# A teaching note on Strong Law of Large Numbers

N Hemachandra
IE and OR Interdisciplinary Programme, IIT Bombay, Mumbai 400 076
nh@iitb.ac.in

September 2, 2003

The purpose of this lecture notes is two fold. Use the occasion of going through a proof of SLLN to demonstrate some very common tools and techniques that one encounters in this subject. We take some results from real analysis as 'given'–we indicate their need and usage rather going into their theory. However, we aim for some completeness as far as the probabilistic tools are concerned. So, we go at (what may seem to many) an excrutiatingly slow pace.

We give some motivation for SLLN in Section 1 for those who intend to use stochastic models and also set the stage for the detailed proof in the next section. There will be need based references to Section 3. The first part of Section 3 has details of some basic probabilistic results like, first Borel-Cantelli lemma, *etc.*, which are quite useful in a course like this. The second part summarizes some frequently used results from real analysis after giving some motivating (counter) examples. In Section 4 we first present a version of second Borel-Cantelli lemma and then use it to state a converse of SLLN: if arthematic average of pairwise independent and identically distributed random variables converge to a limit that is finite, then the limit has to be the mean of these random variables. We also describe a well known example that brings out the subtle difference between the weak law and the strong law. We close by indicating how one can obtain some assymptotic properties of sums of $i.i.d.$ random variables. This leads us to random walks which can also be viewed as an important class of discrete time Markov chains.

# 1 Introduction

Below we reproduce from Billingsley, Probability and Measure, $2^{nd}$ edition, '86, a proof of the Strong Law of Large Numbers, SLLN. This version of SLLN is given by Nasrollah Etemadi in 1981.

We deliberately give lots of additional details. Contrast this with the space Billingsley devotes to the same–hardly one and half a page. There are substantial reasons for this leisurely pace: for most registrants this is the first contact with serious probability theory. We also want to take advantage of this opportunity to demonstrate a couple of ideas that recur later where they will be used routinely without comment: probabilistic arguments like sample path proofs, *etc.*, use of some basic results and techniques from probability theory, impress upon the need for and the way interchange of some operations are handled, *etc.*

The Strong Law shares many ingredients that enables it to qualify as a fundamental result not only in mathematics, but also in science: It is simple to state but non-trivial to prove. It not only explains many natural phenomenon, but in some sense is routinely used in many applications. The application areas pervade physics, biology, economics, statistics and different engineering disciples like communication and computer networks, operations research, *etc.* It is almost always that the first results that one comes across in a model that involves random phenomenon has one thing or another thing to do with SLLN. Imagine asking a person with some basic knowledge of science or engg. to find the probability that a given coin shows head on a toss. The straight forward answer is to patiently toss it a large number of times and find the fraction that showed head. While the justification comes from SLLN, it also indicates the conditions that are needed in arriving at such conclusions.

Now a quick background. Jakob Bernoulli gave the weak law, (see Remark in Section 3), for *i.i.d.* Bernoulli random variables in 1713. Emile Borel gave the strong law (still for Bernoulli random variables) in 1909. One can view each sequence of heads and tails as a sequence of 1's and 0's and thus as a number in the interval $(0, 1]$ in base 2 representation. Then, the set of all coin tosses with a given pattern of heads and tails in the first $n$ tosses correspond to a certain interval in $(0, 1]$ of length $\frac{1}{2^n}$. For example, the set of sequences of tosses that start with a head in their first toss correspond to the interval $(\frac{1}{2}, 1]$. This helps one to reformulate SLLN as the study of properties of such sets of real numbers. The opening section of Billingsley's book has an easily readable development of these ideas culminating in a calculus based proof of weak law of binomial random variables corresponding to the experiment of tossing a fair coin. Kolmogorov showed the popular formulation for a sequence of *i.i.d.* random variables with finite mean in 1930s. This result was arrived at after carefully obtaining estimates of probabilities associated with sums of random variables (which are interesting by themselves) and thus this proof of SLLN is quite involved. Etemadi's proof, by contrast, is elementary if not simple; the tools from probability doesn't go beyond the first Borel-Cantelli lemma. An important feature is its technical acheivement: It relaxes the independence requirement to pairwise independence; see the discussion that follows the proof. The key reason probably it being a sample path argument.

The statement is simple: Suppose $\{X_i\}_{i\geq 1}$ is a sequence of independent and identical *(i.i.d.)* random variables such that $E[X_1] =: \mu \ < \infty$. Then,

$$\frac{\sum_1^n X_i(\omega)}{n} \to \mu \text{ with probability one (or almost surely).}$$

By this we mean that the event where this convergence is guaranteed, *i.e.*, the set of $\omega$ for which the convergence is taking place, is an event of probability one.

