains $81 = 3^4$ cards. Each card has one of 3 colors, one of 3 numbers, one of 3 shapes and one of 3 shades. It so models so the 4-dimensional vector space $\mathbb{Z}_3^4$ which is also called the field $GF(81)$. A collection of three cards is called a "set", if in each of the 3 categories, all three properties either agree or are all different. You randomly pick 3 cards from the 81. What is the probability to draw a set?

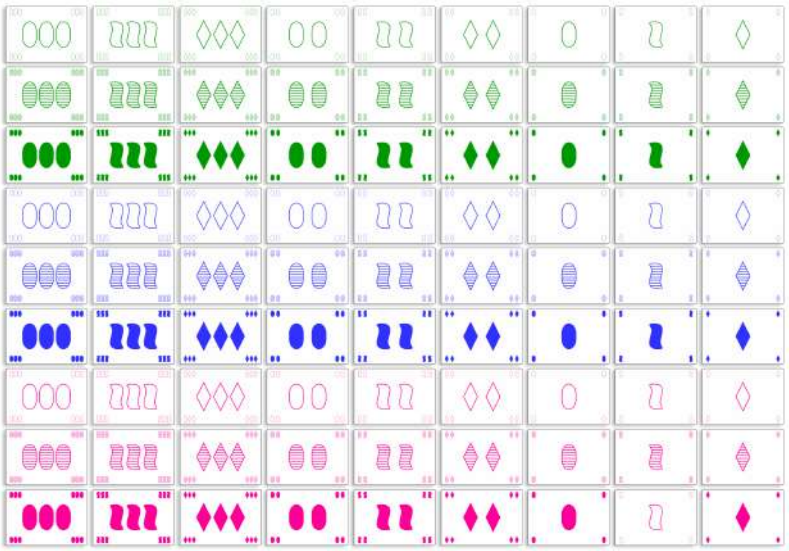


FIGURE 2. The game of set visualizes a 4 dimensional vector space

Problem 1.3: The probability density of a positive integer smaller than n is prime is about $1/\log(n)$ by the prime number theorem. What do you expect is the expected number of **prime twins** smaller than n?

Problem 1.4: a) Alex has three kids, and one of them is a girl. What is the probability that Alex has three girls?
b) Alex has three kids of different age and the oldest is a girl. What is the probability that Alex has three girls?

Problem 1.5: There are three boxes: a box containing two gold coins, a box containing two silver coins, and a box containing one gold coin and one silver coin. The three boxes are shuffled. You pick one box and pick a random coin from it. You notice it to be gold. What is the probability that the other coin from the same box is gold?

PROBABILITY THEORY

MATH 154

Homework 2

Probability spaces

Problem 2.1: Verify the following properties from the axioms.
a) $\mathrm{P}[\emptyset] = 0$.
b) $A \subset B \Rightarrow \mathrm{P}[A] \leq \mathrm{P}[B]$.
c) $\mathrm{P}[\bigcup_n A_n] \leq \sum_n \mathrm{P}[A_n]$.
d) $\mathrm{P}[A^c] = 1 - \mathrm{P}[A]$.
e) $0 \leq \mathrm{P}[A] \leq 1$.
f) $A_1 \subset A_2, \subset \cdots$ with $A_n \in \mathcal{A}$
then $\mathrm{P}[\bigcup_{n=1}^{\infty} A_n] = \lim_{n\to\infty} \mathrm{P}[A_n]$.

Problem 2.2: Let Ω be a set. Let $\mathcal{A}$ be the set of countable or co-countable subsets of Ω.
a) Verify that $\mathcal{A}$ satisfies all the ring axioms of Boolean algebra.
b) Verify that $\mathcal{A}$ is a π-system.
c) Verify that $\mathcal{A}$ is a λ-system.
d) Verify that $\mathcal{A}$ is a σ algebra without using the theorem of Lecture 3.
e) Verify that $\mathcal{A}$ is the smallest σ algebra containing the cofinite topology.

Problem 2.3: Let $\Omega = [0,1]^2$. Let $\mathcal{I} = \{[a,b) \times [c,d)\}$ denote the set of all left-bottom closed right-top open rectangles.
a) Verify that this is a π-system.
b) Verify that $\mathrm{P}[a,b) \times [c,d)] = (d-c)(b-a)$ is a probability measure on this π system.
c) Why can the measure P be extended to the smallest σ-algebra containing $\mathcal{I}$?
d) Under which conditions are two elements in $\mathcal{I}$ independent?

