Kerala School of Mathematics Kozhikode
Summer School on Probability May 17 - June 04, 2010.

Measure Theoretic Probability

Lecture 1: Countable Sample Spaces.

Example 1: Roll a fair die once. Then the set of outcomes consist of $\{1, 2, 3, 4, 5, 6\}$. For example the outcome 4 means that face 4 turns up. Here all the six outcomes have the same probability, namely, 1/6. What are the chances that an odd number turns up. Thus we are asking for the probability of the event $\{1, 3, 5\}$ and we say it is $3/6 = 1/2$.

Example 2: Consider $S = \{1, 2, 3, 4, 5\}$ and choose a permutation of S at random. Here the set of possible outcomes consists of all possible permutations of S. It has 120 elements and each outcome has probability 1/120. What are the chances that the number of cycles in the selected permutation is 5? Recall that, for instance, if σ is the permutation given by $\sigma(1) = 2; \sigma(2) = 3; \sigma(3) = 1; \sigma(4) = 5; \sigma(5) = 4$, then $X(\sigma) = 2$. The event whose probability we are asking for, has just one element, namely, the identity permutation. Hence the probability is 1/120. More generally, letting X denote the number of cycles in the selected permutation, we can see that $P(X = 1) = 24/120; P(X = 2) = 50/120; P(X = 3) = 35/120$ and $P(X = 4) = 10/120$.

Example 3: Let $S = \{1, 2, 3, 4\}$. Select a graph (undirected) on the vertex set S at random. There are $\binom{4}{2} = 6$ possible edges and hence there are $2^6 = 64$ possible graphs. Each has a probability of 1/64 of being selected. What are the chances that the number of connected components in the selected graph is 4? there is only one such graph and hence the probability of the event is 1/64. More generally, if we denote by X the number of connected components in the selected graph, then $P(X = 1) = 38/64; P(X = 2) = 19/64$ and $P(X = 3) = 6/64$. Thus it is most likely that the selected graph is connected. You will learn more about random graphs in the lectures of MGN.

Example 4: Let us classify bolts manufactured by a machine into two classes — good and defective. Pick a bolt at random. What are the chances that it is defective. It is tempting to say 1/2, because there are two possibilities and the event under consideration has only one outcome. However, you

should agree that this is not intuitively acceptable. Afterall, only a small percentage of those manufactured will be defective. Thus depending on the prior knowledge regarding the machine we should associate the probabilities, for instance, the outcome 'defective' could be assigned 1/100 and the outcome 'good' can be assigned 99/100.

Thus the salient features of modelling such experiments is the following. We have a countable set Ω, called the *sample space*. The points of this set, denoted by ω, are the possible *outcomes* of the experiment being analyzed. For each $\omega \in \Omega$, we have a number $p(\omega)$. These numbers are non-negative and they all add to one. The idea is that $p(\omega)$ represents the *probability* (chance) of observing the outcome ω when you perform the experiment. The pair (Ω, p) is called a *probability space*. An *event* is any subset (possibly the empty set) of the sample space. We define the probability of an event A, to be the sum of probabilities of all the outcomes which are in the event, that is, $P(A) = \sum_{\omega \in A} p(\omega)$. Sum over empty set is taken as zero. One interesting case is when all the outcomes are equally likely, that is for any two outcomes, ω_1 and ω_2 we have $p(\omega_1) = p(\omega_2)$. In such a case the set Ω is necessarily finite and $p(\omega) = 1/|\Omega|$ for each ω. Of course, then for any event A we have $P(A) = |A|/|\Omega|$ and thus calculating probabilities reduces to counting problems (this does not mean they are simple problems).

A random variable is any measurement made on the outcome. More precisely, a *random variable* is a real valued function X defined on Ω. Associated with a random variable is its *probability mass function*. It is the function f defined on R as follows: $f(x) = P\{\omega : X(\omega) = x\}$. Clearly, Ω being countable, this function f takes the value zero for all but countably many numbers x. Moreover, all the non-zero values add to one. Associated with a random variable are certain important statistics that give an idea of the random variable, namely, *moments*. The expected value or the mean $E(X)$ of the random variable is the number $\sum xf(x)$. This is defined only when $\sum |x|f(x) < \infty$ and is not defined otherwise. Keep in mind that f is non-zero for only countably many vaues of x and hence this is a countable sum. The provision for absolute convergence is for the following reason. Generally, when you have a series of numbers, $\sum a_n$, then you have a first term and a second term etc. In the present situation, we do not have any pre-determined ordering on the values of the random variable, or to put it differently, there is no pre-determined ordering of the countably many x with $f(x) \neq 0$. The sum, $\sum xf(x)$ depends on how you order these values. Naturally, our definition should not depend on how one orders the numbers and

calculates the sum. By theorems in analysis, if you want a countable sum of numbers to converge and not depend on the order in which they are added, then the series must necessarily be absolutely convergent.

For any integer $k \geq 1$, the k-th moment of X is defined as $E(X^k) = \sum x^k f(x)$. This is defined only when the sum $\sum |x|^k f(x) < \infty$ and not defined otherwise. From now on when we talk about any moment, we assume that it is defined. The number $E(X^2) - (EX)^2$ is called the *variance* of the random variable X, usually denoted by σ^2.

It is not difficult to see that $E(X^k) = \sum_{\omega \in \Omega} X^k(\omega) p(\omega)$. This equation is important for the following reasons. After having defined the expected value of a random variable, we have no business to define the expected value of, say, for instance X^3. This is because $Z = X^3$ is a random variable in its own right and thus you are supposed to find out the p.m.f. g of Z and calculate $\sum zg(z)$. The above equation can be used to show that even if you calculated $\sum x^3 f(x)$, you get the same answer. Also this equation can be used to show that expected value is linear, that is, $E(aX + bY) = aE(X) + bE(Y)$ for any two random variables X, Y (defined on a probability space) and for any two numbers a, b. This is not a trivial fact, simply because knowing the p.m.f s of X and Y will not reveal the p.m.f. of their sum.

One of the important concepts in probability theory is that of conditional probability. For two events A and B in an experiment, we define the *conditional probability of A given B* by $P(A \cap B)/P(B)$ and denoted by $P(A|B)$. This definition is arrived at by the following reasoning. We have performed the experiment (Ω, p) and we have the partial information that an outcome from B has occured; how should we assign probabilities to outcomes now? Suppose $p(\omega|B)$ denotes our modified probability for the outcome ω. Clearly, if $\omega \notin B$, then we should have $p(\omega|B) = 0$. If $\omega_1, \omega_2 \in B$, and if ω_1 is twice more likely than ω_2, it should still remain so. More generally $p(\omega_i|B)/p(\omega_j|B) = p(\omega_i)/p(\omega_j)$ for any two outcomes in B. Of course, we must have $\sum_{\omega \in \Omega} p(\omega|B) = 1$. These prescriptions uniquely determine $p(\omega|B) = p(\omega)/P(B)$ and hence for any event A, $P(A|B) = \sum_{\omega \in A} p(\omega|B)$. This last expression is same as $P(A \cap B)/P(B)$.

It is natural now to define two events A and B to be *independent* if the conditional probability of A given B is same as its unconditional probability, that is, $P(A|B) = P(A)$ which is also the same as saying that $P(A \cap B) =$

$P(A)P(B)$. It is pleasing that this last equation is symmetric in A and B. More generally, n events $A_1, A_2, \cdots, A_n$ are independent if for $1 \leq i_1 < i_2 < \cdots < i_k \leq n$; we have

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

This is equivalent to saying the following. If for each $1 \leq i \leq n$ you take B_i to be either A_i or A_i^c; then

$$P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2) \cdots P(B_n).$$

The basic idea is this. If you tell me that A_1, A_2, A_3 occurred and A_4, A_5 did not occur and ask about the conditional chances of $A_6 \cap A_7^c$, then this conditional probability is same as the unconditional probability of $A_6 \cap A_7^c$.

Concepts for events easily translate to random variables. For example if $X_1, X_2, \cdots X_n$ are random variables defined on a probability space, we say that they are independent if

$$P(X_1 = a_1, X_2 = a_2, \cdots, X_n = a_n) = P(X_1 = a_1)P(X_2 = a_2) \cdots P(X_n = a_n).$$

Note that if $A_1, A_2, \cdots, A_n$ are events in a probability space, then their independence as stated earlier is same as the independence of the random varibles $\{I_{A_i} : 1 \leq i \leq n\}$ as stated above.

Example 5: Let $0 < \alpha < 1$. Toss a coin whose chance of heads is α. $\Omega = \{H, T\}$, $p(H) = \alpha$ and $p(T) = 1 - \alpha$. Let X be the number of heads obtained. Then $X(H) = 1$ and $X(T) = 0$.

Example 6: Toss the above coin 4 times independently. Now Ω consists of sequences of length 4 consisting of H and T. There are a total of 16 outcomes. Since they said that the tosses are independent, our probability should satisfy, for example,

$$p(HTHT) = p(H)p(T)p(H)p(T) = \alpha^2(1-\alpha)^2.$$

You see that here the assignment of probabilities are made in such a way that the hypothesis of independence is satisfied. Again, X, the number of heads obtained is a random variable. For example $X(HTHT) = 2$. $P(X = 2) = \binom{4}{2}\alpha^2(1-\alpha)^2$. Here $E(X) = 4\alpha$ and variance of X is $4\alpha(1-\alpha)$.

In fact if you toss the above coin independently n times and count the number X of heads obtained, then its probability mass function is called Binomial distribution and is given by $f(x) = \binom{n}{x}\alpha^x(1-\alpha)^{n-x}$ for $x = 0, 1, \cdots, n$

and $f(x) = 0$ for other values.

Example 7: Toss the above coin till you get a Head and then stop. Here $\Omega = \{H, TH, TTH, TTTH, \cdots\}$ and $p(T^nH) = (1-p)^n p$ for $n \geq 0$. If X the number of Tails before the Head is obtained then it is a random variable; $X(T^nH) = n$ and p.m.f. of X is given by $f(n) = (1-p)^n p$ for $n = 0, 1, 2\cdots$ and zero otherwise. Chances of having an even number of tails before the Head is $\sum_{n=0,2,4,\cdots} f(n)$.

Of course, in the above explanation, you will notice that we have omitted one outcome which is theoretically possible, namely, T^∞, consisting of all T. However this outcome has probability zero and hence we did not include. Indeed, $p(T^\infty) \leq p(T^n) \leq (1-p)^n$ for every n and hence must be zero.