One interprets this as: The LHS is 'frequency' or 'time average'– it is the limit of the sum of values of $X_i$'s and divided by $n$, for each fixed $\omega$ coming from a sure event. The RHS is the expectation (averaging over $\omega$'s) for any fixed random variable. The SLLN basically says that these two different 'averages' are same. Evolution of a dynamic (physical) system in random environment corresponds to a sequence of $\{X_i(\omega)\}$ for a fixed $\omega$. This sequence of real numbers is called a sample path or a realization or trajectory. One observes the process over time and computes time average. The SLLN guarantees, that under reasonable conditions, this single time average will be the same as the probabilistic averaging over (typically uncountable) realizations. This is exactly what goes on in steady state simulations; to evaluate stationary expectation of certain random variables that could not be evaluated analytically, one simulates the phenomenon and collects appropriate time average statistics for a very large $n$. Done carefully, this is usually a good approximation. Dependence among random variables in dynamical systems could be an issue, where similar laws of Markov chains that follow from SLLN may help. This equality of two averages is in the spirit of ergodic theorems, and in fact, Kolmogorov's SLLN is indeed a special case of ergodic theorem.

Also, the SLLN is, in a sense, the starting point of statistical theory–estimation of parameters, *etc.* Specializing $\{X_i\}_{i\geq 1}$ as $\{I_{\{A\}}\}_{i\geq 1}$, we have that the long run fraction of times a event $A$ occurs in an infinite sequence of independent trails is the probability of occurance of that event. Note that in the widely prevalent axiomatic set up of probability theory, probabilities of events are assigned or assumed as basic building blocks, not derived from more basic entities. In this setting, the SLLN is conceptually important as it gives a satisfying interpretation that the value of the probability of an event *apriori* defined is nothing but 'relative frequency' of this event, which is more intuitive. Indeed, the fact that SLLN, which renders this interpretation, can be proved from axiomatic setting is often viewed not only as an achievement of the axiomatic set up, but also as one that justifies it.

Recall the definition of convergence of a sequence of real numbers, $\{x_n\}_{n\geq 1}$: We require that,

$$\{\ \forall\ \varepsilon > 0, \exists\ N\ \ni\ |x_n - x| < \epsilon\ \ \forall\ n\ \geq\ N\}$$

Here, $N$ typically depends on $\varepsilon$, apart from depending on the sequence itself. We can write this symbolically and hence concisely that convergence happens if,

$$\bigcap_{\varepsilon>0}\bigcup_{N}\bigcap_{n\geq N}\{|x_n - x| < \epsilon\}$$

is non-empty. Note here that the first intersection is an *uncountable* intersection. In the course of proof, we need to evaluate probabilities of these type of events but we do not even know if this uncountable intersection leads us to an event; in view of $\sigma-$additivity of probability, only countable set operations of events lead to an event. However, we can replace it by a sequence of $\varepsilon$s going to zero, say $1/n$, to have the same set, which is now an event. A standard way to prove almost sure convergence is to show that the set (in fact, the event) where the convergence fails is of zero probability. In this context, we use this fact: the event where $\{X_i\}_{i\geq 1}$ fail to converge to $X(\omega)$ is equivalent to the fact that there exists atleast one $\epsilon > 0$, on which $|X_i(\omega) - X(\omega)| \geq \epsilon$ i. o. (with i.o. standing for infinitely often). So, to show almost sure convergence, one shows that the event

$$\cup\{|X_i(\omega) - X(\omega)| \geq \epsilon \text{ i. o.}\}$$

has zero probability, where the union is usually over positive rational $\epsilon$s.

# 2 Proof

Enough to consider $\{X_i^+\}_{i\geq 1}$ where we define,

$$\begin{aligned} X_i^+ &:= \max(X_i, 0) \\ X_i^- &:= \max(-X_i, 0), \end{aligned}$$

the positive and negative parts of $X_i$ such that $X_i = X_i^+ - X_i^-$. This is so because, if we have SLLN for both the $\{X_n^+\}_{n\geq 1}$ and $\{X_n^-\}_{n\geq 1}$, then,

$$\frac{\sum^n X_i}{n} = \frac{\sum X_i^+ - \sum X_i^-}{n} \longrightarrow \mu^+ - \mu^- = \mu$$

where $\mu^+ = E[X_1^+]$ and $\mu^- = E[X_1^-]$. However, before we argue this way, we need to verify that $\{X_i^+\}_{i\geq 1}$ and $\{X_i^-\}_{i\geq 1}$ have inherited the properties assumed of $\{X_i\}_{i\geq 1}$ (*i.e.*, each is an *i.i.d.* sequence with finite mean). This is straight forward.

So, we *can* and *will* assume that the random variables $\{X_i\}_{i\geq 1}$ we consider below are *non-negative*.

