PROBABILISTIC IDEAS AND METHODS IN ANALYTIC NUMBER THEORY (PART I?)

PETE L. CLARK

1. Introduction

These are the extended lecture notes for a talk given in the UGA Number Theory Seminar on November 5, 2008. The aim is to explore the connections between analytic number theory and classical probability, and also to give a brief background to some of the most important probabilistic theorems.

In these notes, Section 2 consists of the number-theoretic discussion, whereas Section 3 provides the probabilistic background (with relatively few proofs). First we draw upon topology and group theory to give some examples of probability spaces. Perhaps the most interesting is the weighted Bernoulli space $\mathbb{B}_p$ which models flipping a – fixed but not necessarily fair – coin infinitely many times. In the sequel, we develop sufficient background to state Chebyshev's Inequality, the Law of Large Numbers, the Central Limit Theorem, and the Khinchin-Kolmogorov Law of the Iterated Logarithm.

The discussion in Section 2 makes reference to these theorems in Section 3 (in the talk itself, the aforementioned theorems were written up in advance on the side board). Therefore from a strictly logical point of view it would make more sense to read Section 3 first and Section 2 second. But almost surely it will be more interesting to do it the other way around!

2. Some Glimpses of Probability in Analytic Number Theory

A very basic problem in number theory is that of the distribution of arithmetic functions – which for our purposes here, will just be functions $f : \mathbb{Z}^+ \to \mathbb{R}$. Ideally, if f is an arithmetic function, we would like to determine the **asymptotic order** of f. That is we wish to find a simple arithmetic function g such that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$, a condition we abbreviate to $f \sim g$.

Remark: We assume here that f and g are nonzero for all sufficiently large n. In many classical situations this need not be the case, but one can at least identify an infinite subset of $\mathbb{Z}^+$ on which f is not zero, and then we want $f(n) \neq 0 \implies g(n) \neq 0$ and $\lim f/g = 1$, where the limit is taken along the nonzero values of f. Moreover, if $\lim_{n\to\infty} f(n) = 0$, then we say that $f \sim 0$, even though this is not a formal consequence of the definition.

Remark: $f \sim g$ is of course an equivalence relation on the set of all ultimately nonzero arithmetic functions. A better statement of the problem, perhaps, is: for

a given function f, find the simplest representative of the asymptotic equivalence class of f.

The example *par excellence* is the **prime number theorem**[1]: if we define $\pi(x)$ to be the number of prime numbers p which are at most x, then we have

$$\pi(x) \sim \frac{x}{\log x} \sim \mathrm{Li}(x) := \int_2^x \frac{dt}{\log t}.$$

Although this is a very deep theorem, in some sense the result is surprisingly simple: most of the other classical arithmetic functions simply do not have enough regularity in order to have a simple asymptotic expansion. For instance, consider any of the following functions:

- $d(n)$, the number of divisors of n.
- $\omega(n)$, the number of distinct prime factors of n.
- $r_2(n) = \#\{(x,y) \in \mathbb{Z}^2 \mid n = x^2 + y^2\}$.
- $\mu(n)$, the Möbius function.

Then:

- $\liminf_{n\to\infty} d(n) = 2$, whereas $\limsup_n d(n) = \infty$.
- $\liminf_{n\to\infty} \omega(n) = 1$,whereas $\limsup_n \omega(n) = \infty$.
- $\liminf_{n\to\infty} r_2(n) = 0$, whereas $\limsup_n r_2(n) = \infty$.
- $\liminf_n \mu(n) = -1$, whereas $\limsup_n \mu(n) = +1$.

Thus for each of these functions (and many others) there is no simple asymptotic. An old trick in analysis is that if the quantity you are looking at doesn't have the nice limiting behavior that you want, try averaging. For an arithmetic function f, we will consider instead

$$\overline{f} : n \mapsto \frac{1}{n}\left(f(1) + \ldots + f(n)\right).$$

If g is another function such that $\overline{f} \sim g$, we say that f is the average order of g. As above, if $\overline{f} \to 0$, we say that the average order of f is 0.

Remark: This usage is not universal. Tenenbaum's authoritative text *Introduction to Analytic and Probabilistic Number Theory* defines "f has average order g" to be $\overline{f} \sim \overline{g}$. To see that these are not the same, under the above definition, then $g(n) = \frac{1}{k+1}n^k$ is an average order for the function $f(n) = n^k$, whereas for Tenenbaum's definition clearly n^k is an average order for n^k. I believe the point is that Tenenbaum is primarily interested in functions which grow so slowly that this difference in definitions does not manifest itself.

Anyway, it is indeed more likely that there is a nice average order:

- The average order of $r_2(n)$ is π.