The following inequality, known as *Markov's* inequality is simple and fundamental. For a non-negative random variable X and $a > 0$, we have $P(X \geq a) \leq E(X)/a$. Indeed

$$E(X) = \sum xf(x) \geq \sum_{x \geq a} xf(x) \geq a \sum_{x \geq a} f(x) = aP(X \geq a).$$

You should appreciate not only its simplicity, but also its generality. There was no assumption about the distribution. Of course if the mean is infinite, this says nothing. This, in turn, gives us, what is called, *Tchebycheff's* inequality. For any random variable X with mean μ and variance σ^2, we have $P(|X - \mu| \geq a) \leq \sigma^2/a^2$. Indeed the event $(|X - \mu| \geq a)$ is same as $(|X - \mu|^2 \geq a^2)$ and now apply Markov to the random variable $(X - \mu)^2$. The same method leads to *Chernoff's* bound. Let X be a random variable with $M(\theta) = E(e^{\theta X})$ finite for all $\theta > 0$. Then for any $\theta > 0$ and any $a > 0$, we have $P(X \geq a) \leq M(\theta)e^{-\theta a}$. Indeed

$$P(X \geq a) = P(e^{\theta X} \geq e^{\theta a}) \leq E(e^{\theta X})\, e^{-\theta a} = M(\theta)e^{-\theta a}.$$

Of course, we are assuming that the function $M(\theta)$ is well defined, that is, the required expectations are finite. At the same time you should appreciate the appearance of an exponential term on the right side. If $M(\theta)$ is nice and does not compensate this exponential term we conclude that the tail probability decays at an exponential rate. Taking advantage of the $\theta > 0$ at our disposal, we can refine this inequality to get *Cramer-Chernoff* inequality. Denote $\Lambda(\theta) = \log M(\theta)$, then we have $P(X \geq a) \leq \exp\{-[a\theta - \Lambda(\theta)]\}$. Since this is true for every $\theta > 0$, denoting $\Lambda^*(a) = \sup_{\theta > 0}[a\theta - \Lambda(\theta)]$, we conclude that for any number $a > 0$, $P(X \geq a) \leq \Lambda^*(a)$. You will learn more about

this in the lectures of SR.

Above inequalities are crucial and play an important role in the lectures of MGN, BR and RS. Here is one use. First note that for independent random random variables, variance adds up. That is, $var(\sum X_i) = \sum var(X_i)$ if the variables are independent. Suppose you toss a coin n times, whose chance of heads in a single toss is p. If X_i denotes one or zero according as you get Heads or Tails in the i-th toss, then these random variables are independent, each having mean p and variance $p(1-p)$. Thus their average S_n/n has mean p and variance $p(1-p)/n$. Tchebycheff's inequality tells us that for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{1}{n}\frac{p(1-p)}{\epsilon^2},$$

a quantity which converges to zero as $n \to \infty$. Do not worry about ϵ, this is held fixed. Clearly, S_n/n is nothing but the observed proportion of heads. Thus what the above result says is the following: as the number of tosses increases, the chances of the observed proportion of heads differing from p by an amount larger than ϵ goes to zero. Since $\epsilon > 0$ is arbitrary, our mathematical theory justifies the intuitive feeling that in a large number of tosses, the proportion of heads is p. This result is known as the *Weak law of large numbers*.

There are several standard practice problems like the matching problem, urn probems, balls and boxes problems. There are several standard p.m.f. like the geometric, negative binomial, Poisson and so on. But we shall not proceed in that direction. The probability spaces that we discussed are also called discrete probability spaces, just to convey that the underlying set of outcomes is at most countably infinite. What if you want to model experiments where the number of outcomes are not countable, for example, the life time of the electric bulb in this room. Of course, you can discretize and say that I count the number of minutes it has worked, but we shall not look at such approximations, we need to provide a good model for the life time. It coud be any positive number. This is what we shall try to understand next with the help of a simple experiment.

Lecture 2: Probability spaces.

Let us consider the experiment of choosing a number at random from the interval $(0, 1]$. Here, every number in this interval is a possible outcome so that the sample space $\Omega = (0, 1]$, a set which is uncountable. If we follow the same simplistic philosophy of assigning numbers $p(\omega)$ to each outcome, we run into difficulties. Fix any number x in $(0, 1]$. Since we are selecting a point at random the chances of it falling in $(0, 1/2]$ and the chances of it falling in $(1/2, 1]$ must be the same and hence each must be equal to $1/2$. Whether x is in $(0, 1/2]$ or $(1/2, 1]$ you conclude that $p(x) \leq 1/2$. You can continue this argument by splitting this interval again to see $p(x) \leq 1/4$ and so on, yielding $p(x) = 0$. If you are not happy about my bringing in the concept of length, you can argue as follows. Since the selection is 'at random' all outcomes must be equally likely. But in such a case, as mentioned earlier, the sample space Ω must be finite. Thus if you try to associate probabilities to outcomes, you observe that each outcome must have probability zero. This, by itself is not bad, but then if you try to define the probability of an event as the sum of probabilities of outcomes in that event, we run into trouble.

We can take clue from the concept of length. You see that points have length zero, but intervals have non-zero length. Perhaps we should not try to associate probabilities to outcomes, we should try to associate, at one stroke, a probability to each event. But then what are events? Imitation being the first choice (also due to lack of any guiding principle at this stage), let us say that an event is simply a subset of $\Omega = (0, 1]$. Thus for every subset A of the unit interval we want to associate a number $P(A)$ which represents the probability of the event – that is $P(A)$ is the chance that the selected number belongs to the set A.

There are some natural requirements on P. For example $P(\emptyset) = 0$ and $P(\Omega) = 1$. In other words, chances of nothing happening is zero and chances of something happening is one. Further if we have disjoint events $A_n, n \geq 1$ with $\cup A_n = A$ then we should have $P(A) = \sum P(A_n)$. This condition is called countable additivity. These conditions are dictated by our intuitive notion of what a probability ought to be. Now there is another condition that reflects the fact that we want to model the specific experiment of picking a point at random. As noted earlier, this amounts to demanding that for any dyadic interval $(k/2^n, (k+1)/2^n]$ the value of the probability should be $1/2^n$. If we can make an assignment of a value $P(A)$ for each event A so that these conditions are satisfied we will have a model for the experiment and can feel happy. Unfortunately our task is a near impossibility due to a theorem proved by the Polish mathematician S.M.Ulam.

Theorem 1: Impossibility of model
(Assume two set theoretical hypotheses : Axiom of choice and Continuum Hypothesis) There is no function P defined for all subsets A of the unit interval such that it is non negative, countably additive, and for dyadic intervals $P(k/2^n, (k+1)/2^n] = 1/2^n$.

Do not bother about the set theoretical hypotheses mentioned above. This is not too important for our present discussion. The moral is that our attempts fail. At this juncture you will surely be tempted to ask: please name one set A for which we fail to associate probability. No, I can not do that. The condition of countable additivity is a collective condition, not an individual condition applying to each event A. This causes problems. To illustrate with a simple analogy, suppose there are 16 students in my class and I have 31 apples. I say: it is not possible for me to give two apples to each of the students. You will surely agree. But can you name one single student to whom I fail to give? No, simply because any person you name I can start giving two apples to him.

Where did we go wrong? Which of our demands is so tough that it can not be met? Let us carefully analyze our demands. Perhaps we should not accept the set theoretic hypotheses mentioned above. May be we should look at other suitable hypotheses that allow us to build a model. Such attempts are made, but are too complicated and, as expected, connected with logic and foundational aspects of set theory. We shall not accept this alternative.

Perhaps we should not demand countable additivity. But then we will run into deep troubles. We will have problems with what we are used to so far. Remember how we calculated the probability of even number of tails before the Head in Example 7? We added up the probabilities of even integers. In spite of this, it is perfectly legitimate for you to say that we abandon countable additivity. You may well argue that the example mentioned above is a special case — after all, demanding finite additivity does not exclude the possibility of countable additivity. You are very right. There are strong arguments in favour of abandoning countable additivity and instead, demand only finite additivity. Attempts at developing models with only finite additivity were performed. The upshot seems to be that such a theory is messy, mathematically difficult, a not so fertile ground for analysis and shows cracks when trying to model independent repetitions of random experiments. We shall not accept this alternative.

The only demand on which we have to compromise is ‘every subset is an

event'. Is it really necessary? What exactly is the purpose of our building a mathematical model? We want to answer simple questions concerning the experiment. So if we can assign probabilities to all the subsets that we are likely to come across in practice then it should serve our purpose. Thus we abandon the assumption that every subset be an event.

Then the question is : which subsets should be events? Clearly each interval should be an event. You might feel that since we know what should be the probability of an interval of the type $(a, b]$ — it should be $(b - a)$ — why not we associate probability to such intervals and their finite disjoint unons? Unfortunately, this collection of sets does not include certain important subsets. For example one may ask what are the chances that the selected number is a rational number? an algebraic number? What are the chances that in the decimal expansion of the selected nuumber each of the digits $0, 1, 2, \cdots, 9$ appears with equal frequency 1/10?. This is a simple question which should be given a meaning and answered. Here frequency of the digit 3 is 1/10 means the following: proportion of the digit 3 in the first n decimal digits must have a limit as $n \to \infty$ and this limit equals 1/10. With a little bit of work, one can show that this set of numbers is neither an interval nor even a countable union of intervals.

Let us see the natural conditions that the collection of events must satisfy. If A is an event then its complement A^c should also be an event. (If you can answer 'what are the chances of something happening', then we should be able to answer 'what are the chances of that something not happening'). If we have countably many events A_n then their union $A = \cup A_n$ should also be an event. (if some one asks you 'what are the chances of this happening' 'what are the chances of that happening' etc then be prepared, he may ask you tomorrow 'what are the chances of one of the things that I mentioned yesterday happening'). Let $\mathcal{B}$ be the smallest class of subsets of Ω satisfying these conditions. We agree to treat sets that belong to this class as events. Sets which are not in this class are not events.

With this change of attitude, to build a mathematical model means assigning probability $P(A)$ to all events A – that is, to sets belonging to the class $\mathcal{B}$. This assignment should satisfy the same demands as earlier. Lo and Behold, we can build a model!

Theorem 2 : Possibility of a model

There is a function P defined for all sets in $\mathcal{B}$ such that it is non negative, countably additive, and for dyadic intervals $P(k/2^n, (k+1)/2^n] = 1/2^n$.

Success at last, but the price we have to pay is that not every subset of sample space is regarded as an event. This is not as bad as it looks like. How many subsets of $(0,1]$ are there? Toooooo many. How many sets are we likely to come across in practice? Verrrrrry few. Let me assure you that every subset of $(0,1]$ that *you can think of* is in this collection $\mathcal{B}$. To produce a set that is not in this collection, one has to work hard; though in fact there are many many sets that do not belong to this collection $\mathcal{B}$ than those that belong!

The question that now arises is : who decides what subsets should be called events? But if you think about it, it should be clear to you that it depends on the experiment under consideration and the questions we are likely to be interested. So there is no decision that can be made *a priori*. We can develop general theory and depending on specific experiment we can specialize the theory.