Consider $Y_i := X_i I_{\{0\leq x_i\leq i\}}$ where, $I_{\{A\}}$ indicator function of event $A$. These are truncated versions of $\{X_n\}_{n\geq 1}$ (*i.e.*, $Y_1$ is same as $X_1$ for values below 1 and set to zero for values above 1, and so on). Look at the associated partial sums,

$$S_n^* := \sum_{i=1}^{n} Y_i \ .$$

For $\alpha > 1$, (temporarily) fixed, let $u_n$ be the integer part of $\alpha^n$, *i.e.*, $u_n := \lfloor \alpha^n \rfloor$.

**Claim** (to be proved below): For any $\epsilon > 0$,

$$\sum_{n\geq 1} P\{|\frac{S_{u_n}^* - E(S_{u_n}^*)}{u_n}| > \epsilon\} < \infty \ . \tag{1}$$

This claim helps us this way: If (1) is true, consider events corresponding to all positive rational $\epsilon$'s and use first *Borel-Cantelli* lemma (see Notes section for some details) to say

$$\frac{S_{u_n}^* - E[S_{u_n}^*]}{u_n} \longrightarrow 0 \text{ almost surely.} \tag{2}$$

Thus, SLLN for a truncated random variables along a sub sequence is on hand. This is because, for a given $\epsilon$, let

$$A_n = \{\frac{S_{u_n}^* - E[S_{u_n}^*]}{u_n} > \epsilon\}$$

so that $\sum P(A_n) < \infty$. The first Borel-Cantelli lemma then ensures us that $P(A_n \text{ i. o.}) = 0$. Let $N_\epsilon$ be the complement of this event so that $P(N_\epsilon) = 1$. On $N_\epsilon$, *i.e.*, for every $\omega \in N_\epsilon$ $A_n$'s occur finitely many times which means that for integers $u_n$ beyond an integer $u_n'$ (this $u_n'$ depends on $\omega$ among other things) we have $|\frac{S_{u_n}^* - E[S_{u_n}^*]}{u_n}| \leq \epsilon$. Now, consider the intersection of such events corresponding to positive rational $\epsilon$s which is also a sure event. On this sure event (2) is thus true.

Now for the proof of the above claim: The idea of proving this claim is to upper bound the LHS of (1) by using *Chebyshev's inequality*–see Notes. Thus, (1) is less than

$$\sum \frac{\text{var}(S_{u_n}^*)}{\epsilon^2 u_n^2} \ . \tag{3}$$

To simplify this bound further, we consider,

$$\begin{aligned}
\text{var}(S_n^*) &= \sum_1^n \text{var}(Y_i) \quad (\text{as } Y_i's \text{ are independent}) \\
&\leq \sum_1^n E(Y_i^2) \quad (\text{as } \text{var}(Z) = E[Z^2] - (E(Z))^2) \\
&= \sum_1^n E[X_i^2 I_{\{X_i \leq i\}}] \\
&= \sum_1^n E[X_1^2 I_{\{X_1 \leq i\}}] \quad (\text{we can use } X_1) \\
&\leq nE[X_1^2 I_{\{X_1 \leq n\}}] \quad (\text{using the fact that } X_1^2 I_{\{X_1 \leq n\}} \geq X_1^2 I_{\{X_1 \leq 1\}}).
\end{aligned}$$

Using this estimate in (3) we have that an upper bound for LHS of (1) is given by

$$\frac{1}{\epsilon^2} E[X_1^2 \sum \frac{1}{u_n} I_{\{X_1 \leq u_n\}}].$$

The order of summation and expectation can be interchanged, as the terms are positive quantities; see Notes again.

While we assumed finite mean for $X_1$, the above bound has $X_1^2$ inside expectation, apart from a summation; so, we need to show that bound is indeed finite. It turns out that a careful argument exploits these indicator functions and presence of decreasing $u_n$'s in the denominator. This is the next step, which is also an example of sample path argument. It looks like this is *the* crucial step in the proof. Sample path arguments, typically take this form: Fix $\omega$ in $X_i(\omega)$; then $X_i(\omega)$ is a real number! Before that, some careful estimates are needed.

For a real number $x$, which will be identified as $X_i(w)$ in the sample path argument below, let $N$ be the smallest $n$ such that $u_n \geq x$. Then, by definition of $u_n$, we have that $\alpha^N \geq x$. Also, we have $y \leq 2\lfloor y \rfloor$ if $y \geq 1$ so that,

$$\alpha^N \leq 2\lfloor \alpha^N \rfloor = 2u_n$$

which means that,

$$\begin{aligned}
x \leq \alpha^N \leq 2u_n \\
\implies \frac{1}{u_n} \leq 2\alpha^{-N} \text{ if } u_n \geq x \ .
\end{aligned}$$

Thus for each $u_n \geq x$ we can find $N$ such that $x \leq \alpha^N \leq 2u_n$. Using this we have,

$$\sum_{u_n \geq x} \frac{1}{u_n} \leq 2 \sum_{n \geq N} \alpha^{-n} \leq \frac{K}{\alpha^N} \leq \frac{K}{x} \quad \text{where} \quad K := \frac{2\alpha}{1-\alpha} \ .$$

Thus, $\sum \frac{1}{u_n} I_{\{X_1 \leq u_n\}} \leq K \frac{1}{X_1}$ for $X_1 > 0$. So, the sum at equation (1) is almost

$$\frac{K}{\epsilon^2} E[X_1] < \infty$$

and the claim is all about this only!