Quick comment: This follows almost immediately from the observation that $\sum_{i=1}^n r_2(i)$

[1]Due to J. Hadamard and C. de la Vallée Pousssin, independently, both in 1896.

counts the lattice points on or inside the closed disk with center $(0,0)$ and radius $\sqrt{n}$.

• The average order of $d(n)$ is $\log n$.

Quick comment: For any two arithmetic functions f and g we have the elementary identity

$$\sum_{i=1}^{N}(f*g)(i) = \sum_{i=1}^{N}\sum_{a \mid i} f(a)g(i/a) = \sum_{a=1}^{N} f(a)\sum_{j=1}^{[\frac{N}{a}]} g(j).$$

Applying this with $f = g = 1$ (constant function), one gets the above result easily.

• The average order of $\omega(n)$ is $\log\log n$.

Comment: Although this is easier than PNT, it was proved later, by Hardy and Ramanujan in the early 20th century. We will return to this later.

• The average order of $\mu(n)$ is 0.

Comment: Although anyone would guess this, in fact it is by far the deepest of the statements. It was proven by Landau in 1909 as a consequence of the prime number theorem. Conversely, it is (comparatively) easy to deduce the prime number theorem from this statement. That is, this is one of a family of statements that is known to be "elementarily equivalent" to PNT.

In this last case our intuition that the average value of the Möbius function should be zero seems motivated by some very rudimentary ideas about probability. Namely, the function takes on values $+1$ at squarefree numbers with an even number of prime divisors and -1 at squarefree numers with an odd number of prime divisors. Evidently the set of positive integers having a bounded number of prime divisors is very sparse (although infinite), so the main contribution comes from integers with many prime divisors. (Indeed, in some sense the previous result says that most integers of order about n have about $\log\log n$ prime divisors, although we have to be careful about this...) It is tempting to regard the parity of a large number as essentially a coin flip, and we do believe that if we flip a fair coin a large number of times, then the number of heads minus the number of tails divided by the total number of flips approaches zero. So we are alluding to a theorem of probability here!

Of course there are two issues here: first, to make precise this result in classical probability. Second, to understand the relationship between the classical result and this arithmetic analogue, because of course there are not literally any random variables here. This is characteristic of the sort of problems we wish to discuss.

In fact we would like to know more:

Maxim of Hard Analysis: For a given function f, instead of (just) finding an asymptotic function for f – i.e., a simple function g such that $f \sim g$ – try to find an explicit error bound: i.e.,

$$f = g + E,$$

where $E = o(g)$.

In the prime number theorem, the classical proof gives such an explicit error function E, but a quite complicated one. For a long time it has been observed empirically that (i) while $\pi(x)$ is asymptotic to both $\frac{x}{\log x}$ and $Li(x)$, the error $|\pi(x)-Li(x)|$ is much smaller than $|\pi(x) - \frac{x}{\log x}|$ and (ii) indeed it seems to be true that the error $|\pi(x) - Li(x)|$ is of order approximately $\sqrt{x}$. A more precise result is:

Theorem 1. *The following assertions (all of which are conjectured but unproven!) are equivalent:*
(i) For every $\epsilon > 0$, $|\pi(x) - Li(x)| = o(x^{1/2+\epsilon})$.
(ii) $|\pi(x) - Li(x)| = O(\sqrt{x}\log x)$.
(iii) The Riemann hypothesis holds.

We have an instance of "almost square root error": much more than just $f \sim g$ we have **almost** $f = g + O(\sqrt{g})$. This is a whole philosophy about such error bounds:

Maxim of Almost Squareroot Error: (i) A sum of n "random" real or complex numbers of absolute value 1 is with high probability not much larger than $\sqrt{n}$.
(ii) Conversely, if such a sum is of order smaller than $\sqrt{n}$, then there is sum deterministic phenomenon behind this extreme cancellation.

In fact our maxim gets tested in some of the other examples above: for instance, it still follows from elementary geometry that

$$r_2(n) = \pi n + O(\sqrt{n}).$$

Moreover Sierpinski was able to reduce the error to $O(n^{\frac{1}{3}})$ and there has been much further work. Hardy and Landau independently showed that the infimum of all δ such that the error is $O(n^\delta)$ is at least $\frac{1}{4}$. Recently Cappell and Shaneson released a preprint showing that the error is $O_\epsilon(n^{\frac{1}{4}+\epsilon})$ for all $\epsilon > 0$.

Why are we doing better than squareroot error? The argument for squareroot error works for the number of lattice points in dilates of any planar region bounded by a sufficiently nice (e.g. piecewise C^1) curve. However, if instead we took the square centered at the origin and with side length $2N$, we find that its area is $4N^2$ whereas the number of lattice points is $(2N+1)^2 = 4N^2 + 4N + 1$, so the error is $4N + 1$: i.e., no better than square root error. For a general body, there are theorems which assert that for a sufficiently general rotation the error terms become smaller. But a circle is already completely symmetric about rotations, so this extra symmetry leads to better error boundds.