This brings us to the basic frame work of measure theoretic probability. We should have a triple $(\Omega, \mathcal{A}, P)$. Here Ω is a non-empty set — sample space, describing the possible outcomes of the experiment under consideration. $\mathcal{A}$ is a family of subsets of Ω – sets in this class are to be called events. The family $\mathcal{A}$ should satisfy the conditions: (i) $\emptyset \in \mathcal{A}$, $\Omega \in \mathcal{A}$; (ii) if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$; if we have a sequence of sets A_n each belonging to $\mathcal{A}$, then $\cup A_n \in \mathcal{A}$. P is defined on $\mathcal{A}$ and should satisfy the following: (i) $P(\emptyset) = 0$, $P(\Omega) = 1$ and (ii) if (A_n) is a sequence of disjoint events, then $P(\cup A_n) = \sum P(A_n)$. This setup is general enough to allow discussion of all chance experiments.

Such a triple is called a *probability space*. Such a family of subsets $\mathcal{A}$ is called a *σ-field* of subsets of Ω. Such a P is called *a probability* on $\mathcal{A}$. Before we start any analysis, we should explain how to get these objects.

Getting sample space is easily settled. It depends on the experiment you are interested in. If it is tossing a coin twenty times, then Ω consists of all sequences of length twenty consisting of the symbols H and T. If it is selecting a point at random from the unit interval $(0,1]$, then Ω consists of the set of all points in this interval. If you are trying to model the motion of a particle which moves continuously on the real line, then Ω consists of the set of all continuous functions $x : [0,\infty) \to R$. The understanding is that, $x(t)$ denotes the position of the particle at time t and the function x itself describes one possible motion of the particle, one outcome. Theoretically, Ω

can be any non-empty set.

The next order of business is to see examples of σ-fields and to explain how to construct σ-fields. For any Ω the family of all its subsets is a σ-field on Ω. At the other extreme is the family consisting of just the two sets: $\emptyset$ and Ω. This family is also a σ-field. You can think of several in between. For example take any set $A \subset \Omega$ and consider the family that consists of the four sets: $\emptyset$, A, A^c and Ω. More generally take any finite or countable partition of Ω and consider the family which consists of precisely those sets that can be obtained as union of some sets in this partition.

Here is another one. Consider those sets which are countable *or* those sets whose complements are countable. This family is also a σ-field. This is called the countable-cocountable σ-field. Cocountable set means complement of a countable set and hence this nomenclature. Of course, if Ω itself is countable, then this family indeed consists of all subsets of Ω. However, if Ω is uncountable then this family does not consist of all subsets. Here is more satisfying answer to constructing σ-fields.

Theorem 3A: Getting σ-fields
Let Ω be any non-empty set and let $\mathcal{C}$ be any family of subsets of Ω. Then there is a σ-field $\mathcal{A}$ of subsets of Ω which includes all sets from the given family $\mathcal{C}$ and which is smallest such. That is, if $\mathcal{B}$ is any σ-field of subsets of Ω that includes the given family $\mathcal{C}$ then $\mathcal{A} \subset \mathcal{B}$.

Clearly the σ-field $\mathcal{A}$ of the theorem above is unique. It is called the σ-field generated by the family $\mathcal{C}$ denoted by $\sigma(\mathcal{C})$. Proof of the theorem is simple. Consider all possible σ-fields on the set Ω that include the given class of sets $\mathcal{C}$. There is at least one such — namely the class of all subsets of Ω. Collect in a bag just those sets A which belong to all these σ-fields. This bag of sets is a σ-field, it includes the given class $\mathcal{C}$. This is what we are looking for.

For example, you can take R for Ω and the collection of all intervals for the family $\mathcal{C}$. The resulting σ-field of subsets of R is called the *Borel σ-field* on R and sets in this family are called *Borel sets*. It is easy to show that every singleton set, and hence every countable set is a Borel set. Thus theoretically σ-field could be any $\sigma(\mathcal{C})$.

Finally, how do we construct probabilities. Even if we take Ω to be the real line and $\mathcal{A}$ to be its Borel σ-field, how are we ever going to construct

probabilities on this σ-field. Remember, to construct a probability we have to assign a number $P(A)$ to each Borel set A. We must also show that it is countably additive. Definitely a tall order – especially under our present circumstances where we did not even see all the Borel sets. First we make a useful definition.

Say that a family $\mathcal{F}$ of subsets of Ω is a *field* if the following properties hold: (i) $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$; (ii) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$; and (iii) if A and B are in $\mathcal{F}$ then their union $A \cup B$ is also in $\mathcal{F}$. Thus a field differs from a σ-field only in one respect – namely, a field is closed under finite unions whereas a σ-field is closed under countable unions. This actually makes a major difference.

Suppose that $\mathcal{F}$ is a field of subsets of Ω and P is defined on $\mathcal{F}$. Say that P is *a probability on* $\mathcal{F}$ if the following hold: (i) $P(\emptyset) = 0$ and $P(\Omega) = 1$; (ii) Whenever we have a sequence A_n of pairwise disjoint sets in $\mathcal{F}$ whose union A is also in $\mathcal{F}$, then $P(A) = \sum P(A_n)$. Note that as opposed to the definition of countable additivity on a σ-field, here, in case of a field, we have to say that if each A_n is in the field *and also that their union is in the field* then countable additivity holds. This is because, it is quite possible that the union may not be in the field and so its probability may not be defined.

Theorem 4: Caratheodory Extension Theorem
Let $\mathcal{F}$ be a field of subsets of Ω and $\mathcal{A} = \sigma(\mathcal{F})$. Suppose that P is a probability on $\mathcal{F}$. Then P has a unique extension as a probability to $\mathcal{A}$. This means that we have a probability P_1 on $\mathcal{A}$ such that for sets $F \in \mathcal{F}$, $P_1(F) = P(F)$. Further there is only one such extension.

Usually the extension is also denoted by P. Yes, mathematically speaking the domain of definition of P is sets in $\mathcal{F}$ whereas the domain of definition of P_1 is sets in $\mathcal{A}$. So there is a clear distinction. But you agree that notation should help us to understand things rather than being a burden. As long as you and I understand what we mean (and can successfully communicate with others), it is fine.

What is the importance of this theorem for us? It tremendously reduces our job of constructing probabilities. If you want to construct a probability on a σ-field $\mathcal{A}$ then you need not really catch hold of each set in $\mathcal{A}$ and prescribe $P(A)$. You choose *any* convenient field $\mathcal{F}$ so that $\sigma(\mathcal{F}) = \mathcal{A}$ and just construct probability on the field you have chosen. The theorem above assures that there is then a unique probability on $\mathcal{A}$ which agrees with what

you prescribed on the field. Thus you did indeed exhibit a probability on the σ-field itself. Thanks to Caratheodory, our job of constructing probabilities on a σ-field is very much simplified.

The question still remains. Can you execute this idea in any concrete case? Yes. This is what we do now. What is more concrete than R and its Borel σ-field? We now recall the definition of distribution function. To start with, suppose that indeed P is a probability on $(R, \mathcal{B})$. Let us define a function on the real line as follows: $F(x) = P(-\infty, x]$. Then the function F satisfies the following conditions. (i) F is monotone nondecreasing, that is, if $x \leq y$ then $F(x) \leq F(y)$. (ii) F is right continuous, that is, if $x_n \downarrow x$ then $F(x_n) \downarrow F(x)$. (iii) $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to +\infty$. Such a function F as above is called a a distribution function, *d.f.* for short. The above properties follow from the following.

Theorem 5 : Continuity of Probability
Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let (A_n) be a sequence of events (that is, sets in $\mathcal{A}$). If $A_n \uparrow A$ then $P(A_n) \uparrow P(A)$. If $A_n \downarrow A$ then $P(A_n) \downarrow P(A)$.

Proof is not difficult. Define, $B_1 = A_1$, $B_2 = A_2 - A_1$, $\cdots$, $B_k = A_k - A_{k-1}$, $\cdots$; These are disjoint with union A and hence $P(A) = \sum_n P(B_k)$. By definition of sum of series the right side is nothing but $\lim_n \sum_1^n P(B_k)$. Recognize this finite sum as $P(A_n)$ to complete the proof.

To prove the other part you can use $A_n^c \uparrow A^c$. Or can argue as follows. Suffices to show $P(A_n - A) \downarrow 0$. Observe that $A_1 - A = (A_1 - A_2) \cup (A_2 - A_3) \cup \cdots$ so that $P(A_1 - A) = \sum P(A_n - A_{n+1})$ is finite. But then the tail sum of the series on the right, which is precisely $P(A_n - A)$, goes to zero.

Theorem 6: Basic Existence Theorem of Probability Theory
(i) If P is a probability on the Borel σ-field $\mathcal{B}$ of R and we define $F(x) = P(-\infty, x]$, then F is a *d.f.*
(ii) Conversely, given a *d.f.* F on R, there is a unique probability P on $(R, \mathcal{B})$ such that for every $x \in R$, $F(x) = P(-\infty, x]$.

This is a very satisfying theorem for several reasons. Firstly, it produces examples of probabilites on the Borel σ-field of the real line. You only need to take a *d.f.* and you have a probability associated with it on the Borel σ-field. Secondly, it produces *all* probabilities on the real line. Caratheodory already reduced our job of constructing probabilities on a σ-field. He assured

us that if we can do it on a field that generates the σ-field then it admits a unique extension to the entire σ-field. The theorem above reduces the job still further. Do not bother even about a field. Just give me a *d.f.*, that is enough. Then we can show a probability on the Borel σ-field. It is easy to construct *d.f.*s on R.

In the next lecture we shall see some indication of proofs of these theorems and then proceed to the further development of the theory.

Lecture 3: Probabilities on R and Random variables.

Continuing the discussion of the last lecture, first we outline the proof of Theorem 6.

This is an application of the Caratheodory extension theorem. Read the statement of the theorem to convince yourself that you have no alternative but to define $P(-\infty, b] = F(b)$. Since you are searching for a probability, you are forced to put $P(a, b] = F(b) - F(a)$ for any $-\infty \leq a \leq b \leq \infty$, with the understanding $F(+\infty) = 1$ and $F(-\infty) = 0$. So consider the collection $\mathcal{S}$ of sets of the form $(a, b]$ with $-\infty \leq a \leq b \leq +\infty$. The convention is that $(a, \infty]$ is interpreted as (a, ∞). Thus for example $(-\infty, +\infty] = R$. We can write the proof without such conventions, but you have to break your arguments case-wise. Define P for sets in the class as suggested above.

Step 1: We claim that P so defined is countably additive on the class of sets $\mathcal{S}$ — that is, if $I \in \mathcal{S}$ is a countable disjoint union of sets $I_n \in \mathcal{S}$ then $P(I) = \sum P(I_n)$. This will be done in a series of three observations.

1^o. If we have a finite number of disjoint intervals $(a_n, b_n]$ for $1 \leq n \leq k$, all contained in $(a, b]$ then $F(b) - F(a) \geq \sum[F(b_n) - F(a_n)]$. Indeed, since we have only finitely many intervals, rearrange them if necessary and assume $a_1 < a_2 < a_3 < \cdots a_k$. Keep in mind the telescopic sum

$$F(b) - F(a) = [F(b) - F(b_k)] + [F(b_k) - F(a_k)] + [F(a_k) - F(b_{k-1})] + \cdots$$

and ignore the unnecessary terms – all terms being positive – you get the desired inequality.