Recall, $\frac{E[S^*_{u_n}]}{u_n} = \frac{\sum^{u_n} E[Y_i]}{u_n}$. Now, limit of $\frac{\sum^{u_n} E[Y_i]}{u_n}$, being along the subsequence $\{u_n\}$, has the same limit as $\frac{\sum^{n} E[Y_i]}{n}$, if the later exists. But by Cesàro summation lemma, (see Notes), $\frac{\sum^{n} E[Y_i]}{n}$ has the same limit as $E[Y_k]$, if the later exists. Now, we have, $E[Y_k] = E[X_k I_{X_k \leq k}] = E[X_1 I_{X_1 \leq k}]$ (why ?). By the very definition of truncation, we have that these integrands (which are positive) monotonically increase to $X_1$. So, by Monotone Convergence Theorem, MCT, (see Notes), we can interchange operations, $E$ and lim, to have the limit as $E[X_1]$, *i.e.*, $\mu$. We thus have,

$$\frac{S^*_{u_n}}{u_n} \longrightarrow E[X_1] \text{ almost surely.}$$

Roughly we can say that we have SLLN *'along a given subsequence for truncated random variables'*. That leaves us with two tasks: drop truncation and subsequences restriction: We take up the first one now. For this, we need to figure out how each $X_i$ defers from its truncated avatar $Y_i$.

$$\begin{aligned}
\textstyle\sum_i P(X_i \neq Y_i) &= \textstyle\sum_i P(X_i > i) \\
&= \textstyle\sum_i \; P(X_1 > i) \quad \text{(why ?)} \\
&\leq \textstyle\int_0^\infty P(X_1 > t)\, \mathrm{d}t \\
&= E[X_1] \; < \; \infty \quad \text{(since } X_1 \text{ is positive–see notes).}
\end{aligned}$$

Another application of Borel Cantelli lemma now gives

$$\frac{S_{u_n}}{u_n} \longrightarrow E[X_1] \text{ with probability 1.}$$

We leave the couple of details needed here; you have to supply them–its a question of a sequence of events *not* happening infinitely many times is an event of probability one $\cdots$.

For the (almost) last step to clinch the proof, if $u_n \leq k \leq u_{n+1}$, then,

$$\begin{array}{rccccc}
 & S_{u_n} & \leq & S_k & \leq & S_{u_{n+1}} \quad \text{(why ?)} \\
\Rightarrow & \frac{S_{u_n}}{u_{n+1}} & \leq & \frac{S_k}{k} & \leq & \frac{S_{u_{n+1}}}{u_n} \\
\Rightarrow & \frac{u_n}{u_{n+1}} \cdot \frac{S_{u_n}}{u_n} & \leq & \frac{S_k}{k} & \leq & \frac{S_{u_{n+1}}}{u_{n+1}} \cdot \frac{u_{n+1}}{u_n}.
\end{array}$$

Using a standard $\epsilon - \delta$ argument, prove that $\frac{u_{n+1}}{u_n} \to \alpha$. This gives us,

$$\frac{1}{\alpha} E[X_1] \leq \frac{S_k}{k} \leq \alpha \, E[X_1] \text{ with probability 1.}$$

Now look at the event that is obtained by intersecting such probability one events in a countable fashion (considering rational $\alpha > 1$, for example), to claim that

$$\frac{S_k}{k} \to E[X_1] \quad \text{with probability one.}$$

**Remark 1** *Pairwise independence* is enough for this proof to go through. This is because we have used the fact that $\mathrm{var}(S^*_n) = \sum \mathrm{var}(Y_i)$ and for this to hold, it is enough that $X_i$'s are pairwise independent.

**Remark 2** One of the above steps involves the fact that if $X_n \to X$ *a.s.* and $P(Y_n \neq X_n \text{ i.o.}) = 0$, then $Y_n \to X$ *a.s.* You should show this.

**Remark 3** The *weak law* claims that for any $\epsilon > 0$, $P(|\frac{S_n}{n} - \mu| \geq \epsilon) \to 0$. The interpretation is that for any given $\epsilon$ (however small) the probability of time average $\frac{S_n}{n}$ deviating from $\mu$ by $\epsilon$ goes to zero, as $n$ increases. The sequence $\{\frac{S_n}{n}\}_{n\geq 1}$ is said to 'converge in probability' to $\mu$. As names indicate, the Strong law implies, but is not implied by, the weak law.