I don't have time to talk about the Dirichlet divisor problem, which is somewhat lucky because the situation there is quite similar to $r_2(n)$ – "Gauss' circle problem" – and I'm not how to justify the better than squareroot error in this case.

Let's look back at the average value of $r_2(n)$: it is the irrational number π. This is a refinement of the joke that the typical American family has 2.5 children. Of course no number n has exactly π representations as sums of squares. In fact if $a^2 + b^2 = n$ then also $(\pm a)^2 + (\pm b)^2 = n$ and $(\pm b)^2 + (\pm a)^2 = n$, so – unless n is

twice a square, which is a neglible set – then whenever $r_2(n) > 0$ we in fact have $r_2(n) \geq 8$. Since the average value is π, it follows that the density of the set of numbers n which are sums of two squares is less than $\frac{1}{2}$, since $\frac{1}{2} \cdot 8 = 4 > \pi$.

But in fact the density of the set of n for which $r_2(n)$ is positive is 0. This is something that one can prove from the Two Squares Theorem together with the prieme number theorem in arithmetic progressions (PNTAP). Moreover, it suggests that we may be missing something by considering only the average order.

Definition: We say that an arithmetic function f has **normal order** g if for all $\epsilon > 0$, the set of n such that $|f(n)-g(n)| > \epsilon|g(n)|$ has density zero. One can chheck that the normal order of a function is well-determined up to asymptotic equivalence.

It follows from that $r_2(n)$, being a function which is zero on a set of density one, has normal order zero.

On the other hand, it turns out that the normal order of $\omega(n)$ *is* equal to its average order: $\log\log n$. What is different here? In order to answer the question, we need some way of measuring the deviation of a function from its mean value. There is a statistical quantity, the **variance**, which gives an upper bound for the deviation from the mean value, via **Chebyshev's Inequality**.

Hardy and Ramanujan were the first to show that the normal order of $\omega(n)$ is $\log\log n$. Their arguments were improved and made more explicitly probabilistic by Turan and later by Erdös-Kac. The final result is a triumph of both the hard-analytic spirit and probabilistic ideas:

Theorem 2. *(Erdös-Kac) For any real numbers $a < b$, we have*

$$\lim_{N\to\infty} \frac{1}{N} \mid \left\{ n \leq N \mid a \leq \frac{\omega(n) - \log\log n}{\sqrt{\log\log n}} \leq b \right\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2} dt.$$

This is a remarkable result: it transplants the most hallowed result of statistics – the **Central Limit Theorem** – into the area of analytic number theory!

Finally, let us come back to the Mertens function $M(n) = \sum_{i=1}^n \mu(n)$. It has long been known that:

Theorem 3. *The Riemann hypothesis is equivalent to: for all $\epsilon > 0$,*

$$M(n) = O_\epsilon(n^{\frac{1}{2}+\epsilon}).$$

An interesting wrinkle here is that in the 1890's Franz Mertens conjectured an even stronger statement:

Conjecture 4. *For all n $M(n) \leq n$.*

He tested this based on numerical data which is, by modern standards, amusingly meager (up to about $n = 10^4$). The point here is that why in the world would you expect such a good bound? In fact this was disproved by te Riele and Odlyzko in 1985:

Theorem 5. *There is an explicit constant $C_1 > 1$ such that $\limsup_n \frac{M(n)}{\sqrt{n}} > 1$.*

Stieltjes claimed in 1895 to prove something only slightly weaker: there exists an absolute constant C such that $M(n) \leq Cn$ for all n. His proof was never published, and again it seems very unlikely (but it is a testament to the difficulty of the problem that we cannot as yet disprove this!).

The question here is as follows: suppose for the sake of argument that $M(n)$ really were a sum of independent, identically distributed random variables. What would we expect its upper order to be? Is this "small enough" to prove the Riemann hypothesis? Is there any reason we should expect the error to be better than random?

The Central Limit Theorem tells us that if we were in a random i.i.d. situation, then indeed the error would be, with probability 1, $O(n^{\frac{1}{2}+\epsilon})$ for any $\epsilon > 0$ and that it will not be $O(\sqrt{n})$. How much larger than $\sqrt{n}$ would randomness predict? The answer is given by the sensational Khinchin-Kolmogorov law of the iterated logarithm.

The question here is not whether the Mertens function is really a sum of random variables: it obviously isn't. The question is how close to random it is, and whether we can prove or even believe it is close to random. I have seen a paper that suggests that the true upper order of $M(n)$ is $\sqrt{n} \log\log\log n$. It would be amazing to have some good reason to believe such a precise bound.