Problem 2.4: Verify the following properties. The first four are known as **Keynes postulates**, the fifth is called **Bayes Theorem**.
1) $\mathrm{P}[A|B] \geq 0$.
2) $\mathrm{P}[A|A] = 1$.
3) $\mathrm{P}[A|B] + \mathrm{P}[A^c|B] = 1$.
4) $\mathrm{P}[A \cap B|C] = \mathrm{P}[A|C] \cdot \mathrm{P}[B|A \cap C]$.
5) $\mathrm{P}[A|B| = \mathrm{P}[B|A]\mathrm{P}[A]/\mathrm{P}[B]$.

Problem 2.5: Prove the ΠΣΛ **sorority theorem** in the text. It states "The smallest λ-system $\mathcal{A}$ containing a π-system $\mathcal{I}$ is the smallest σ algebra containing $\mathcal{I}$."

FIGURE 1. To the left an example of a ΠΣΛ chapter (in this case Oxford MS). To the right, a brooch from BU in the shape of a Marguerite daisy (or $A \cap B$ when intersecting two sets in a Venn Diagram) also in the order of the mathematical order ΠΛΣ: to check that we have a σ-algebra, we have to check it is a π-system and a λ-system.

PROBABILITY THEORY

MATH 154

Homework 3

Random Variables

Problem 3.1: The **Gamma distribution** with shape $\alpha > 0$ and rate $\lambda > 0$ has support on $[0, \infty)$. It is used in econometrics. The probability density function is

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \lambda^{\alpha} .$$

a) What distribution do we get in the case $\alpha = 1$?
b) Verify that f satisfies the properties of a PDF.
c) Compute the expectation $\mathrm{E}[X]$ and variance $\mathrm{Var}[X]$.
d) Compute the moment generating function $M_X(t)$.
e) Why is a Gamma distributed random variable in $\mathcal{L}^p$ for all p?

Problem 3.2: Verify that for $\theta > 0$ the **Maxwell distribution**

$$f(x) = \frac{4}{\sqrt{\pi}} \theta^{3/2} x^2 e^{-\theta x^2}$$

is a PDF of a probability distribution on $\mathbb{R}^+ = [0, \infty)$. This distribution can model the speed distribution of molecules in thermal equilibrium. Now compute its expectation $\mathrm{E}[X] = \int_0^\infty x f(x)\, dx$.

Problem 3.3: **Benford's law** deals with the statistics of the first significant digit in data. Simon Newcomb found the law in 1881 and Frank Benford made significant progress to understand it in 1938. The distribution appears also in naturally occurring sequences. For example, if you look at the first digit of the sequence 2^n then the first significant digit k appears with probability $p_k = \log_{10}(1 + 1/k)$. The digit 1 for example occurs with about $\log_{10}(2) = 0.30$ which is 30 percent.
a) What is its expectation and variance of the distribution?
b) Verify that the sequence 2^n produces this distribution.

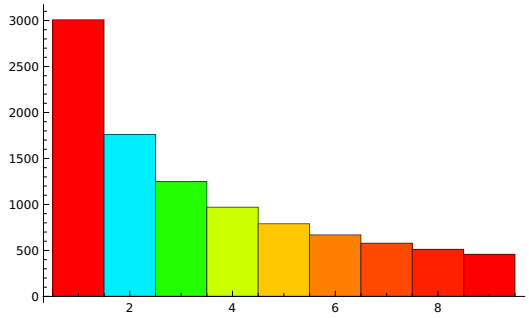


FIGURE 1. The Benford distribution for the first significant digit. It is computed with Histogram[Table[First[IntegerDigits[2^n]], {n, 1, 10000}], 10]

Problem 3.4: For a centered Cauchy distributed random variable, the probability density is $(\frac{1}{\pi}/(1+x^2)$. As seen in class you can generate random variables with this distribution. Define $X(x) = x$ on $(\Omega = \mathbb{R}, \mathcal{B}, P = f(x)dx)$.
a) Check that the random variable X is not in $\mathcal{L}^1$.
b) Look up the definition of convergence in the sense of Cauchy and verify that the expectation of the distribution in this generalized sense.
c) What can you say about the variance and higher moments or moment generating function of a Cauchy distributed random variable?
d) Why again does $Cot(PiRandom[])$ generate Cauchy distributed random variables?