Let $A_n(\epsilon) := \{|\frac{S_n}{n} - \mu| > \epsilon\}$ and $B_m(\epsilon) := \cup_{n\geq m} A_n(\epsilon)$. Recall that we can write SLLN as $P(\cup_\epsilon \cap_m B_m(\epsilon)) = 0$ and hence as, $P(\cap_m B_m(\epsilon)) = 0 \; \forall \; \epsilon > 0$.

For a fixed $\epsilon > 0$, note that $B_m(\epsilon)$ is a decreasing sequence of events with limit $\cap_m B_m(\epsilon)$, call it $A(\epsilon)$. So, we then have that SLLN holds for a sequence of random variables, *iff* $P(A(\epsilon)) = 0 \; \forall \; \epsilon > 0$. By continuity of probability measure, (see Notes) this is true, *iff*, $P(B_m(\epsilon)) \to 0$ for all $\epsilon > 0$. This is an useful characterization of almost sure convergence of $\{\frac{S_n}{n}\}_{n\geq 1}$.

Since, $A_n(\epsilon) \subseteq B_n(\epsilon)$, we then have that $P(A_n(\epsilon)) \to 0$, *i.e.*, weak law follows from the Strong law. A careful examination of the events $B_m(\epsilon)$ and $A_n(\epsilon)$ will indicate why weak may not imply Strong Law; see Section 4 for an illuminating example.

**Remark 4** As of now, in above we can't drop "same law" assumption. But there are versions that do not demand 'same law' assumption. For example, one is due to Cantelli that states that Strong law is true for zero mean independent random variables with bounded fourth moments. The proof is not that difficult; one uses Markov's inequality (which generalizes Chebyshev's inequality).

# 3 Notes (please read!)

We collect here some details of arguments that are useful for us. In Part 1, we give some results from probability that we have used and they will also turn out to be useful later. In Part 2, we list some results from analysis. We try to motivate them occasionally by indicating what can go wrong, if imposed conditions are not met.

## 3.1 Part 1

**1** *Continuity of probability measure:* Consider a sequence of events, $A_1, A_2, \cdots$, where $A_1 \subseteq A_2 \subseteq A_3, \cdots$. The 'last' set is $\cup A_n$ and hence is an event. The continuity of probability (from above) lemma says that $P(\cup A_n) = \lim P(A_n)$. It is easy to see why this is true: Set $B_1 = A_1$ and for $n \geq 2$, define $B_n = A_n \setminus A_{n-1}$; $B_i$'s are 'rings'. Observe two things: $B_i$'s are disjoint and that $\cup_1^n B_i = \cup_1^n A_i$ for $n = 1, \cdots, \infty$. So, $P(\cup A_n) = P(\cup B_n) = \sum P(B_n) = \lim \sum^n P(B_i) = \lim[P(A_1) + \sum_2^n P(A_i) - P(A_{i-1})] = \lim P(A_n)$. The last but one step uses the easily provable fact that $P(A \setminus B) = P(A) - P(B)$ if $B \subseteq A$, while the last one follows from the telescoping of the summation.

Similar result is true for a decreasing sequences of events also: If $A_1, A_2, \cdots$, where $A_1 \supseteq A_2 \supseteq A_3, \cdots$ then, $P(\cap A_n) = \lim P(A_n)$. To prove this, work with complements of these events and use the fact that we are working with a probability (finite) measure. (Such a continuity from below may not hold for measures that are not finite; consider the (counter) example given below while discussing MCT.)

**2** *First Borel-Cantelli lemma:* If for a sequence of events, $A_1, A_2, \cdots$, we have that $\sum P(A_i) < \infty$, then, $P(A_i \text{ infinitely often}) = 0$.

First observe that

$$\{\omega : \omega \in A_i \text{ for infinitely many } i\} = \{\omega : \omega \in \cap_n \cup_{m \geq n} A_m\}.$$

So,

$$P(A_i \text{ infinitely often}) = P(\cap_n \cup_{m \geq n} A_m) \leq P(\cup_{m \geq n} A_m) \leq \sum_{m \geq n} P(A_m).$$

Since this is true for any $n$, take limit as $n \to \infty$ and the result follows from the fact that $\sum P(A_i) < \infty$.