2^o. If we have intervals (a_n, b_n) for $1 \leq n \leq k$, whose union includes $[a, b]$ then $F(b) - F(a) \leq \sum[F(b_n) - F(a_n)]$. Indeed this is clearly true for $k = 1$ because F is non-decreasing. Suppose it is true for some k. If you have $k + 1$ intervals take the interval that includes the point b – without loss of generality assume it to be the $(k+1)$-st interval. Observe that the remaining k intervals cover $[a, a_{k+1}]$. Use induction hypothesis and add one more term and simplify.

3^o. Finally, let $I_n = (a_n, b_n]$ be disjoint with union $I = (a, b]$. Need to show that $F(b) - F(a) = \sum[F(b_n) - F(a_n)]$. By observation 1^0, the inequality $F(b) - F(a) \geq \sum[F(b_n) - F(a_n)]$ holds for each finite sum on the right side and hence holds for the total sum. To get the reverse inequality, fix $\epsilon > 0$. Choose $a' \in (a, b]$ with $F(a') < F(a) + \epsilon$. Even if $a = -\infty$,

note that a' is finite. In case $b = \infty$, choose $b' \in (a, b)$ with $b' \geq a'$ so that $F(b') > F(\infty) - \epsilon = 1 - \epsilon$. If $b \neq \infty$ just take $b' = b$. For each n, choose $b'_n > b_n$ so that $F(b'_n) < F(b_n) + \epsilon/2^n$. Of course if $b_n = \infty$ just take $b'_n = b_n$. Clearly the intervals (a_n, b'_n) cover the compact interval $[a', b']$. Take a finitely many of these intervals that cover $[a', b']$. Apply observation 2^0 to see

$$F(b') - F(a') \leq \sum [F(b'_n) - F(a_n)] \leq \sum [F(b_n) - F(a_n)] + \epsilon$$

But in turn, $F(b) - F(a) \leq F(b') - F(a') + 2\epsilon$. Since $\epsilon > 0$ is arbitrary desired inequality is established.

Step 2: This collection of sets $\mathcal{S}$ has three interesting properties ; (1) $\emptyset, R \in \mathcal{S}$. (2) $A, B \in \mathcal{S}$ implies $A \cap B \in \mathcal{S}$. (3) If $A \in \mathcal{S}$ then A^c is a finite disjoint union of sets in $\mathcal{S}$. Such a collection of sets is called a *semifield* of subsets of R. What you name it is unnecessary for us now. Let the collection of finite disjoint unions of sets in this class be denoted by $\mathcal{F}$.

Using the above three properties of $\mathcal{S}$ we verify that $\mathcal{F}$ is a field. Clearly $\emptyset$ and R are already in $\mathcal{S}$ and hence in $\mathcal{F}$. If A and B are in $\mathcal{F}$ then each is a finite disjoint sets in $\mathcal{S}$, say, $A = \cup A_i$ and $B = \cup B_j$ so that $A \cap B = \cup_{i,j} (A_i \cap B_j)$, is also a disjoint union of sets in $\mathcal{S}$ and hence is in $\mathcal{F}$. Finally, if A is in $\mathcal{F}$, say, $A = \cup A_i$, union of disjoint sets in $\mathcal{S}$, then $A^c = \cap A_i^c$. But A^c is a finite disjoint union of sets in $\mathcal{S}$ and hence is in $\mathcal{F}$. But then the intersection is also in $\mathcal{F}$ from what has been established already. This shows $\mathcal{F}$ is indeed a field of sets.

Define for $A \in \mathcal{F}$ which is a disjoint union of finitely many sets $A_n \in \mathcal{S}$, $P(A) = \sum P(A_n)$. Here is the execution of the argument sketched but not worked out in the lecture. Firstly, this is a good definition because if the same A is a disjoint union of finitely many sets B_m, $1 \leq m \leq l$ from $\mathcal{S}$ then

$$\sum_n P(A_n) = \sum_n \sum_m P(A_n \cap B_m) = \sum_m \sum_n P(A_n \cap B_m) = \sum_m P(B_m)$$

Here the first equality is from the facts that for each n, A_n is the union of disjoint sets $\{A_n \cap B_m : 1 \leq m \leq l\}$, these are all in $\mathcal{S}$, and by step 1, P is additive on this class; second equality is just interchange of the order of sums; and the third equality is from the fact that for each m, B_m is the union (now over n) of the disjoint sets $A_n \cap B_m$.

This P is actually countably additive on $\mathcal{F}$. Indeed let A_n be disjoint sets from $\mathcal{F}$ whose union, say, A is again in $\mathcal{F}$. Each of these sets in turn are disjoint union of finitely many sets from $\mathcal{S}$, say $A_n = \cup_j B_{n,j}$ and $A = \cup B_i$. Of course for different n, you may need different number of sets – that is the range of j in the union may depend on n. Let it be so. Now, using additivity

of P from step 1 and the fact that intersection of two sets in $\mathcal{S}$ is again in $\mathcal{S}$, we get

$$\begin{aligned} P(A) &= \sum_i P(B_i) && \text{by definition of } P(A) \\ &= \sum_i \sum_n \sum_j P(B_i \cap B_{n,j}) && B_i \text{ is disjoint union of } B_i \cap B_{n,j} \\ &= \sum_n \sum_i \sum_j P(B_i \cap B_{n,j}) && \text{interchange of order of sum} \\ &= \sum_n P(A_n) && A_n \text{ is dijoint union over } i,\ j \text{ of } B_i \cap B_{n,j} \end{aligned}$$

To complete the proof, appeal to Caratheodory and verify that $\sigma(\mathcal{F})$ is precisely the Borel σ-field of R.

Caratheodory extension theorem has a natural proof, but lengthy and we shall not discuss. We should note that, in general, there are many many sets in the σ-field generated by a given family $\mathcal{C}$ – just as the collection of Borel sets of R has. It is not easy to *see* all of them. The natural question that arises is: If I do not see all the sets in my σ-field how am I going to handle them and in particular, define probabilities? Coming to think of it, aren't there too many real numbers – too many to be listed. Nonetheless we had no problem in doing calculus. There are several ways of handling σ-fields. Here is one way.

Say that a family $\mathcal{M}$ of subsets of Ω is a *monotone class* if the following holds: if $A_n \uparrow A$ or $A_n \downarrow A$ and each $A_n \in \mathcal{M}$ then $A \in \mathcal{M}$.

For example if Ω is the set of real numbers, then the collection of all intervals (empty, degenerate, open, closed, semi-open) is a monotone class. Again, like theorem (3A), given any class $\mathcal{C}$ of subsets of Ω, one can show that there is a smallest monotone class $\mathcal{M}(\mathcal{C})$ of subsets of Ω that includes the given class.

Theorem 7 : Monotone Class Theorem, MCT
If $\mathcal{F}$ is a field of subsets of Ω then $\sigma(\mathcal{F}) = \mathcal{M}(\mathcal{F})$.

We shall not prove this, though the proof is simple. Instead, we shall see how it helps us.

Theorem 8: A Uniqueness Theorem for Probabilities
Let $\mathcal{F}$ be a field of subsets of Ω and $\mathcal{A} = \sigma(\mathcal{F})$. Suppose that P and Q are two probabilities on $\mathcal{A}$. If they agree on $\mathcal{F}$ then they agree on $\mathcal{A}$. In otherwords if $P(F) = Q(F)$ for every set $F \in \mathcal{F}$, then $P(A) = Q(A)$ for

every set $A \in \mathcal{A}$.

Proof is very simple. The class of sets $A \in \mathcal{A}$ such that $P(A) = Q(A)$ is a monotone classs by theorem 5 and this class includes the field; so that it includes $\mathcal{M}(\mathcal{F}) = \sigma(\mathcal{F}) = \mathcal{A}$ by MCT. But of course this class can not be larger than $\mathcal{A}$ and hence equals it, completing the proof. The beauty is that you need not have a list of all sets in the σ-field before you, nor is it necessary for you to see all sets in the σ-field. Usually fields are simple objects and explicit calculations can be made. That is, you can actually catch hold of all sets in $\mathcal{F}$ and show that the hypothesis is true. If you did that, then there is no need for you to catch hold of all sets in the σ-field to arrive at the conclusion. We will see more applications of the MCT as we go along.

Thus the ingredients for chance experiment consist of a set Ω which describes the set of all possible outcomes of the experiment; a σ-field $\mathcal{A}$ of subsets of Ω which consists of the collection of events, and a (countably additive) probability P defined on the collection of events. The next question is : what should be a random variable. Generally, we perform an experiment and make a measurement. This measurement is called random variable. For example the experiment may consist of tossing a coin 20 times, and you may want to count the number of heads obtained. Here the number of heads is the random variable. Thus a random variable is nothing but measurement, that is, assignment of numerical value to each outcome. Mathematically it is just a real valued function defined on the sample space Ω.

In dealing with general chance experiments, we have to be a little more careful. We may endup with objects about which we can not answer even simple questions. For example, suppose X is a measurement – that is, a real valued function on Ω. We are interested in the chances that the measurement is no more than 23. To calculate this, we should collect the set A of all sample points for which the measurement is at most 23 and then calculate the probability of this event A. This is fine, but the only problem is that I have already used the word 'event' for the set A. How do I know that this set A is in my σ-field? There is nothing so far to tell us that $A \in \mathcal{A}$. In fact, it need not be. In such a case we can not answer the simple question raised above. Is there any point admitting such measurements, if we can not answer even simple questions regarding them. We shall not admit such measurements as random variables.

Thus for us a *random variable* is a real valued function X defined on Ω such that for every real number a, the set $\{\omega : X(\omega) \leq a\}$ is in $\mathcal{A}$. . Of

course, the concept depends only on the σ-field and not on the probability P (as it should be – because measurement depends on the sample points and not on what probability model is assumed). We shall denote by L the collection of all random vaiables. We shall denote by $\mathcal{E}$ the collection of those random variables which take only finitely many values, such random variables are called *simple random variables*.

There are several questions/doubts that arise immediately. First, how do you know that your demand above is good enough to answer all possible simple questions. By properties of inverse images, clearly those sets $B \subset R$ such that $X^{-1}(B) \in \mathcal{A}$ is a σ-field and if it includes sets of the form $(-\infty, a)$ then it includes all Borel sets. As a result, $X \in L$ iff for every Borel set $B \subset R$, we have $X^{-1}(B) \in \mathcal{A}$.

Second, my measurement X is very important to me, does not satisfy your requirement, how can you exclude it from discussion? The point is well made. But then the answer is: your choice of σ-field is wrong. Your initial choice of events does not take into account the fact that you need to answer questions regarding this particular X. You should start with a better σ-field which makes X a random variable.

Third, how do I know that you are admitting enough random variables or not? Imagine for a moment that our definition is so bad that only constant functions are random variables, no matter what $\mathcal{A}$ is. Well, there are enough random variables if there are enough events, as we will see shortly.