3. A Probability Cheat Sheet

3.1. Examples of probability spaces.

Let $X = (X, \mathcal{A}, P)$ be a probability space. For simplicity, let us assume that $\mathcal{A}$ contains each singleton subset $\{x\}$ of X. Thus, if X is at most countable, $\mathcal{A} = 2^X$, so the σ-algebra plays no essential role.

Example: Let X be a locally compact Hausdorff space, and let $\mathcal{B}$ be the Borel algebra, i.e., the σ-algebra generated by the open subsets. A **regular Borel measure** μ on X is a measure – with values in $[0, \infty]$ – on $(X, \mathcal{B})$ which satisfies the following additional properties:
(i) **inner regularity**: for every measurable set Y, $\mu(Y) = \sup \mu(K)$ as K ranges over all compact subsets of Y.
(ii) **outer regularity**: for every measurable set Y, $\mu(Y) = \inf \mu(U)$ as U ranges over all open sets containing Y.

A Borel measure is called **locally finite** if every point has a neighborhood of finite measure, and a **Radon measure** is a locally finite regular Borel measure. Note that local finiteness is considerably weaker than $\mu(X) = 1$, so that any regular Borel probability measure is necessarily a Radon measure.

Theorem 6. *Let G be a locally compact Hausdorff topological group.*
*a) There exists a Radon measure μ_L (resp. μ_R) on G such that for all $g \in G$ and all Borel subsets Y of G, $\mu(gY) = \mu(Y)$ (resp. $\mu(Yg) = \mu(Y)$). In other words, μ_L (resp. μ_R) is **left-translation invariant**) (resp. **right-translation invariant**).*

b) If ν is any left-translation (resp. right-translation invariant) Radon measure

on X, there exists a positive constant C such that $\nu = C\mu_L$ (resp. $\nu = C\mu_R$). Any such measure is called a ***left Haar measure*** *(resp.* ***right Haar measure****) on G.*

c) G is said to be ***unimodular*** *if there exists a single measure μ which is simultaneously a left Haar measure and a right Haar measure. For a unimodular group we simply write μ and refer to it as the* ***Haar measure****. By part b), this holds iff for any one left Haar measure μ_L and any one right Haar measure μ_R, there exists a constant C such that $\mu_L = C\mu_R$. Evidently abelian groups are unimodular. Moreover, so is any compact group.*

d) For a left Haar measure μ_L, $\mu_L(G) < \infty$ iff G is compact. Thus for a compact group there is a unique normalization of μ with $\mu(G) = 1$.

Application: There are no countably infinite compact Hausdorff topological groups.[2]

An **atom** in a probability space X is an element $x \in X$ such that $P(\{x\}) > 0$.[3]

Lemma 7. *The subset $\mathbb{A}$ of atoms in any probability space X is countable.*

We can therefore define the **atomic mass** of a probability space X as

$$P(\mathbb{A}) = \sum_{x \in \mathbb{A}} P(x).$$

A probability space is **discrete** if its atomic mass is 1. Such spaces are completely described by a function $P : X \to [0, 1]$ such that: $P(x) = 0$ off of a countable set $\{x_i\}$, and

$$\sum_{i \in I} P(x_i) = 1.$$

A space is **continuous**, or **nonatomic**, if it has no atoms.

Example: Let G be any connected Lie group, μ a Haar measure on G, $Y \subset G$ a regular-closed subset – i.e., Y is the closure of its interior – with $0 < \mu(Y) < \infty$. Then $P = \frac{1}{\mu(Y)} \cdot \mu$ is a continuous probability measure on Y. For instance, the normalized Lebesgue measure on an interval $[a, b]$ is an example of this.

Example: We can put a measure on $\mathbb{Z}^+$ by taking $\mu(n) = \frac{1}{2^n}$. The space X can be viewed as the space of possible outcomes of flipping a fair coin n times until we get heads.

Example (uniform measure on a finite space): As a simple example, let X be a finite set of cardinality n. Then we can define the probability of every singleton subset to be $\frac{1}{n}$, thus for any $Y \subset X$, $P(Y) = \frac{\#Y}{\#X}$. Thus we have made a connection between probability and combinatorics. It is very simple one, but that is one of its strengths!

Example (Finite Bernoulli space): Let $G = \{\pm 1\}^n$, so G is a finite abelian group of order $N = 2^n$. The Haar measure on G corresponds to flipping a fair coin n times

[2]This also follows from the Baire category theorem.

[3]In the future we shall write $P(x)$ instead of the correct but tedious $P(\{x\})$, trusting that no confusion will arise.

and recording the sequence of results.