Problem 3.5: The **support** K of the law μ of a random variable is the largest closed subset of $\mathbb{R}$ such that $\mu((x-a, x+a)) > 0$ for every $x \in K$ and $a > 0$.
a) There are absolutely continuous distribution functions for which the support is a Cantor set on $[0, 1]$. Construct one. (Note that this can not be the standard Cantor set because the Standard Cantor set has measure zero.)
b) There are singular continuous distributions for which the support is $[0, 1]$. Construct one.
c) There are pure point distributions for which the support is $[0, 1]$. Construct one.
d) Verify that for every closed set K in $[0, 1]$ there exists a measure which has K as support.

PROBABILITY THEORY

MATH 154

Homework 4

Independence

Problem 4.1: Ana and Bob own a business OAP (a pun to $(\Omega, \mathcal{A}, \mathrm{P})$) which builds custom designed dice. The customer wants a dice with a given probability distribution. OAP delivers 3D printed dice.
Ana spends 40 percent of her day in meetings, while Bob spends 25 percent of his day in meetings. They schedule their meetings independently.
a) What is the probability that both meet at the same time?
b) What is the probability that Ana has a meeting during a time that Bob has a meeting?
c) What is the probability that Bob has a meeting during a time when Ana has a meeting?
d) Is the event that both have a meeting at the same time independent of the event that both have no meeting at the same time?

Figure 1. Palindromes Ana and Bob meet to discuss the design of new dice. (AI generated picture)

Problem 4.2: True or False? (Please give justifications).
1) If A, B are independent, then A, B^c are independent.
2) If $\mathrm{P}[B] > 0$, and A, B are independent, then $\mathrm{P}[A|B] = \mathrm{P}[A]$.
3) If A, B are independent and B, C are independent then A, C are.
4) If A, B, C are independent, then $A + B$ is independent of C.
5) If A, B, C are independent, then $A \cap B$ is independent of C.
6) If A, B, C are independent then $A \cup B$ is independent of C.
7) Two disjoint sets A, B are independent if and only if $P[A] = 0$ or $P[B] = 0$.
8) $\emptyset$ is independent of any other set.
9) Ω is independent of any other set.
10) If A is independent to itself, then $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$.

Problem 4.3: If $(\Omega, \mathcal{A}, \mathrm{P})$ has a $\mathcal{P}$ trivial σ-algebra, you might think that $\mathcal{A}$ is the trivial σ-algebra. This is not the case as you verify here with an example:
Verify that the σ algebra of cocountable or countable sets in $\Omega = [0, 1]$ is $\mathcal{P}$-trivial, if $\mathcal{P} = \lambda$ is the probability Lebesgue measure on $[0, 1]$

Problem 4.4: In all of this problem, all random variables are bounded $\mathcal{L}^\infty$.
a) Verify that if X, Y are independent and n, m are positive integers, then X^n, Y^m are independent.
b) Verify that $X \cdot Y = \langle X, Y \rangle = \mathrm{E}[XY]$ defines an inner product on $\mathcal{L}^2$. Define $|X| = \sqrt{\langle X, X \rangle}$. Check **Cauchy-Schwarz** $|\langle X, Y \rangle| \leq |X||Y|$.
c) We have seen that if X, Y are independent $\mathcal{L}^2$ random variables, then $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$. Can you reverse this? Does the condition $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$ imply that X, Y are independent?
d) What about asking that $\mathrm{E}[X^n Y^m] = \mathrm{E}[X^n]\mathrm{E}[Y^m]$ for all $n, m > 0$? Does this imply that X, Y are independent?

Problem 4.5: a) Verify that the moment generating function of the Cauchy distribution does not exist.
b) Compute the characteristic function $\phi_X(t)$ of a Cauchy distributed random variable.
c) Compute the characteristic function of the Gaussian distribution with probability density function $f(x) = e^{-x^2}/\sqrt{\pi}$.
d) Find a probability space and a random variable X such that $\phi_X(t) = \cos(t)$.