Finally, how do I know that the definition has reasonable properties. For example if you have admitted two measurements for discussion (as random variables) it is only fair that their sum or product etc, which is again a measurement based on my experiment, should be admitted too.

Here are two theorems that answer all these and other questions in a satisfactory way.

Theorem 3B : Getting σ-fields
Let Ω be any set and Φ be any collection of real valued functions on Ω. Then there is a σ-field $\mathcal{A}$ of subsets of Ω such that each function in Φ is a random variable w.r.t. this σ-field and it is the smallest such.

Clearly, the σ-field of the above theorem is unique. It is named as the σ-field generated by the class Φ, denoted $\sigma(\Phi)$.

Lecture 4/5: Random variables and Expectation.

We shall first understand the nature of the collection of random variables.

Theorem 9 : Structure of Random Variables

(i) The space L is a real vector space, that is, if X and Y are in L, then so is $\alpha X + \beta Y$ for any real numbers α and β. L is an algebra, that is, besides being a vector space it is closed under multiplication. L is a lattice too, that is, the (pointwise) maximum and minimum of two random variables is again a random variable. Constant functions are in L. Thus in particular, for any random variable X, both $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$ are random variables. Clearly, $X = X^+ - X^-$.

(ii) If $X_n \in L$ and $X_n \uparrow X$ or $X_n \downarrow X$ pointwise and X is real valued, then $X \in L$. In particular, pointwise lim sup and lim inf of any sequence of random variables X_n are in L provided they are real valued. If a sequence of random variables (X_n) converges pointwise to an X, then X is a random variable.

(iii) $I_A \in L$ iff $A \in \mathcal{A}$. The collection of simple random variables is also a vector space, algebra and lattice – as in (i) above.

(iv) Given any bounded random variable X there is a sequence (s_n) of simple random variables such that $s_n \uparrow X$ uniformly. Given any non-negative random variable X, there is a sequence (s_n) of simple non-negative random variables such that $s_n \uparrow X$ point wise.

(v) Suppose that $\mathcal{F}$ is a field of sets such that $\mathcal{A} = \sigma(\mathcal{F})$. Suppose that Φ is a vector space of real valued functions on Ω closed under monotone limits — which means that if each $X_n \in \Phi$ and $X_n \uparrow X$ or $X_n \downarrow X$ pointwise and X is real valued, then $X \in \Phi$. Suppose that for each $A \in \mathcal{F}$, we have $I_A \in \mathcal{F}$. Then $L \subset \Phi$.

It is rather awkward that we needed to assume in (ii) above that the lim sup etc are real valued. Actually this is unnecessary. But we have to admit extended real valued functions to make things precise. We are not doing this at present. The last part of the theorem above helps us in dealing with the collection of all random variables just as the MCT helps us in dealing with all sets in the σ-field.

Here is the proof. To prove X is a random variable it suffices to show that for any number a, the set $X^{-1}(-\infty, a)$ is in our σ-field, because then, $X^{-1}(-\infty, a] = \bigcap X^{-1}(-\infty, a + \frac{1}{n})$ will also be in our σ-field.

(i) For any number a, $(X + Y < a) = \cup_r \{(X < r) \cap (Y < a - r)\}$ where

the union runs over the countable set of rationals r. Thus if X and Y are random variables then so is $X+Y$. For $\alpha > 0$, $(\alpha X < a) = (X < a/\alpha)$ and for $\alpha < 0$, $(\alpha X < a) = (X > a/\alpha)$ to conclude that αX is a random variable if X is so. Observe that $(X^2 \leq a)$ is empty set if $a < 0$, where as, it equals $(-\sqrt{a} \leq X \leq +\sqrt{a})$ for $a > 0$ showing that X^2 is a random variable if X is so. Thus in particular combining these it follows that if X and Y are random variables then so is $XY = [(X+Y)^2 - X^2 - Y^2]/2$. Observe that $(X \wedge Y < a) = (X < a) \cup (Y < a)$ and $(X \vee Y < a) = (X < a) \cap (Y < a)$ so that $X \wedge Y$ and $X \vee Y$ are random variables if X and Y are so. Since constant functions are visibly random variables this proves (i).

(ii) If $X_n \uparrow X$ then $(X \leq a) = \cap(X_n \leq a)$. If $X_n \downarrow X$ then $(X < a) = \cup(X_n < a)$ to prove the first statement. Let X_n be random variables. Fix n. For each $m > n$, $\max\{X_n, X_{n+1}, \cdots, X_m\}$ is a random variable by (i) and increases to, say, Y_n as $m \uparrow \infty$. By the first statement each Y_n is a random variable. But then $Y_n \downarrow \limsup X_n$ so that this last one is also a random variable. Similar argument holds for liminf.

(iii) Since $(I_A < 1) = A^c$ it follows that I_A is a random variable iff A is in our σ-field. For the second statement, you only need to verify that sums, products, scalar multiples of simple functions are simple again.

(iv) For convenience assume that $0 \leq X \leq 1$. Define $s_n(\omega)$ as follows: if $0 \leq X(\omega) \leq 1/2^n$ then $s_n(\omega) = 0$ and for $1 \leq k \leq 2^n - 1$ put $s_n(\omega) = k/2^n$ in case $k/2^n < X(\omega) \leq (k+1)/2^n$. This sequence does.

Given non-negative X and an integer $n \geq 1$ define s_n as follows: $s_n(\omega) = 0$ in case $0 \leq X(\omega) \leq 1/2^n$ and $s_n(\omega) = n$ in case $X(\omega) > n$. For $1 \leq k \leq n2^n - 1$, in case $k/2^n < X(\omega) \leq (k+1)/2^n$ put $s_n(\omega) = k/2^n$. This sequence does.

(v) First observe that the class of sets A such that $I_A \in \Phi$ is a monotone class and includes $\mathcal{F}$ — all this by hypothesis. So by MCT, we conclude that $I_A \in \Phi$ for each $A \in \mathcal{A}$. Since Φ is a vector space, it must include every simple random variable. By hypothesis and (iv) it must include every nonnegative random variable too. For any random variable X, since $X = X^+ - X^-$ we conclude that X must also be in Φ to complete proof.

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space. Suppose that s is a simple non-negative random variable say $s = \sum a_i I_{A_i}$ where for each i, $a_i \geq 0$. We define $E(s) = \sum a_i P(A_i)$. If X is a non-negative random variable then we

put $E(X) = \sup\{E(s) : s \text{ simple }, 0 \leq s \leq X\}$. If X is any random variable, we say that X is integrable iff $E(X^+) < \infty$ and $E(X^-) < \infty$ and in such a case we put $E(X) = E(X^+) - E(X^-)$. We shall also write $\int X dP$ instead of $E(X)$. We denote the collection of integrable random variables by $L^1(\Omega, \mathcal{A}, P)$, in short, L^1 when other things are understood.

The first question that should be answered is the reasons to adapt this definition. Suppose we have $s = I_A$, then our definition gives $E(s) = P(A)$. This stands to reason because s takes two values one and zero with probabilities $P(A)$ and $1-P(A)$ so that, using what we learnt in lecture 1, we should indeed have $E(s) = P(A)$. Since expectation is, in a sense, average value it should be linear. So for non-negative simple functions it should be as defined above. Another way of thinking of it is as follows: say s takes the distinct values $a_i \geq 0$ and let $A_i = \{\omega : s(\omega) = a_i\}$ so that $s = \sum a_i I_{A_i}$. Then as you can see the random variable s takes the value a_i with probability $P(A_i)$ and hence its expectation should be $\sum a_i P(A_i)$. This is what we get from our definition too. Of course what is not clear is whether any representation of simple function will lead to the same answer. That is, one needs to show that the formula for $E(s)$ does not depend on the representation used in the definition.

It is clear that if we have two measurements $X \leq Y$ then the average value of X should not exceed the average value of Y. Thus in particular for every simple non-negative $s \leq X$ we must have $E(s) \leq E(X)$. Thus $E(X)$ should not fall below the value we have prescribed. The best way at this stage is to see if we can get an upper bound for the possible value of $E(X)$ and then decide what to do. If luckily these two values coincide then there is nothing for us to decide. We did not do this (though this path can also be pursued carefully). Since there is no *a priori* reason why $E(X)$ should exceed this lower bound (if so by how much? etc) we have declared this lower bound as $E(X)$. If you take a general random variable X, note that $X = X^+ - X^-$ so that linearity justifies the definition we gave. Of course it is not clear whether the expectation so defined has reasonable properties at all. The next theorem says that all properties you expect from *expectation* hold with this definition.

Theorem 10: Properties of Expectation

(i) If a simple nonnegative random variable s is expressed in two different ways as $\sum a_i I_{A_i}$ with $a_i \geq 0$ as well as $\sum b_j I_{B_j}$ with $b_j \geq 0$ then we have $\sum a_i P(A_i) = \sum b_j P(B_j)$. In other words $E(s)$ for simple nonnegative random variables is well defined. $E(s)$ is nonnegative. It is linear in the sense that $E(\alpha s + \beta t) = \alpha E(s) + \beta E(t)$ for simple nonnegative s and t and non-negative

reals α and β. Also $E(s)$ is monotone, if $s \leq t$ then $E(s) \leq E(t)$.

(ii) The clause defining integral for non-negative random variables extends the definition given for simple non-negative random variables earlier. $E(X)$ is monotone on the collection of non-negative random variables.

(iii) (Monotone convergence theorem ; **MCT**) If $X_n \geq 0$ and $X_n \uparrow X$ then $E(X_n) \uparrow E(X)$.

(iv) If $X \geq 0$, $Y \geq 0$ and $\alpha, \beta \geq 0$ then $E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$.

(v) The final clause defining integral for general (integrable) random variables extends the definition for non-negative integrable random variables. L^1 is a vector space and expectation is linear on this space and is also monotone. Further, $X \in L^1$ iff $|X| \in L^1$. Finally, $|\int X dP| \leq \int |X| dP$.

(vi) (**Fatou's** lemma) $\int \liminf X_n dP \leq \liminf \int X_n dP$ for nonnegative random variables X_n,

(vii) (**Lebesgue's** Dominated convergence theorem; **DCT**) If $X_n \in L^1$ for each n; $X_n \to X$ pointwise ; there is $Y \in L^1$ such that $|X_n| \leq Y$ pointwise for all n, then $X \in L^1$ and $\int |X_n - X| dP \to 0$ and $\int X_n dP \to \int X dP$.

(viii) Let X be a non-negative random variable. Then $P(X \neq 0) = 0$ iff for every set $A \in \mathcal{A}$, $E(XI_A) = 0$. As a consequence, for two integrable random variables X and Y, $P(X \neq Y) = 0$ iff for every set $A \in \mathcal{A}$, $E(XI_A) = E(YI_A)$.

Here is the proof.