Example (Biased finite Bernoulli space): Fix a number p, $0 \leq p \leq 1$. Let $X = \{\pm 1\}^n$ be the same set as before, but this time we wish to model a sequence of n coin flips of a biased coin, whose probability of heads (say 1 = "heads") is p. Thus we do not wish to assign all atoms the same weight. Indeed, if $x \in X$ is a sequence which has k heads, we want $P(x) = p^k(1-p)^(n-k)$. (Of course we appeal the binomial theorem to ensure that this discrete measure has total mass 1.) Note that if $p = 0$ or 1 the entire mass is concentrated in a single atom.

Example (**Bernoulli space**): Let $\mathbb{B} = \prod_{i=1}^{\infty} \{\pm 1\}$. Endowing each factor with the discrete topology, X carries a natural topology, which is (by Tychonoff) compact, Hausdorff and totally disconnected. Therefore its Haar measure P is a probability measure. Note that P is continuous "even though" the topology on X is totally disconnected.

Suppose wish to construct a **biased Bernoulli space** $\mathbb{B}_p$ with the same underlying set $\prod_{i=1}^{\infty} \{\pm 1\}$ as the Bernoulii space, but which is supposed to model the situation of an infinite sequence of flips of a biased coin which has probability p of heads. The cleanest way to view this is an an instance of a product of probability spaces, a construction which can, happily, be made in complete generality:

Let $(X_i, \mathcal{A}_i, P_i)_{i \in I}$ be a nonempty family of probability spaces. Let $X = \prod_i X_i$ be the Cartesian product. By a **cylindrical set** we mean a subset Y of X which is of the form $\prod_i Y_i$, with $Y_i \subset X_i$ for all i and $Y_i = X_i$ for all but finitely many i's. Let $\mathcal{A}$ be the σ-algebra on X generated by the cylindrical sets.

Theorem 8. *With notation as above, there exists a unique probability measure P on $(X, \mathcal{A})$ such that for every cylindrical set $Y = \prod_i Y_i$, $P(Y) = \prod_i P_i(Y_i)$.*

An optimal proof of this theorem has been given by S. Saeki in 1996 article in the American Mathematical Monthly (Vol. 103 (1996), p. 682-683).

Applying this with $I = \mathbb{Z}^+$ and each (X_i, P_i) equal to the two point space $\{pm1\}$ with $P(1) = p$, $P(-1) = 1-p$, we get a construction of the weighted Bernoulli space.

Example: If L/K is any Galois extension of fields, possibly infinite, then the Krull topology endows $G = \mathrm{Aut}(L/K)$ with the structure of a compact, totally disconnected topological group, which therefore has a Haar measure. It is only relatively recently that this measure has been given serious attention, but it is now a major part of the branch of mathematics known as **field arithmetic**.

3.2. **Random variables and distribution functions.**

A (real-valued) **random variable** on X is a measurable function $f : X \to \mathbb{R}$. Recall that this means that the preimage of every Borel subset of $\mathbb{R}$ is an element of our fixed σ-algebra $\mathcal{A}$.

Example: If $A \subset X$ is an event, then its characteristic function $\mathbf{1})_A$ is a random variable.

In particular, for each $a \in \mathbb{R}$ we have an event

$$f_a := \{x \ \in X \mid f(x) \leq a\}.$$

We abbreviate this to $[f \leq a]$.

By taking probabilities, we get a function

$$F : \mathbb{R} \to [0,1], \ t \in \mathbb{R} \mapsto P(f \leq t).$$

F is called the **distribution function** of F. It is easy to see that it has the following properties:

(DF1) F is nondecreasing: $t_1 \leq t_2 \implies F(t_1) \leq F(t_2)$.
(DF2) $F(-\infty) := \lim_{t \to -\infty} F(t) = 0$, $F(\infty) := \lim_{t \to \infty} F(t) = 1$.
(DF3) F is right-continuous at each point: for all t, $\lim_{h \to 0^+} F(t+h) = F(t)$.

Note that as a consequence of (DF1), F is in fact continuous except possibly for jump discontinuities at a countable set of points and is differentiable except on a set of measure zero. Conversely:

Theorem 9. *Let $F : \mathbb{R} \to \mathbb{R}$ be any function satisfying (DF1), (DF2) and (DF3) above. Then there is a probability space X and a random variable $f : X \to \mathbb{R}$ such that F is the distribution function of f.*

Proof: Indeed we can take $X = \mathbb{R}$, $\mathcal{A}$ to be the standard Borel σ-algebra, and let P be the Lebesgue-Stieltjes measure on $(X, \mathcal{A})$ determined by F: i.e.,

$$\int_{\mathbb{R}} g dP = \int_{\mathbb{R}} g dF.$$

Let $f : \mathbb{R} \to \mathbb{R}$ be the identity function. Then the P-probability that $f(x) \leq x$ is

$$\int_{-\infty}^{x} dF = F(x) - F(-\infty) = F(x).$$

A distribution function F is **absolutely continuous** if there exsits a Borel measurable $f : \mathbb{R} \to \mathbb{R}$ such that

$$F(x) = \int_{-\infty}^{x} f(t) dt.$$

The function F is called the **density function** for F. By the Radon-Nikodym theorem, the function f is uniquely determined as an element of $L^1(\mathbb{R})$, i.e., is unique up to almost everywhere equality. Conversely, if $g \in L^1(\mathbb{R})$ is a function which is non-negative and has $\int_{\mathbb{R}} g = 1$, then integrating against g yields an absolutely continuous distribution function G.