PROBABILITY THEORY

MATH 154

Homework 5

Tail algebra

Problem 5.1: Bond percolation in 3 dimensions. Ω is the set of all subgraphs of the lattice $\mathbb{Z}^3$ with nearest neighbor connections. Look at σ-algebras $\mathcal{A}_e$ generated by the random variable $X_e(\omega) = 1_{\{e \in E(\omega)\}}$. The assumption $\mathrm{P}[\{X_e = 1\}] = p, \mathrm{P}[\{X_e = 0\}]$ defines a probability space in which $\{X_e\}_{e \in E}$ are independent.
a) What theorem does assure that we have a probability measure on Ω that is translation invariant?
b) The event A consists of all graphs for which there is an infinite cluster. Verify that $\mathrm{P}_p[A] \leq \mathrm{P}_q[A]$ if $p \leq q$.
c) Conclude there is a threshold p_c so that $p > p_c$ gives an infinite cluster and $p < p_e$ none with probability 1.
d) Hit the literature: what is currently the best estimate for p_c?

Jensen

Problem 5.2: a) Formulate Jensen inequality in the case $f(x) = |x|$ and show that it implies the calculus identity $|\int_0^1 f(x)\, dx| \leq \int_0^1 |f(x)|\, dx$ for a continuous function on $[0, 1]$.
b) It implies the geometric-arithmetic mean inequality $\sqrt{ab} \leq (a + b)/2$.
c) Jensen's inequality can explain risk aversion and motivate portfolio optimization. Let ϕ be a concave utility function. ($-\phi$ is convex). What does Jensen tell you about the expected utility?

Entropy

Problem 5.3: We study entropy $S(\mathcal{A})$ calculus for a finite σ-algebra.
a) Single variable: $f(x) = x \log(1/x)$ is concave. The limit $f(0) = 0$ exists.
b) Multi: the uniform distribution on $\{1, \ldots, n\}$ has maximal entropy.
c) Let $\mathcal{A}_X$ be the σ-algebra of a random variable $X \in \mathcal{S}$ and $\mathcal{A}_{X,Y}$ the σ algebra of two random variables $X, Y \in \mathcal{S}$. Show that if X, Y are independent, then $S(\mathcal{A}_{X,Y}) = S(\mathcal{A}_X) + S(\mathcal{A}_Y)$.

Chebychev

Problem 5.4: You own an insurance company that gets random claims at random times. In order to have enough reserves, you want to estimate how large claims will be in the future. Your staff tells you the mean and standard deviation of the historical claim distribution but you do not know the distribution.
a) Why does Chebyshev's inequality imply that at least 89 percent of future claims will be within three standard deviations away from the mean?
b) Build a similar rule of thumb to see that $\boxed{}$ percent of future claims are within two standard deviations from the mean. Explain.
c) Fill in the box: 96 percent of future claims are within $\boxed{}$ standard deviations from the mean. Explain.

Problem 5.5: A probability space and random variable X defines what one calls a **null hypothesis**, the assumption that an effect does not exist. Assume you measure $X = c$ and that c is larger than the expectation, then the P**-value** of this experiment is defined as $\mathrm{P}[X \geq c]$. If the P-value is < 0.05, one considers the result as **significant** and rejects the null-hypothesis. If the P-value is > 0.05, one fails to reject the null hypothesis.
a) Assume a hypothesis is that X is exponentially distributed. You measure $X = 2$. What is the p-value?
b) Estimate the p-value using Chebyshev's inequality.
c) Having a p-value smaller than 0.05 is considered the gold standard for "statistical significance". Discuss the following strategy: we repeat an experiment a couple of times until the P-value is smaller than 5 percent. You label the early runs as warm-up-test runs and publish the paper.
d) Is it true that if you make a measurement and see the P value is larger than 0.05 that the non-significance means that the effect does exist? Explain in an example.

[1]

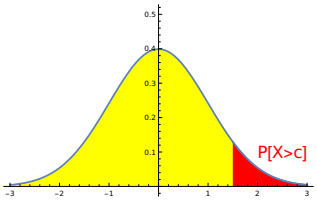


Figure 1. P-Value.

Oliver Knill, knill@math.harvard.edu, Math 154, Spring, 2025

[1] If c was smaller than the expectation, we would define the P-value as $\mathrm{P}[X \leq c]$.