(i) We shall show well-defined-ness shortly. Accepting this, the other parts can be observed easily as follows. Suppose that $s = \sum_1^n a_i I_{A_i}$ and $t = \sum_1^m b_j I_{B_j}$ then clearly $s + t = \sum_1^{n+m} c_i I_{C_i}$ where $c_i = a_i$, $C_i = A_i$ for $1 \leq i \leq n$ and $c_i = b_{i-n}$, $C_i = B_{i-n}$ for $n + 1 \leq i \leq n + m$. Now use the definitions of expectation to see $E(s + t) = E(s) + E(t)$. If $s \leq t$, then $t - s$ is a simple function and $E(t) = E(s) + E(t - s) \geq E(s)$.

(ii) Just to avoid confusion, temporarily use $\tilde{E}$ for expectation defined for non-negative random variables. Let X be simple nonnegative random variable. Note that X itself is a candidate for s in the definition of $\tilde{E}(X)$. Thus $\tilde{E}(X) \geq E(X)$. On the other hand if we take any other non-negative simple $s \leq X$ then monotonicity of expectation, observed in (i) above, yields that $E(s) \leq E(X)$ showing that $\tilde{E}(X) = E(X)$. Monotonicity of expectation is immediate from definition.

(iii) Start observing that if s is a simple nonnegative random variable and Ω_n are sets in the σ-field increasing to Ω then $E(sI_{\Omega_n}) \uparrow E(s)$. This is done by direct computation. If $s = \sum c_i I_{B_i}$ (finite sum), then $sI_{\Omega_n} = \sum c_i I_{B_i \cap \Omega_n}$

and hence

$$E(sI_{\Omega_n}) = \sum c_i P(B_i \cap \Omega_n) \uparrow \sum c_i P(B_i) = E(s)$$

For the proof of MCT, let $X_n \uparrow X$ – all being nonnegative. We already know that $E(X_n) \leq E(X)$ and increases with n, so that $\lim E(X_n) \leq E(X)$. Towards the other inequality fix any simple nonnegative $s \leq X$. Suffices to show that $E(s) \leq \lim E(X_n)$. Again fix $0 < \alpha < 1$. Suffices to show that $\alpha E(s) = E(\alpha s) \leq \lim E(X_n)$. Let $\Omega_n = \{\omega : \alpha s(\omega) \leq X_n(\omega)\}$, so that $\Omega_n \uparrow \Omega$. By monotonicity of expectation, we have $E(\alpha s I_{\Omega_n}) \leq E(X_n I_{\Omega_n}) \leq E(X_n)$. Take limits and use the observation made at the beginning to complete the proof.

(iv) Get simple nonnegative $s_n \uparrow X$ and $t_n \uparrow Y$. From (i), $E(s_n + t_n) = E(s_n) + E(t_n)$, use MCT for all the three terms noting that $s_n + t_n \uparrow X + Y$ to complete the proof.

(v) That the definition extends from non-negative variables is clear since, if $X \geq 0$ then $X^+ = X$ and $X^- = 0$ so that the new definition is same as the old one. First note that X is integrable iff both $E(X^+)$ and $E(X^-)$ are finite – equivalently, $E(|X|) = E(X^+) + E(X^-)$ is finite. Thus note that $X \in L^1$ iff $|X| \in L^1$. Further by definition, $|E(X)| = |E(X^+) - E(X^-)|$ is at most $E(X^+) + E(X^-)$ which is $E(|X|)$.

In particular, if X and Y are integrable, then so are $|X|$ and $|Y|$ and hence their sum because of additivity of expectation on non-negative random variables. Since integral is monotone on non-negative random variables and since $|X + Y| \leq |X| + |Y|$ we conclude that $E(|X + Y|)$ is finite and hence $X + Y$ is integrable. Finally, since

$$(X+Y)^+ - (X+Y)^- \quad = X + Y = \quad X^+ - X^- + Y^+ - Y^-;$$

we conclude that

$$(X+Y)^+ + X^- + Y^- = (X+Y)^- + X^+ + Y^+.$$

Now take expectation on both sides and since all terms are finite, rearrange them carefully to get $E(X + Y) = E(X) + E(Y)$. Monotonicity is routine.

(vi) If $Y_n = \inf\{X_n, X_{n+1}, \cdots\}$ then $Y_n \leq X_n$. So $E(Y_n) \leq E(X_n)$ for all n which implies that $\liminf E(Y_n) \leq \liminf E(X_n)$. Note that Y_n are increasing to $\liminf X_n$ and so by MCT, the left side is actually $E(\liminf X_n)$ to complete the proof.

(vii) The hypotheses imply that for the limit random variable X too we have $|X| \leq Y$ and hence X is integrable. So are $|X_n - X|$. The non-negative

random variables $2Y - |X_n - X|$ converge pointwise to $2Y$. An application of Fatou gives $\int 2Y dP \leq \liminf \int (2Y - |X_n - X|) dP$. Linearity of integral and finiteness of the integrals involved gives after cancellation of the Y integral $\limsup \int |X_n - X| dP \leq 0$. This completes the proof.

We shall now show the well-defined-ness of the integral for non-negative simple random variables. Just to understand the issue, note that we have $s = 3I_{[0,4]}$ is same as $3I_{[0,2]} + 3I_{(2,4]}$. This is also same as $2I_{[0,3]} + I_{(2,4]} + I_{[0,2]} + I_{(3,4]}$ and so on. To proceed with the proof, let s be a simple random variable. Let $v_1, v_2, \cdots, v_k$ be the distinct values of s and $V_i = s^{-1}(v_i)$ so that we have $s = \sum v_i I_{V_i}$. Consequently $E(s) = \sum v_i P(V_i) = \alpha$, say. We show any representation of s leads to the same value.

Suppose that s is represented as $s = \sum a_i I_{A_i}$ where the sets A_i are non-empty and form a partition of Ω. Clearly then each a_i is a possible value of s and hence must be a v_j for some j. Further $V_j = \sum\limits_{i:A_i \subset V_j} A_i$ and consequently $P(V_j) = \sum\limits_{i:A_i \subset V_j} P(A_i)$. Thus

$$\sum a_i P(A_i) = \sum_j \sum_{i:A_i \subset V_j} a_i P(A_i) = \sum_j v_j \sum_{i:A_i \subset V_j} P(A_i) = \sum_j v_j P(V_j) = \alpha$$

Suppose finally that s is represented as $s = \sum_{i=1}^m b_i I_{B_i}$. The sets (B_j) may not be disjoint. First we claim that there is a finite partition A_j, $1 \leq j \leq k$ such that each B_i is a union of sets in the partition. In fact, consider all sets of the form $\cap_{i=1}^m C_i$ where each C_i is either B_i or its complement. Enumerate the non-empty sets you have so obtained. Clearly you get finitely many such sets, they are disjoint, and each B_i is union of some of these sets and all these sets are in your σ-field. Let these be A_i, $1 \leq j \leq k$. Denote $a_j = \sum\limits_{i:A_j \subset B_i} b_i$. The representation of s tells us that a_j is the value of s on the set A_j. In other words $s = \sum a_j I_{A_j}$ and hence from what was observed earlier, we get $\sum a_j P(A_j) = \alpha$. Note also that $P(B_i) = \sum\limits_{j:A_j \subset B_i} P(A_j)$. As a consequence,

$$\sum_i b_i P(B_i) = \sum_i b_i \sum_{j:A_j \subset B_i} P(A_j) = \sum_j P(A_j) \sum_{i:A_j \subset B_i} b_i = \sum_j a_j P(A_j) = \alpha$$

completing the proof.

There are some important details that need to be attended to. There are no new ideas – only extensions of the ideas discussed so far. This is simply a fine tuning of the development so far.

infinities

We need to discuss random variables that may take values $\pm\infty$ also. Not that we invite them, but they may crop up in our calculations. We realized it while dealing with lim sup and lim inf of random variables. There are other reasons too. Suppose you take i.i.d normal observations and wish to study the properties of the observed averages. There is no reason why the observed averages can not go to ∞. In fact they do (of course with probability zero).

Extended real line means the set $\{-\infty\} \cup R \cup \{+\infty\}$, with the understanding that $-\infty < x < +\infty$ for all real numbers x. Usually $+\infty$ is also denoted as ∞. A function taking values in this set is said to be an extended real valued function. Such a function X is said to be an extended real random variable if for every real number x, $\{\omega : X(\omega) \leq x\} \in \mathcal{A}$ – as in the case of real random variables. We have the common sense arithmetic on this extended real line. For every real x, $\pm\infty + x = \pm\infty$; $\infty + \infty = \infty$; $-\infty - \infty = -\infty$. We agree NOT to talk about $\infty - \infty$. Also $x.\infty = \infty$ or 0 or $-\infty$ according as $x > 0$ or $x = 0$ or $x < 0$.

With this arithmetic we can define sums of two extended random variables provided at no sample point one of them assumes value ∞ and the other assumes the value $-\infty$. Then the sum will also be an extended real random variable. Limsups and lim infs are all (extended real) random variables. Integration can also be defined. Of course one still takes the same old definition for simple random variables – that is, the values $\pm\infty$ are not allowed for simple random variables.

σ-finite measures

Sometimes it is necessary to look at things defined on σ-fields which do satisfy countable additivity and are non-negative but the value for whole space is not necessarily one.

Suppose that Ω is a set and $\mathcal{A}$ is a σ-field of subsets of Ω. A function μ defined on $\mathcal{A}$ with values in $[0, \infty)$ is said to be a *finite measure* if $\mu(\emptyset) = 0$, and μ is countably additive. Clearly if $\mu(\Omega) = c$ then for every set $A \in \mathcal{A}$ we have $\mu(A) \leq c$.

A function μ defined on $\mathcal{A}$ with values in $[0, \infty]$ is said to be a σ-finite measure if $\mu(\emptyset) = 0$, μ is countably additive and Ω is a union of countably many sets in $\mathcal{A}$ each having finite measure – that is, each having finite μ value.

The extension theorem of Caratheodory is immediate for finite measures. It is also easy to deduce it for measures which are σ-finite on the field $\mathcal{F}$ – that is, when μ given on $\mathcal{F}$ is such that there exists a sequence of sets in $\mathcal{F}$ of finite μ value whose union is Ω then it can be extended to $\sigma(\mathcal{F})$. Since the definition of random variable does not depend on the probability/measure there is no new problem. We can define integral as in the case of probability

and all (nearly) those theorems are valid. But of course some care is needed. For example if μ is σ-finite then $A_n \downarrow A$ need not imply $\mu(A_n) \downarrow \mu(A)$. Similarly if μ is not finite (but σ-finite) then constant function $\mathbf{1}$ is not integrable whereas on a probability space the constant functions are always integrable.

There is a unique σ-finite measure λ on the Borel σ-field of the real line such that for intervals it coincides with length – that is for $a < b$, we have $\lambda(a, b] = b - a$. This is called the Lebesgue measure. For instance, the standard normal density is a density w.r.t this measure.

complex random variables

If you have a random variable then $M(t) = E(e^{tX})$ is the moment generating function of the random variable X. This function played a role in the inequalities in lecture 1. The only problem is that the required expectation may not exist and hence this function may not be defined for all values of t. However if we change slightly, $\varphi(t) = E(e^{itX})$ always exists and as you know this is called the characterstic function of the random variable X. This function helps in the study of the random variable X and in particular in the Central limit theorem, a topic JM would discuss. In fact an analysis involving characterstic functions will put at our disposal all the results of complex function theory. So there is advantage in discussing complex valued functions.