Example: (a) $X = \{\pm 1\}$ with equal probability. Let $f : X \to \mathbb{R}$ be the identity function, which is a random variable. Its distribution function is piecewise constant: it starts at 0, and then jumps to $\frac{1}{2}$ at $x = -1$, and then jumps to 1 at $x = 1$.
(b) Let $X = \{\pm 1\}^\infty$ be the unweighted Bernoulli space. Then $p_n(x) := x_n$ is a random variable. The distribution functions are the same as in part (a).
(c) Let $X = \{\pm 1\}^\infty$ be the weighted Bernoulli space with probability p. With p_n

defined as above, its distribution function is now the saltatory function with jumps $(-1, p)$, $(1, 1-p)$.

Example (**Gaussian distribution**): Let μ and $\sigma^2 \geq 0$ be two fixed real constants. Then we define

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

3.3. **Expected values.**

Expectation: Let $f : X \to \mathbb{R}$ be a random variable. Its **expectation**, or **expected value**, is simply

$$E(f) := \int_X f dP,$$

provided this integral exists and is finite (which it will if f is, e.g., bounded).

Example: The expected value of any of the functions p_n on the p-weighted Bernoulli space is

$$E(p_n) = 1 \cdot p + (-1) \cdot (1-p) = 2p - 1.$$

Lemma 10. *If $B \subset \mathbb{R}$ is a Borel set, and f is a random variable with distribution function F, then*

$$P(f \in B) = \int_B dF.$$

Theorem 11. *Let g be a Borel measurable function on $\mathbb{R}$, and $f : X \to \mathbb{R}$ a random variable on the probability space X. Then the expectation of the composite function $g \circ f$ exists iff*

$$\int_{\mathbb{R}} |g(x)| dF_x(x) < \infty,$$

in which case we have

$$E(g \circ f) = \int_{\mathbb{R}} g(x) dF_X(x).$$

Corollary 12. *If in Theorem 11 g is continuous, then $E(f \circ g)$ exists iff the improper Stieltjes integral*

$$\int_{\mathbb{R}} |g(x)| dF_X(x) < \infty,$$

in which case

$$E(g \circ f) = \int_{\mathbb{R}} g(x) dF_X(x).$$

Corollary 13. *If f is a random variable with distribution function F, then Ef exists iff the two improper Stieltjes integrals*

$$\int_0^\infty x dF(x), \ \int_{-\infty}^0 x dF(x)$$

are both finite, in which case

$$E(f) = \int_{\mathbb{R}} x dF(x).$$

So the expected value of f can be computed from its distribution function.

We say a random variable f is **discrete** if there exists an infinite sequence of real numbers $\{x_n\}$ such that with $p_n := P(f = x_n) = p_n$, then $\sum_n p_n = 1$. The distribution function of a discrete random variable is locally constant, with a jump of p_n at each point x_n.

Corollary 14. *If f is a discrete random variable with finite expectation, then*

$$E(f) = \sum_{n=1}^{\infty} x_n p_n.$$

Corollary 15. *If f is a random variable with an absolutely continuous distribution function F with density $dF = g$, then*

$$E(f) = \int_{\mathbb{R}} x g(x) dx.$$

Example: Take $g(x) = G(\mu, \sigma^2)$, the Gaussian. Then the random variable variable $f(x) = x$ has expected value μ.

The nth **moment** of a random variable f with distribution function F is

$$Ef^n = \int_{-\infty}^{\infty} x^n dF(x),$$

provided its exists. The nth **central moment** of f is

$$E(f - E(f))^n = \int_{-\infty}^{\infty} (x - Ef)^n dF(x).$$

In particular, the **variance** of f is

$$\mathrm{Var}(f) = E(f - E(f))^2.$$

3.4. **Chebyshev Inequalities.**

Theorem 16. *(Generalized Chebyshev Inequality) Let f be a random variable on X and $g : \mathbb{R} \to \mathbb{R}$ be a non-negative measurable function which is non-decreasing on the range of f. Then for any real number t we have*

$$P(f \geq t) \leq \frac{1}{g(t)} \int_X (g \circ f)(x) dP(x).$$

Proof: Let $A_t := \{x \in X \mid f(x) \geq t\}$ and $\mathbf{1}_{A_t}$ be the characteristic function of A_t. Then for all $t \in \mathbb{R},\ x \in X$ we have

$$0 \leq g(t)\mathbf{1}_{A_t} \leq g(f(x))\mathbf{1}_{A_t} \leq g(f(x)).$$

Integrating these inequalities we get

$$0 \leq g(t)P(f \geq t) \leq \int_X g(f(x)) dP(x).$$

If $P(f \geq t) = 0$, fantastic. Otherwise, dividing by it, we get the conclusion.