PROBABILITY THEORY

MATH 154

Homework 6

Stochastic Convergence

Problem 6.1: Consider the random variables $X_n(x) = \cos(nx)$ on $[-\pi, \pi], \mathcal{B}, dx/(2\pi)$.
a) By writing $\cos(nx) = \mathrm{Re}(e^{inx})$ and using a geometric series, verify that

$$S_n(x) = \frac{2\sin((n+1/2)x)}{\sin(x/2)} - 1 \ .$$

This is $D_n(x) - 1$, where $D_n(x)$ is called the Dirichlet kernel.
b) Verify that $||S_n(x)||_1 + 1 \geq \frac{2}{\pi}\log(2n+1)$.
c) First recollect from class why the assumptions of the weak law are satisfied and restate the conclusion of that theorem about S_n/n. This should verify that $S_n/n \to 0$ in $\mathcal{L}^1$ and so in probability.
d) Given a continuous even function f with $E[f] = 0$, the expectation $a_n = E[fX_n]$ is called the n'th Fourier coefficient and $g(x) = \sum_n a_n X_n(x)$ is the cos-Fourier series of n. The formula $\sum_{n=1}^{\infty} a_n^2 = ||f||_2$ is called Parceval's identity. What geometric condition does assure it and what famous geometric theorem does it generalize?

Problem 6.2: a) Give an example of a sequence of random variables $X_n \to X$ for which we have convergence in probability but not complete convergence.
b) Give an example of a sequence of random variables X_n, where $X_n \to X$ in probability but where $X_n \to X$ in $\mathcal{L}^1$ does not happen.
c) Give an example of a sequence of random variables where $X_n \to X$ in $\mathcal{L}^1$ but where the convergence is not in $\mathcal{L}^2$.

Problem 6.3: a) Is there for $1 \leq p < \infty$ a relation between L^p convergence and convergence almost everywhere?
b) Is there a relation between L^∞ convergence and convergence almost everywhere?
c) Is there a relation between complete convergence and L^p convergence for $p < \infty$?
d) Is there a relation between complete convergence and L^∞ convergence?

Law of Large numbers

Problem 6.4: The n'th Chebyshev polynomial is defined as $X_n = T_n(x) = \cos(n \arccos(x))$. We have $T_n(\cos(t) = \cos(nt)$.
a) Verify that $T_n(x)$ is a polynomial of degree n and write down $T_n(x)$ for $n = 1, 2, 3, 4$.
b) We look at $T_n(x)$ as a random variable on the probability space $(\Omega = [-1, 1], \mathcal{B}, \mathrm{P} = \frac{1}{\pi\sqrt{1-x^2}})$. Check that the later indeed is a probability space.
c) Demonstrate (by showing all conditions) that you can use the weak law of large numbers to establish that $(1/n)S_n$ converges in probability to 0.

Problem 6.5: Let $X_n(\omega)$ be the n'th binary digit of $\omega \in [0, 1]$.
a) Investigate the convergence $\frac{S_n}{n} \to m$ in probability.
b) Verify that S_n has the Binomial distribution $p_k = \begin{pmatrix} n \\ k \end{pmatrix} 1/2^n$.
c) Show directly and then use the weak law to see that $S_n/n \to 0$ in probability.
d) Verify that $S_n/\sqrt{n}$ does not go to zero in L^2.
e) Verify also that $S_n/\sqrt{n}$ does not converge to 0 in distribution.

PROBABILITY THEORY

MATH 154

Homework 7

Strong Law and Birkhoff

Problem 7.1: Let $(\Omega = [0,1]^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, \mathrm{P}^{\mathbb{N}})$ denote the standard product Lebesgue probability space. Consider for $n \geq 1$ the sequence $a_n = 1/(n\log(n))$. Define $X_n(x) = n1_{[0,a_n/2](x_n)} - n1_{[1-a_n/2,1](x_n)}$. In other words, we have a sequence of random variables that take values $n, -n, 0$.
a) Check that X_n is a sequence of independent random variables of zero mean and variance $n/(\log(n))$.
b) Check that the proof of the weak law of large numbers still works so that $\mathrm{P}[S_n/n \geq \epsilon] \to 0$.
c) Verify that $\sum_n \mathrm{P}[\{X_n = n\}]$ diverges and conclude that with probability 1, we have $|S_n/n| \geq 1/2$ infinitely many often.
d) Conclude that X_n does not satisfy the strong law of large numbers.

Problem 7.2: Use the notes to write down the proof of the maximal ergodic theorem of Hopf. Make sure you understand every step.