Note that if X is a complex valued function on a set Ω, then there are two uniquely defined real valued functions X_1 and X_2 such that $X(\omega) = X_1(\omega) + iX_2(\omega)$. Simply take $X_1(\omega)$ as real part and $X_2(\omega)$ as the imaginary part of $X(\omega)$. If we have a σ-field $\mathcal{A}$ on Ω then we say that X is a random variable iff both X_1 and X_2 are so. We already know when to say a real valued function is a random variable. If moreover we have a probability P on $\mathcal{A}$ then we say that X is integrable if both X_1 and X_2 are so and then we define $E(X) = E(X_1) + iE(X_2)$.

All the results for expectation remain valid. Of course there is no linear order on complex numbers and hence MCT does not make sense. But the fact $|E(X)| \leq E|X|$ and the DCT remain valid.

almost everywhere

While dealing with random variables we should not ignore the existence of probability. For instance $X \leq Y$ if for every sample point ω, $X(\omega) \leq Y(\omega)$ is a statement that does not take into account the existence of a probability. From a probabilistic point of view, if we have a property that depends on sample points then we should agree that it holds provided those points where it fails has probability zero. This is what we discuss now.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. For two random variables X and Y, say that $X \leq Y$ *a.e.* if $P[\omega : X(\omega) > Y(\omega)] = 0$. Say that $X = Y$ *a.e.* if $P[\omega : X(\omega) \neq Y(\omega)] = 0$. Say that $X_n \uparrow$ *a.e.* if $P[\omega : X_n(\omega) > X_{n+1}(\omega)$ for some $\omega] = 0$. Say that $X_n \to X$ *a.e.* if $P[\omega : X_n(\omega) \not\to X(\omega)] = 0$. Here *a.e.* is abbreviation for almost everywhere. Observe that to make sense of these definitions, we must make sure that the sets in braces are indeed events. But once you recognize the problem, it is easy to sort out. Most of the theorems on integration can be fine tuned by putting *a.e.* For example, MCT can be restated as follows : if $X_n \uparrow X$ *a.e.* and $X_n \geq 0$ *a.e.* then $E(X_n) \uparrow E(X)$. DCT can be restated as follows : If $X_n \in L^1$ for each n; $X_n \to X$ *a.e.* ; there is $Y \in L^1$ such that $|X_n| \leq Y$ *a.e.* for all n, then $X \in L^1$ and $\int |X_n - X| dP \to 0$ and $\int X_n dP \to \int X dP$.

If you take the distribution function $F(x) = 0$ for $x \leq 0$; $F(x) = x$ for $0 \leq x \leq 1$ and $F(x) = 1$ for $x \geq 1$ you will get a probability on $(R, \mathcal{B})$ which gives zero value for the complement of $[0, 1]$. So you can actually think of it as a probability on $[0, 1]$ and the Borel σ-field restricted to this interval, that is, the collection of sets $\{B \cap [0, 1] : B \in \mathcal{B}\}$. This is called the Lebesgue measure (λ) on the unit interval. Suppose you take a continuous function f on $[0, 1]$ and calculate the Riemann integral we knew, $(R)\int_0^1 f(x)dx$. We can also regard f as a random variable on $[0, 1]$ and we can calculate the integral $\int f d\lambda$ as developed now. Do we get the same answer? Yes, indeed the Riemann sums can be regarded as actually integrals of simple functions and by MCT (or DCT) they converge to $\int f d\lambda$. But on the other hand the limit of these Riemann sums is, by definition, the Riemann integral.

Does Riemann integral always equal the integral (w.r.t.λ) developed above? Not always, especially when you go to 'improper Riemann integrals'. For example consider the following function on $[0, 1]$. Put $f(x) = (-1)^{(n-1)}(n + 1)$ on the interval $(1/(n + 1), 1/n]$ for $n \geq 1$. Put $f(0) = 0$. Then the improper Riemann integral $\int_0^1 f(x)dx$ exists and equals $\log 2$. However, the integral $\int f d\lambda$ as developed above does not exist. This is not surprising because in the above development if X is integrable, then $|X|$ is integrable whereas for Riemann integral this is not true.

Lecture 6/7: Product spaces and convergence notions.

How do we construct new models of probability spaces from known models? Suppose that you have one probability space $(\Omega, \mathcal{A}, P)$.

Here is one way. Take one sub σ-field $\mathcal{B}$ of $\mathcal{A}$ and restrict your probability P to $\mathcal{B}$ and denote this by Q. Here sub σ-field means that $\mathcal{B}$ is a σ-field of subsets of Ω and every set which is in $\mathcal{B}$ is in $\mathcal{A}$. Then $(\Omega, \mathcal{B}, Q)$ is again a probability space. This will be useful if you already have partial information about the outcome of the experiment – for example you may already know which event in $\mathcal{B}$ actually occurred. Specifically, I may make an observation on a standard normal variable and may anounce to you absolute value of the outcome. There is unceratinty since you still do not know the sign of the observation. This set up will be useful to you when you define conditional expectation and martingales. How do you explain integration in this new model? If X is a random variable in this new model, it is a random variable in the original model and $\int X dQ = \int X dP$. This equalty means that if one side exists, so does the other side and they are equal.

Here is another way. Pick a set $\Omega_0 \in \mathcal{A}$ and restrict your probability to this set and normalize. More precisely take $\mathcal{B} = \{B \subset \Omega_0 : B \in \mathcal{A}\}$. This is a σ-field of subsets of Ω_0. Define for sets B in this class $Q(B) = P(B)/P(\Omega_0)$. Of course, for this to make sense we must assume that $P(\Omega_0) > 0$. I may perform the experiment and already tell you that the the observation is in Ω_0. There is no point in hanging on to the initial assumption that the chance of the observation falling in Ω_0 is $P(\Omega_0)$. We must modify our model to incorporate the given information and make $P(\Omega_0) = 1$. There are several reasons why this Q fits this job. This is conditional probability *given* Ω_0. If X is a random variable in the new model, then the function $\widetilde{X}$ defined as X on the set Ω_0 and zero on $\Omega - \Omega_0$ is a random variable in the original model and further $\int X dQ = \int \widetilde{X} dP/P(\Omega_0)$.

Here is another way. Take a nonnegative random variable X with $E(X) = 1$ and put $Q(A) = \int X I_A \, dP$ for sets $A \in \mathcal{A}$. Then Q is a probability on $\mathcal{A}$ and thus $(\Omega, \mathcal{A}, Q)$ is a probability space. Here X is called density of Q w.r.t P. Thus, for example, normal density is actually density – the only trouble is that it is not density w.r.t. another probability, but it is density w.r.t. the Lebesgue measure which is a σ-finite measure on R. For any random variable Z, we have $\int Z dQ = \int ZX dP$, in the sense that if one side exists, so does the other side and both are equal.

Here is yet another way. Suppose that we have a set Ω' and a σ-field $\mathcal{A}'$

of its subsets, and a map T defined on Ω with values in Ω' such that for every set $A' \in \mathcal{A}'$, we have $T^{-1}(A') \in \mathcal{A}$. Then we can define a probability on $\mathcal{A}'$ by putting $Q(A') = P(T^{-1}A')$. This is indeed a probability on $\mathcal{A}'$ giving us a new probability space $(\Omega', \mathcal{A}', Q)$. This Q is called the induced probability on $\mathcal{A}'$ induced by T, or also the distribution of T. In particular if Ω' happens to be the real line with its Borel σ-field and T is a random variable X, then this Q is called the distribution of the random variable X. This arises in statistics quite often. The original space corresponds to the sample from a population and T is what is called *statistic* based on the sample. If X is a random variable on Ω' then $\widetilde{X}(\omega) = X(T(\omega))$ is a random variable on Ω and $\int X dQ = \int \widetilde{X} dP$.

We shall now describe another way of manufacturing new spaces out of old ones. This is a standard construction – product spaces. For two groups you can consider the product group, for vector spaces you can consider the product vector space etc. For probability spaces also we can construct product probability space. This is not just for the sake of imitation. This will be the main tool for getting independent random variables – or equivalently this models two independent experiments. This can aso be regarded as imitation of defining areas from lengths. Note that if we have a rectangle then its area is the product of lengths of its sides. Of course when we said length we had usual length in mind. But we can, if we wish, use different scales for the two axes. Why not?

Suppose we have two probability spaces $(\Omega_1, \mathcal{A}_1, P_1)$ and $(\Omega_2, \mathcal{A}_2, P_2)$. Define $\Omega = \Omega_1 \times \Omega_2$, that is, cartesian product of the two sets Ω_1 and Ω_2. For sets $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$, consider the rectangle $A_1 \times A_2$. This is a subset of Ω. We do not discuss rectangles if sides are not in the respective σ-fields. The collection of all such rectangles as A_1 varies in $\mathcal{A}_1$ and A_2 varies in $\mathcal{A}_2$ is denoted by $\mathcal{R}$. Let $\mathcal{A} = \sigma(\mathcal{R})$. This is also denoted as $\mathcal{A}_1 \otimes \mathcal{A}_2$.

Theorem 11 : Product probability
There is a unique probability P on $\mathcal{A}$ such that for rectangles we have $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$.

This probability is called the product probability and is also denoted by $P_1 \otimes P_2$. Thus we have a new probability space

$$(\Omega, \mathcal{A}, P) = (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, P_1 \otimes P_2) = \otimes_{i=1}^{2} (\Omega_i, \mathcal{A}_i, P_i)$$

Proof of the theorem is simple. For any set A in the product space, and

$\omega_1 \in \Omega_1$, let

$$A(\omega_1) = \{\omega_2 \in \Omega_2 : (\omega_1, \omega_2) \in A\}.$$

If $A = A_1 \times A_2 \in \mathcal{R}$ then $A(\omega_1)$ is either A_2 or $\emptyset$ according as $\omega_1 \in A_1$ or not. Thus it is in $\mathcal{A}_2$. This holds also for sets in the field consisting of finite disjoint unions of sets in $\mathcal{R}$. MCT now shows that for every $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$ the set $A(\omega_1) \in \mathcal{A}_2$. Thus the function $l(\omega_1) = P_2(A(\omega_1))$ makes sense. Similar argument as above shows that this is a random variable on Ω_1. Thus $Q(A) = \int l dP_1$ makes sense. A careful application of Monotone convergence theorem now shows that this Q is indeed a probability on the product σ-field with the stated condition.

To prove uniqueness, if there are two such, then they agree on sets in $\mathcal{R}$ by the stipulation mentioned. Hence they agree on sets which are finite disjoint union of sets in $\mathcal{R}$. This is a field generating the product σ-field and hence they must agree there too.