Applying this result with the function

$$g(t) = \left\{t^2,\ t \geq 0\right\},\ \{0,\ t \leq 0\}.$$

and $f \mapsto |f|$, we get:

$$P(|f| \geq t) \leq \frac{1}{t^2} Ef^2.$$

Replacing $|f|$ with $|f - Ef|$ and t with ϵ, we get

Corollary 17. *(Chebyshev's Inequality) Let $f : X \to \mathbb{R}$ be a random variable with finite expectation $\mu = Ef$. Then, for any $\epsilon > 0$,*

$$P(|f - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(f)}{\epsilon^2}.$$

Among other things, this justifies the name "variance": it provides an upper bonud for the deviation of a random variable from its expected value.

Taking $g(t) = \{t,\ t \geq 0\},\ \{0, t < 0\}$ and replacing f by $|f|$, t by ϵ, we get

Corollary 18. *(Markov's Inequality) For any $\epsilon > 0$, we have*

$$P(|f| \geq \epsilon) \leq \frac{1}{\epsilon} E|f|.$$

3.5. **Independence.**

Two events $A, B \in \mathcal{A}$ are said to be **independent** if

$$P(A \cap B) = P(A)P(B).$$

An equivalent but somehow more psychologically striking formulation is as follows: for any event B with $P(B) > 0$, we define $P(A|B)$, the conditional probability of A given B, as

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

Then independence of A and B is equivalent to either of the following statements:

$$P(A|B) = P(A),$$

$$P(B|A) = P(B).$$

Given a set $\{A_i\}_{i \in I}$ of events, we say they are (mutually) independent if for every finite subset $J \subset I$,

$$P(\cap_{j \in J} A_j) = \prod_{j \in J} P(A_j).$$

Moreover, if we have an indexed collection $\{B_\alpha\}$ of families of events, we say that the families of events are independent if for every choice of one event from each family, the set of events is independent.

Now suppose given a collection $\{f_i\}_{i \in I}$ of random variables on the space X. We say they are **independent** if for every finite subset $\{1, \ldots, n\} \subset I$ and real numbers $r_1, \ldots, r_n$, then

$$P(f_1 \leq r_1, \ldots, f_n \leq r_n) = \prod_{i=1}^{n} P(f_i \leq r_i).$$

An equivalent condition is that if $B_1, \ldots, B_n \subset \mathbb{R}$ are Borel sets, then

$$P(f_1 \in B_1, \ldots, f_n \in B_n) = \prod_{i=1}^{n} P(f_i \leq B_i).$$

Proposition 19. *Let $f_1, \ldots, f_n$ be a set of random variables on X. Let $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$ be Borel measurable functions. Then $g_1 \circ f_1, \ldots, g_n \circ f_n$ are also independent random variables.*

Theorem 20. *Let $f_1, \ldots, f_n$ be independent random variables, $g_1, \ldots, g_n$ are Borel measurable functions such that $E(g_i \circ f_i)$ exists for all i. Then $E(\prod_{i=1}^n (g_i \circ f_i))$ exists and*

$$E(\prod_{i=1}^{n} g_i \circ f_i) = \prod_{i=1}^{n} E(g_i \circ f_i).$$

Corollary 21. *If $f_1, \ldots, f_n$ are independent random variables on X with finite second moments, then*

$$\mathrm{Var}(\sum_{i=1}^{n} f_i) = \sum_{i=1}^{n} \mathrm{Var}(f_i).$$

Proof:

$$\mathrm{Var}(\sum_{i=1}^{n} f_i) = E(\sum_{i=1}^{n} f_i - E(\sum_{i=1}^{n} f_i)^2 = E(\sum_{i=1}^{n} f_i - Ef_i)^2$$

$$= \sum_{i=1}^{n} E(f_i - E(f_i))^2 + \sum_{j \neq k} E((f_i - E(f_i))(f_j - E(f_j)) = \sum_{i=1}^{n} \mathrm{Var}(f_i) + 0 = \sum_{i=1}^{n} \mathrm{Var}(f_i).$$

Proposition 22. *Let $\{X_i\}_{i \in I}$ be a nonempty family of probability spaces, and $\{f_i : X_i \to \mathbb{R}\}_{i \in I}$ be a family of random variables. Let $X = \prod_i X_i$ be the product probability space and denote by π_i the canonical projection $X \to X_i$. For all i, let $F_i = f_i \circ \pi_i : X \to \mathbb{R}$. Then the random variables F_i are independent.*

This is essentially a tautological consequence of the defining property of the product probability space.