Problem 7.3: Use the notes to write down the proof of the Birkhoff ergodic theorem. Make sure you understand every step.

Problem 7.4: Write down a paragraph about the history of Birkhoff's ergodic theorem. Especially make a connection with Harvard.

Problem 7.5: Given a real number α let $T : \mathbb{T} = \mathbb{R}/\mathbb{Z} \to \mathbb{T}$ be defined as $T(x) = x + \alpha$. A continuous function $f : \mathbb{T} \to \mathbb{R}$ defines so a sequence of random variables $X_n(x) = f(T^n(x)) = f(x + n\alpha)$.
a) If there exists a continuous function g such that $f(x) = g(x+\alpha) - g(x)$, we call f a coboundary). What can you say about the growth rate of S_n if f is a coboundary?
b) The sum S_n is also known as a Weyl sum. Assume f is continuous with $\int_0^1 f(x)\, dx = 0$ and that α is irrational. What does the Birkhoff ergodic theorem say about S_n/n?
c) Assume α is irrational. Are the random variables X_n independent? Are the random variables decorrelated? Can you use the strong law of large numbers to estimate S_n? Can you use the weak law of large numbers to estimate S_n?
d) Look up what happens if α has the Diophantine property $|\alpha - p/q| \leq 1/q^2$ for all rational numbers p/q. (An example is if α is the golden mean.) There is a result that assures that S_n stays bounded in this Diophantine case if f is continuous. Find that result and state it.

PROBABILITY THEORY

MATH 154

Homework 8

Transformation

Problem 8.1: a) Check that the automorphisms of a probability space form a group. There is a subset of ergodic automorphisms. Investigate whether (i) ergodic, (ii) weakly mixing, (iii) mixing automorphisms form a subgroup.
b) For every $T \in \mathrm{Aut}(\Omega, \mathcal{A}, \mathrm{P})$ we have a unitary transformation $U : \mathcal{L}^2 \to \mathcal{L}^2$ given by $Uf = f(T)$. Check the orthogonality condition $\langle Uf, Ug \rangle = \langle f, g \rangle$.
c) Classical mechanics is the theory of automorphisms of probability spaces, where the unitary evolution is given by a dynamics $Uf = f(T)$. Quantum mechanics allows for a larger automorphism group consisting of all unitary operator $Uf = e^{itA} f$ with a self-adjoint operator A on the Hilbert space $\mathcal{L}^2(\Omega)$. Assume our probability space is finite. What is its classical automorphism group? What is its quantum automorphism group?

Problem 8.2: Show that if a measure-preserving transformation T has the property that for any $A, B \in \mathcal{A}$ there is m such that $\mathrm{P}[A \cap T^{-n}(B)] = \mathrm{P}[A]\mathrm{P}[B]$ for all $n \geq m$, then $\mathcal{A}$ is a trivial algebra.

Ergodicity

Problem 8.3: Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space, and let $T : \Omega \to \Omega$ be a measure-preserving transformation. Verify that the following conditions are equivalent:
(i) T is ergodic
(ii) If $A \in \mathcal{A}$ and $\mathrm{P}[T^{-1}(A) \Delta A] = 0$, then $\mathrm{P}[A] = 0$ or $\mathrm{P}[A] = 1$.
(iii) If $A \in \mathcal{A}$ satisfies $\mathrm{P}[A] > 0$ then $\mathrm{P}[\bigcup_n T^{-n}(A)] = 1$.
(iv) If $A, B \in \mathcal{A}$ satisfy $\mathrm{P}[A] > 0, \mathrm{P}[B] > 0$ then there is n such that $\mathrm{P}[T^{-n}(A) \cap B] > 0$.
Instead of checking all 12 possible ordered pairs, use the Merry-Go-Round proof technique: $(i) \to (ii) \to (iii) \to (iv) \to (i)$.