**3** *Chebyshev's inequality:* Convince yourself that for a non-negative random variable $Z$ and a positive real number $a$, the following are true:

$$E[Z^2] = E[Z^2 I_{\{Z>a\}}] + E[Z^2 I_{\{Z \leq a\}}] \geq E[Z^2 I_{\{Z>a\}}] \geq E[a^2 I_{\{Z>a\}}] = a^2 P(Z > a).$$

Taking $|Z - E[Z]|$ as $Z$ in above we have Chebyshev's inequality:

$$\text{For } \epsilon > 0,\ P(|Z - E[Z]| > \epsilon) \leq \frac{\text{var}(Z)}{\epsilon^2}.$$

**4** We give a sample path proof of the fact that for a non-negative random variable $Z$, $E[Z] = \int_0^\infty P(Z > z)dz$. For a real number $z$ define an indicator function (the notation is slightly different now),

$$I(z, \omega) := 1 \text{ if } Z(\omega) > z \text{ and zero elsewhere.}$$

For a fixed $\omega$, draw a picture with $Z(\omega)$ on $x$-axis and $I(z, \omega)$ on $y$-axis; we have a rectangle of unit height. What is its length? Now, we can write $Z(\omega) = \int I(z, \omega)dz$ and hence, $E[Z] = E[\int I(z, \omega)dz]$. Since the integrands are positive, we can interchange the order of them–see below (at worst, both side can be $\infty$). Thus,

$$E[Z] = \int_0^\infty E[I(z, \omega)]dz = \int_0^\infty P(Z > z)dz.$$

## 3.2 Part 2

**1** As above, we will be occasionally required to 'change the order of integration' in iterated integrals. The question is will the answer be same. Consider a function $f(\cdot, \cdot)$ defined on $Z_+^2$, the set of positive integers in plane, as $f(n, n) = 1, f(n, n+1) = -1$ and $f(\cdot, \cdot) = 0$ elsewhere. Integrate first w.r.t. first co-ordinate and then w.r.t. second co-ordinate, (*i.e.*, along 'horizontal lines' first and then along 'vertical lines'). Perform the integrations in the other order; we have different answers. Incidentally, in the discrete setting here, integration is summation. This simple example shows that we *do* require conditions to change the order of integrations.

Observe that in above $f^+$ and $f^-$, the positive and negative parts of $f$, are such that their double integrals are infinite. One such condition is as above: it doesn't matter if the integrand is non-negative. For functions that are not so, the condition roughly states that the area in plane of the modulus function should be finite: $\int \int |f(x, y)|da < \infty$. In practice, find iterated integral of $|f|$ in a certain order, say w.r.t. $x$ first and then w.r.t. $y$; an upper bound that can be easily calculated is also enough. If this turns out to be finite, then all the

three integrals are of same value and are finite, by above. So, the order while integrating $f$ doesn't matter. This result goes by the name, Fubini-Tonelli theorem.

**2** Quite a few times we will be looking at the limit of integrals of functions. It may be advantageous to compute the limiting function first and then find its integral. Again the question is if both these two ways give the same answer. Note that integrating a function is analogous to finding the expectation of a random variable. Consider a sequence of right angled triangles on the interval $(0, 1)$. Let the $n^{th}$ one have a height of $2n$ at point $x = 0$ and the length of base be $\frac{1}{n}$. Let $f_n(x)$ 'describe' the $n^{th}$ triangle so that $\int f_n(x)dx = 1$, $\forall\, n$ so that $\lim \int f_n(x)dx = 1$. On the other hand, convince yourself that $\lim f_n(x) = 0$ so that $\int \lim f_n(x)dx = 0$. So, these interchange of operations may not preserve the final value and we need conditions that do preserve it. We list here the ones we frequently use:

1) Suppose, for non-negative functions $\{f_n(\cdot)\}$, we have that $f_n(x) \uparrow f(x)$ for each point $x$. Then, we can interchange the order of these integration and taking limit operations. This result is known as Monotone Convergence Theorem. We need increasing functions: consider $f_n(x) = I_{\{x>n\}}$. Each associated integral has a infinite value while the limiting function is zero!

2) Suppose for a sequence of functions $\{f_n(\cdot)\}$ we have that $f_n(\cdot) \to f(\cdot)$, almost everywhere (*i.e.*, $f_n$s converge pointwise to $f$ almost everywhere) along with $|f_n(x)| \leq g(x)$ for an integrable function $g(x)$, *i.e.*, $\int g(x)dx < \infty$. Then, we can interchange these operations. This is Dominated Convergence Theorem. As a consequence, for a almost everywhere convergent $f_n(\cdot)$'s that are uniformly bounded on a finite interval and zero elsewhere, the order of integration and taking limit doesn't matter. This particular result is known as Bounded Convergence Theorem.

Convince yourself that the triangles that we considered above, doesn't fit into these two settings.

**3** *Cesàro summation lemma:* Suppose $a_n \to a$. Then, $\frac{\sum_1^n a_i}{n} \to a$. So, if $a_n$'s converge, so do their 'time averages'. As in the proof of SLLN, this is how it is typically used. The intuition here is that, since $a_n$s converge to $a$, for all large enough $n$, $a_n \sim a$. So, beyond a large integer $N$, $\sum_N^{N+n} \sim na$ and hence the numerator in time average is $na$ plus sum of initial $a_i$s. Dividing by $n$ we will have the result; the detailed proof shows that in the limit, these approximations are correct.