As a special case, if we take all X_i's to be the same space S, we get $X = S^I$. Then, for any family of random variables $f_i : S \to \mathbb{R}$, the pullback family $F_i : S^I \to \mathbb{R}$ is independent. As an even more special case, we can take one fixed random variable $f : S \to \mathbb{R}$ and take each $f_i = f$.

This is the case for the random variables p_n on the weighted Bernoulli space $\mathbb{B}_p$. In particular, the sequence $\{p_n\}$ is an instance of a family of variables which is **independent and identically distributed**: henceforth **i.i.d.**.

4. Laws of Large Numbers

Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of random variables on a probability space X. Let f be a random variable. There are several different senses in which we may have $f_n \to f$. For simplicity, we consider only two.

We say that f_n converges to f **in probability** if for all $\epsilon > 0$,

$$\lim_{n \to \infty} P(|f_n - f| \geq \epsilon) = 0.$$

(This concept may be familiar to the student of measure theory, under the name "convergence in measure.") We abbreviate convergence in measure as

$$f_n \xrightarrow{P} f.$$

It can be shown that if $f_n \xrightarrow{P} f$ and also $f_n \xrightarrow{P} g$, then $f = g$ P-a.e.

Another form of convergence is almost everywhere pointwise convergence. This just means that there exists a null set[4] $Y \subset X$ such that for all $x \in X \setminus Y$, $f_n(x) \to f(x)$. One also describes this as **almost sure** convergence.

Basic results of measure theory apply to give that almost sure convergence implies convergence in probability, whereas convergence in probability implies that there is a subsequence which converges almost surely. In contrast, by virtue of its hypothesis of independence, the following belongs to the theory of probability:

Theorem 23. *(Kolmogorov) For an* independent *sequence of random variables $\{f_n\}$, converges in probability and almost everywhere convergence are independent.*

Theorem 24. *(Law of Large Numbers) Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of random variables, which are independent and identically distributed (i.i.d.). We assume moreover that Ef_1 exists, in which case all the Ef_i's exist and have a common value, say μ. Define*

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} f_i.$$

Then $\overline{X}_n \to \mu$.

Important remark: The point of the statement of the theorem is that we have defined precisely two versions of convergence: convergence in probability and almost everywhere convergence, and according to Kolmogorov's theorem, under the hypothesis of independence they are equivalent. To be more precise, the version of Theorem 24 in which the conclusion is convergence in probability is called the **weak law of large numbers**, whereas the version in which the conclusion is almost everywhere convergence is called the **strong law of large numbers**, and is due to Kolmogorov.

This applies in particular to the random variables π_n on the weighted Bernoulli space $\mathbb{B}_p$. Define $S_n = \sum_{i=1}^{n} \pi_n$: this is the running total of the number of heads minus the number of tails after flipping coins. Then the strong law of large numbers says that, with probability 1, we have

$$\overline{X}_n = \frac{S_n}{n} \sim (2p-1)n.$$

But now remember our First Maxim: when given an aymptotic formula, ask instead for a formula with an explicit error term. For instance, suppose we have a coin and are trying to test out the assumption that it is fair – i.e., that $p = \frac{1}{2}$. Clearly just a statement that $S_n \to 0$ is not helpful here.

The following theorem is the epicenter of the philosophy of square-root error:

Theorem 25. *(Central Limit Theorem) Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables, with common expectation μ and variance σ^2. Then, for any $z \in \mathbb{R}$, we*

[4]Let us say a null set of a measure space is a set which is a subset of a set of measure zero. If the measure is complete, then null sets are themselves measurable, but in general we work with Borel measures rather than their completions.

have

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx.$$

In particular the central limit theorem asserts that with probability one, for any $\epsilon > 0$ we have

$$S_n = \mu \cdot n + o(n^{\frac{1}{2}+\epsilon}),$$

and also that the probability that $|S_n - \mu \cdot n|$ is as large as $C\sqrt{n}$ is positive for any $C > 0$.

The following theorem – one of the real jewels in the theory – does even better.

Theorem 26. *(Khinchin-Kolmogorov's Law of the Iterated Logarithm) Let f_n be a sequence of i.i.d. random variables on a space X with common variance σ^2 and common expected value $\mu = 0$, and put $S_n = \sum_{i=1}^{n} f_i$. Then:*

$$P\left(\limsup_{n\to\infty} \frac{S_n}{\sigma\sqrt{2n\log\log n}} = 1\right) = 1.$$

Similarly

$$P\left(\liminf_{n\to\infty} \frac{S_n}{\sigma\sqrt{2n\log\log n}} = -1\right) = 1.$$