Proof. $\boxed{(i) \to (ii)}$ $\mathrm{P}[T^{-1}(A)\Delta A] = 0$ means that $T(A) = A$ up to a measure zero. By definition A has measure 0 or 1. $\boxed{(ii) \to (iii)}$ The set $B = \bigcup_n T^{-n}(A)$ is invariant and so has measure 0 or 1. Since it contains A which has positive measure, it has measure 1.
$\boxed{(iii) \to (iv)}$ If there existed a set B which never can be reached, then B would be disjoint of $T^{-n}(A)$. But $\mathrm{P}[\bigcup_n T^{-n}(A)] = 1$.
$\boxed{(iv) \to (i)}$ Assume $T^{-1}(A) = A$ and A has measure different from one. Then take $B = A^c$. □

Weak mixing

Problem 8.4: a) In the proof showing that T is mixing implies T^2 is mixing, we use the following Lemma from calculus or real analysis: the following two things are equivalent:
(i) $c_n \geq 0$ is a bounded sequence with $\frac{1}{n}\sum_{k=1}^n |c_k| \to 0$.
(ii) There exists a set J of density 1 in $\mathbb{N}$ on which $\lim_{j\in J} |c_k| \to 0$.
b) Use a) to verify that if $c_n \geq 0$ is a bounded sequence $\frac{1}{n}\sum_{k=1}^n |c_k| \to 0$ is equivalent to $\frac{1}{n}\sum_{k-1}^n |c_k|^2 \to 0$.
c) Conclude that weakly mixing can be rephrased as the property $\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} |\mathrm{P}[A \cap T^{-k}(B)] - \mathrm{P}[A]\mathrm{P}[B]|^2 = 0$ for all $A, B \in \mathcal{A}$.

Mixing

Problem 8.5: a) Prove the following result of Rényi: A dynamical system T is mixing if and only if $\mu(A \cap T^{-n}A) \to \mu(A)^2$ for $n \to \infty$.
b) State and give a proof of the Riemann-Lebesgue lemma. Why does this lemma imply that T has only absolutely continuous spectrum, then T is mixing? (Use a).

PROBABILITY THEORY

MATH 154

Homework 9

Central Limit

Problem 9.1: a) Find again the characteristic function ϕ_X of a standard Cauchy distributed random variable X. (We have done it before. Maybe try to do it without a computer algebra system using residue calculus.)
b) Deduce that if you take two independent standard Cauchy distributed random variables X, Y, then $(X + Y)/2$ is again standard Cauchy distributed.

Problem 9.2: a) Verify that the differential entropy of the Cauchy distribution with density $1/(\pi(1 + x^2))$ is $\log(4\pi)$. Mathematica gives wrongly $\log\left((1 + \sqrt{\pi})^2 \pi\right)$!
b) As a flashback, recall how the expectation $\mathrm{E}[X]$ of a Cauchy distribution X is defined in a renormalized way by subtracting two infinite quantities.
c) Verify that the renormalized variance $\lim_{n\to\infty} \frac{1}{n} \int_{-n}^{n} x^2 f(x)\, dx$ exists for the Cauchy distribution. What is its value?

Problem 9.3: a) Compute the entropy of the standard distribution $N(0, 1)$. We have sketched it in class.
b) What is bigger, the entropy of the Cauchy distribution or the entropy of the standard normal distribution? c) Compute the entropy of the exponential distribution on $[0, \infty)$.

Problem 9.4: We work here with measures on $(\mathbb{R}, \mathcal{B})$.
a) Assume $d\mu(x) = f(x)\, dx$ and $d\nu(x) = g(x)\, dx$ are absolutely continuous probability measures. The convolution $f * g(x) = \int_{\mathbb{R}} f(y)g(x-y)\, dy$ defines a new measure $d\mu * d\nu = f * g\, dx$. Verify

$$\int f * gh(z)\, dz = \int \int h(x+y)f(y)\, dyg(z)\, dz\ .$$

b) Conclude that

$$d\mu * d\nu(A) = \int_{\mathbb{R}} \int_{\mathbb{R}} 1_A(x+y)\, d\mu(x)\, d\mu(y)$$

c) Verify that the transformation

$$T_\lambda(\mu)(A) = \int_{\mathbb{R}} \int_{\mathbb{R}} 1_A(\frac{x+y}{\sqrt{2}})\, d\mu(x)\, d\mu(y)$$

on the space of all Borel probability measures on $(\mathbb{R}, \mathcal{B})$ satisfying $\int x\, d\mu(x) = 0$ has a unique fixed point.

Problem 9.5: We have seen that the central limit theorem implies the de Moivre central limit theorem so that in principle we do not need to prove it again. Write down a proof of the de Moivre central limit theorem. You have the following options: a) using the Stirling approximation formula $n! \sim \sqrt{2\pi n}(n/e)^n$ for the factorial.
b) using characteristic functions, essentially repeating the general proof.