A couple of remarks: 1) There is a corresponding continuous time version of this result– summation is replaced by integral. 2) A general version has an increasing sequence of integers, say $b_n$'s, in place of $n$'s with their differences suitably 'weighing' $a_n$'s. 3) The converse is not true: just consider the sequence with 0 and 1 in alternating places.

# 4 Further material

**1** *A Second Borel-Cantelli lemma:* The first lemma can be stated as $P(A_n \text{ i.o.}) > 0 \Rightarrow \sum P(A_i) = \infty$. Let us see if the converse is true. We use a popular probability triple: Take $\Omega$ as $(0, 1)$. The probability function, $P$, assigns probability of $b - a$ to interval $(a, b)$ for all reals $a$ and $b$. In fact, such intervals and many, many, many more sets are now in the $\sigma$-field (such sets are called Lebesgue sets, we will not go into the technicalities now). Consider events $A_n := I_{\{0,\frac{1}{n}\}}, n = 1, 2, \cdots$. We have $\sum P(A_n) = \infty$, but $P(A_n \text{ i.o.}) = 0$. So, some conditions are definitely needed for the converse to hold.

A version of second lemma: Let $A_1, A_2, \cdots$ be a sequence of pairwise independent events

such that $\sum P(A_i)$ diverges. Then, $P(A_i \text{ i. o.}) = 1$. To prove this, introduce a random variable $N_n := I_{\{A_1\}} + \cdots + I_{\{A_n\}}$ that counts the number of occurrences of events among $A_1, \cdots, A_n$. This is useful as $\{A_n \text{ i.o. }\} = \{\omega : \sup_n N_n(\omega) = \infty\}$. So, we are through if we show that $P(\sup_n N_n < \infty) = 0$. Let, $p_i = P(A_i),\ i = 1, 2, \cdots$, so that,

$$E[N_n] =: m_n = \sum_1^n p_i, \text{ and } Var(N_n) = \sum_1^n Var(I_{\{A_i\}}) = \sum_1^n p_k(1-p_k) \leq \sum_1^n p_k = m_n \ .$$

Also, for a real number $r < m_n$, $P(N_n \leq r) \leq P(|N_n - m_n| \geq m_n - r) \leq \frac{Var(N_n)}{(m_n - r)^2}$. For a given $r$, we then have that $\lim_n P(N_n \leq r) \to 0$, using the above estimate of $Var(N_n)$ and the fact that $m_n \to \infty$ (and hence eventually, $m_n > r$).

From $P(\sup_n N_n \leq r) \leq P(N_n \leq r)$, we have that $P(\sup_n N_n \leq r) = 0$ for any $r$. Now take a union over $r = 1, 2, \cdots$, to have the desired result: $P(\sup_n N_n < \infty) = 0$.

So, with the extra condition of pairwise independence, we have that the probability is not only positive, but also achieves the maximum possible value of 1. There are further extensions to this result. The pair of Borel-Cantelli lemmas are the first of a series of results that go by the name 'zero-one laws'.

**2** *A converse to SLLN:* Suppose for an pairwise independent sequence of identically distributed random variables, $\{X_n\}_{n\geq 1}$, we have that $\frac{S_n}{n} \to \mu$ almost surely. Then, $\mu = E[X_1]$.

In words, if time average of such a sequence 'stabilizes', then we can identify that limit as the mean of these random variables. In this setting, *a priori* we do not know if these random variables have finite mean.

The proof is easy: We have $\frac{X_n}{n} = \frac{S_n - S_{n-1}}{n} \to 0$ *a.s.* We next claim that $\sum P(|\frac{X_n}{n}| > 1) < \infty$; if not, we can apply the above Borel-Cantelli second lemma to say that $P(|\frac{X_n}{n}| > 1 \text{ i.o.}) = 1$ contradicting that $\frac{X_n}{n} \to 0$ *a.s.* Then,

$$E[|X_1|] \leq 1 + \sum P(|X_1| > n) = 1 + \sum P(|X_n| > n) < \infty.$$

Now that $X_1$ is integrable, we are in the setting of SLLN; so, apply it to $\{X_n\}_{n\geq 1}$ to say that their time average goes to the mean of $X_1$ and thus identify the (necessarily unique) limit as $\mu$.

**3** *Weak law doesn't imply Strong Law:* We recall the definitions. If $X_n \to X$ *a.s.*, then there is an event of probability one, say $A$, on which this convergence takes place. In words, for any $\omega \in A$ and any given $\epsilon > 0$, we can find an integer $N$, such that $|X_n(\omega) - X(\omega)| < \epsilon$ for *all* integers $n$ greater than $N$. Usually, $N$ depends on the pair $(\omega, \epsilon)$.

A sequence $\{X_n\}_{n\geq 1}$ is said to converge to $X$ in probability, if for every $\epsilon > 0$, $P(|X_n - X| \geq \epsilon) \to 0$. So, as $n$ increases, the probability of 'undesirable events' where $X_n$'s fail to converge reduces and the limit of them is zero. Only these probabilities decrease-there is no guarantee that for a given sample path $\omega$, $X_n(\omega) \to X(\omega)$.

The celebrated example below brings this out clearly. Sample points $\omega$ play a cat-and-mouse game with us: for a fixed $\omega$, $|X_n(\omega) - X(\omega)| < \epsilon$ may be true for a certain $n$ and may be also for a *finite* number of subsequent integers. Then, we will have an integer $n'$ propping up where $|X_{n'}(\omega) - X(\omega)| \geq \epsilon$. We may even have a finite string of integers where the inequality is of greater type, only to find another finite set of integers starting from some $n''$ such that $|X_{n''}(\omega) - X(\omega)| < \epsilon$. The sign of the inequality does flip-flops *without* end, with the result that on this sample path there is no convergence. This happens for almost all sample paths and hence there is no convergence with probability one. Meanwhile, for a

given $\omega$, whenever the inequality is of greater than sign (as when $|X_{n'}(\omega) - X(\omega)| \geq \epsilon$) it contributes to these (undesirable) probabilities. Since these probabilities also decrease we have convergence in probability.

In short, the sequence of events of $\{\omega : |X_n(\omega) - X(\omega)| \geq \epsilon\}_{n\geq 1}$ keep changing with $n$ and in the process they sweep the entire $\Omega$ leading to non-convergence; however, they also become 'progressively smaller' giving rise to convergence in probability. Recall the role of events $B_m(\epsilon)$ and $A_n(\epsilon)$ in Remark 2 after the proof.

The probability space $\Omega$ is the interval $(0,\ 1)$ with probability of an interval being its length. We define a sequence of indicator functions $\{X_n\}_{n\geq 1}$ that capture intervals of length $\frac{1}{n}$ in such a way that when these intervals are put side-to-side, they cover the interval $(0, 1)$ infinite number of times (since $\sum \frac{1}{n}$ diverges). These infinite number of covers can be obtained by wrapping $(0, 1)$ around infinitely often: those intervals that 'overshoot' point 1 are 'split' at point 1 and 'continued' from point 0, as in $X_4, \cdots$.

$X_1 = I_{\{(0,\ 1)\}},\ X_2 = I_{\{(0,\ \frac{1}{2})\}},\ X_3 = I_{\{(\frac{1}{2},\ \frac{5}{6})\}},\ X_4 = I_{\{(\frac{5}{6},\ 1)\ \cup\ (0,\ \frac{3}{12})\}},\ X_5 = I_{\{(\frac{1}{12},\ \frac{1}{12}+\frac{1}{5}\}},\ \cdots,$

$X_{11} = I_{\{(a,\ 1)\ \cup\ (0,\ b)\}},$ where $a = \frac{1}{12} + \sum_5^{10} \frac{1}{n}$ and $b = \frac{1}{12} + \sum_5^{11} \frac{1}{n} - 1,$

$\cdots\cdots$.

However, the situation is not so bleak: In countable spaces, both these types of convergences are same. Also, whenever a sequence converges in probability, we can find a subsequence that converges to the same limit in the almost sure fashion.

**4** *A word about sums of random variables:* Assymptotic properties of partial sums are important. A basic result, when $\{X_n\}_{n\geq 1}$ is an *i.i.d.* sequence with $E[X_1] =: \mu \in R$, is this:

$$\begin{aligned}
&(a)\ \mu > 0 \ \Leftrightarrow\ S_n \to \infty\ a.s.,\\
&(b)\ \mu < 0 \ \Leftrightarrow\ S_n \to -\infty\ a.s.,\\
&(c)\ \mu = 0 \ \Leftrightarrow\ \liminf S_n = -\infty\ and\ \limsup S_n = \infty\ a.s.
\end{aligned}$$

Also, $\mu = 0$ is equivalent to: for every $\epsilon > 0$, $P(|S_n| \leq \epsilon \text{ i.o.}) = 1$. Of course, for any given sequence $\{X_n\}_{n\geq 1}$ only one of these situations prevail.

A important particular case is when $P(X_1 = 1) = \frac{1}{2} = P(X_1 = -1)$; we then have the well known random walk on intergers $\cdots, -1, 0, 1, \cdots$. A gambling interpretation is that you bet and win a rupee if a fair coin shows $H$ on each toss but loose a rupee if a $T$ shows up; $S_n$ is your cummulative gain/loss by $n^{th}$ round. Then $(c)$ holds and one can interpret this: You gain infinitely large amounts infinitely many times, loose infinitely large amounts infinitely many times but you also play with no money on hand infinitely many times!

Polya obtained in early 20s this type of results for random walks in $d$-dimensions. The above more general result and other charecterizations (which we didn't list) seem to be largely due to Chung and Fuchs.

We will later understand this particular result viewing $\{S_n\}_{n\geq 0}$ as a Markov chain.