It is necessary to explain how to calculate expectations w.r.t. this new probability if we know how to calculate in the given spaces. In calculus, you evaluated double integrals using what is called repeated integration – or to put it differently, you explained how to integrate function of two variables if we know how to integrate functions of one variable. Of course to integrate one function of two variables, you need to integrate several functions of one variable. This is possible in our present situation too.

First we introduce some notation, extending a notation introduced above for sets. Suppose that we have a function f on the product set Ω. For each $\omega_1 \in \Omega_1$ we denote by f_{ω_1} the function defined on Ω_2 by the formula $f_{\omega_1}(\omega_2) = f(\omega_1, \omega_2)$. This is called the vertical section of f at ω_1. Similarly we can define for each $\omega_2 \in \Omega_2$ a function f^{ω_2} on Ω_1 by the formula $f^{\omega_2}(\omega_1) = f(\omega_1, \omega_2)$. This is called the horizontal section of f at ω_2. Of course these are nothing but the same f with appropriate variable fixed.

Theorem 12 : Fubini

(i) Let X be a non-negative random variable on the product space. Then for each ω_1, X_{ω_1} is a random variable on Ω_2 and the map $\omega_1 \longrightarrow E_{P_2}(X_{\omega_1})$ is a random variable on Ω_1. Similarly for each ω_2, X^{ω_2} is a random variable on Ω_1 and the map $\omega_2 \longrightarrow E_{P_1}(X^{\omega_2})$ is a random variable on Ω_2. Finally $E_P(X) = E_{P_1}(E_{P_2}(X_{\omega_1})) = E_{P_2}(E_{P_1}(X_{\omega_2}))$.

(ii) Let X be a P- integrable random variable on the product space. Then for each ω_1, X_{ω_1} is a random variable on Ω_2 and also for *a.e.*, ω_1 it is P_2 inte-

grable and the map $\omega_1 \longrightarrow E_{P_2}(X_{\omega_1})$ is a random variable on Ω_1 with the provision that for those ω_1 for which X_{ω_1} is not P_2 integrable we take the value to be zero. Similarly for each ω_2, X^{ω_2} is a random variable on Ω_1 and also for $a.e.$ ω_2, it is P_1 integrable and the map $\omega_2 \longrightarrow E_{P_1}(X^{\omega_2})$ is a random variable on Ω_2 with the provision that for those ω_2 for which X_{ω_2} is not P_1 integrable we take the value to be zero. Finally $E_P(X) = E_{P_1}(E_{P_2}(X_{\omega_1})) = E_{P_2}(E_{P_1}(X_{\omega_2}))$.

To prove the first part, you show by hand calculation the result holds for indicators of sets in $\mathcal{R}$ and hence for indicators of sets which are finite disjoint unions of sets in $\mathcal{R}$ and use monotone class theorem to show that it holds for all sets in $\mathcal{A}$. This is achieved by applying monotone convergence theorem appropriately because, every nonnegative random variable is an increasing limit of nonnegative simple ones.

The second part follows by carefully applying first part to X^+ and X^-. You only need to note that if a nonnegtive random variable is integrable, then it is finite almost everywhere.

The moral is that you can integrate the variables one by one as you did in calculus. The conclusion is also written as

$$\int X\, dP = \int \left(\int X_{\omega_1}\, dP_2 \right) dP_1 = \int \left(\int X^{\omega_2}\, dP_1 \right) dP_2.$$

It should be noted that for nonnegative random variables we can always talk about expectation – at the worst it may turn out to be ∞. That is why we did not make fuss about integrability in the first part. For general random variables even the symbol $E(X)$ will not make sense unless you are sure that the random variable is integrable. This is one instance where you end up with extended random variables after one integration – remember your original X is real valued.

This theorem is very powerful and we shall see applications soon. The nuisance is that in the second part we demanded that X be P- integrable. You may wonder whether verification of this hypothesis is easy at all. If we need to verify this, then whether we succeeded in explaining integration w.r.t P using integrations w.r.t P_1 and P_2. You are right. The verification of the hypothesis, namely $P-$integrability of X does seem to need integration w.r.t P. However the two parts of the theorem are to be taken together. Given a random variable X on the product space we consider first of all $|X|$ and apply part one. In part one there is no hypothesis. See whether $|X|$ is P-integrable by performing the two repeated integrals in your own convenient order – on $|X|$. If it is P-integrable then apply part two of the theorem to calculate

$E_P(X)$ by performing the repeated integrals in your own convenient order – now on X.

We define two random variables X and Y on a probability space to be independent if for any numbers a and b;

$$P(X \leq a; Y \leq b) = P(X \leq a)P(Y \leq b).$$

This is also same as saying that for any two Borel sets A and B;

$$P(X \in A; Y \in B) = P(X \in A)P(Y \in B).$$

Also recall that for a random variable X, its distribution is the probability on R defined by $\mu(B) = P(X^{-1}(B)$. This is same as saying that μ is the probability induced on R by X. This is also same as saying that μ is the probability on R corresponding to the distribution function F of X.

Suppose we take μ to be the standard normal probability on the real line (with Borel σ-field). It is the probability on $(R, \mathcal{B})$ such that for every $a \in R$, $\mu(-\infty, a] = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$. Consider the product space by taking each space to be $(R, \mathcal{B}, \mu)$. we get the product space $(R^2, \mathcal{B}^2, P)$ where $P = \mu \otimes \mu$. Now suppose that on the product space we consider the random variables : $X(a, b) = a$ and $Y(a, b) = b$ – that is, the co-ordinate random variables. These are indeed random variables. It is immediate that for any two numbers x and y ; $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$. In other words these random variables are independent. Further the distribution of each is standard normal, that is, for each number x, $P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} exp(-t^2/2)\, dt$. Thus we have two independent standard normal variables.

When we built probability models and made our definition of distribution and random variables it was not clear whether we could manufacture two independent random variables. But as seen above, we can. Nothing special about standard normal. You can take any two distributions you want. Again there is nothing special about two, you can manufacture any number. The concept of product can be extended to many spaces. In particular, we can get a probability space which can support independent random variables with given distributions.

Here are simple but powerful and useful facts.

Theorem 13:

(i) **Markov**'s inequality : For a non-negative random variable X and any $a > 0$, $P(X > a) \leq EX/a$.

(ii) **Tchebycheff**'s inequality : For any random variable X with mean μ and finite variance σ^2 , $P(|X-\mu| > a) \leq \sigma^2/a^2$.

(iii) **Borel-Cantelli** lemma : If A_n are events with $\sum P(A_n) < \infty$ then $P(\limsup\ A_n) = 0$. If the events are independent and if $\sum P(A_n) = \infty$ then $P(\limsup\ A_n) = 1$.

Proofs of the inequalities are as in the discrete case. For the Borel cantelli note that, by definition of limsup;

$$P(\limsup A_n) \leq P(\bigcup_{n\geq m} A_n) \leq \sum_{n\geq m} P(A_n),$$

for every m, tail sum of a convergent series and hence must be zero. The last part is argued as follows. Fix any integers $m < k$.

$$P(\bigcap_m^k A_n^c) = \prod_m^k P(A_n^c) = \prod_m^k [1 - P(A_n)] \leq \prod_m^k e^{-P(A_n)} = e^{-\sum_m^k P(A_n)}$$

which converges to zero as $k \to \infty$, because $\sum P(A_n) = \infty$. Thus for each m, $P(\bigcap_{n\geq m} A_n^c) = 0$ giving us $P(\liminf A_n^c) = 0$ and hence $P(\limsup A_n) = 1$.

We conclude our discussion of measure theoretic preliminaries with concepts of convergence for random variables. Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let (X_n) be a sequence of random variables and X be a random variable.

Say that $X_n \to X$ *everywhere* if for every sample point ω, $X_n(\omega) \to X(\omega)$. This concept has no reference to the underlying probability and could have been defined for functions on any set.

Here is a better concept that takes probability into account. Say that $X_n \to X$ *a.e.* if the set of sample points ω for which $X_n(\omega) \to X(\omega)$ has probability one. This concept involves collecting all sample points ω for which the sequence of numbers $\{X_n(\omega)\}$ converges to the number $X(\omega)$ and checking whether this event has probability one.

Here is a weaker concept you came across in WLLN. Say that $X_n \to X$ *in probability* if for every $\epsilon > 0$ we have $P(|X_n - X| > \epsilon) \to 0$. Here you need not actually collect sample points. If you knew the joint distribution of X_n and X, you can calculate the probability above for each n (ϵ being fixed) and see if it converges to zero. Here the chances of X_n and X differing by an amount larger than a *preassigned* quantity goes to zero. On the face of it, it appears as if here also you are actually collecting sample points ω for which $|X_n(\omega) - X(\omega)| > \epsilon$ and calculating its probability Yes, but this can be done

using only the joint distribution of (X_n, X), without going to sample space. Also this involves only two random variables, namely X_n and X. If you look at the *a.e* concept above, to see whether a sample point ω is to be collected or not depends on whether $X_n(\omega)$ converges to $X(\omega)$ or not. This involves looking at *all* your random variables at this point ω.

Here are yet other very useful concepts. Fix $1 \leq p < \infty$. Let L_p be the collection of all random variables X such that $\int |X|^p dP < \infty$. Here is a notion of convergence for random variables in the space L_p. Say that $X_n \to X$ *in* L_p if $\int |X_n - X|^p \to 0$. This is called L_p convergence or convergence in the p-th mean. If $p = 2$, it is called mean square convergence.

The first question that naturally arises is: why so many concepts? It depends on the problem at hand. We will use several of these, convergence in probability and a.e. in WLLN and SLLN; L^p convergence in martingales to be discussed by BR. Next question is the relation between these concepts and their uses. Here are some basic results that explain the relations.

Theorem 14 : Relations between convergence notions

Let $(\Omega, \mathcal{A}, P)$ be a probability space.

(i) $X_n \to X$ a.e. implies $X_n \to X$ in probability, but the converse is not always true.

(ii) Let $1 \leq p < \infty$. If $X_n \to X$ in L_p then $X_n \to X$ in probability, but the converse need not be true even if all the random variables are in the L_p we are talking about.

To prove part (i), fix $\epsilon > 0$.

$$P(|X_n - X| > \epsilon) \leq P(|X_m - X| > \epsilon \text{ for some } m \geq n)$$

Events on the right side are decreasing and their intersection is contained in the set $(X_n \not\to X)$ and hence has probability zero. For proof of (ii) one uses the Markov inequality, $P(|X_n - X| > \epsilon) \leq E|X_n - X|^p/\epsilon^p$. For the other parts, one can construct examples.

It can be shown that the space L_p with $d(X, Y) = (\int |X - Y|^p)^{1/p}$ is indeed a metric space and is complete, that is, every Cauchy sequence converges.

Lecture 8: Kolmogorov zero-one law.

Today we shall discuss the following phenomenon.

Suppose (X_n) is a sequence of independent random variables defined on a probability space. They are independent as well as random, yet they collectively exhibit 'typical behaviour' on several matters.

First we start with understanding independence. Throughout, a probability space $(\Omega, \mathcal{A}, P)$ is fixed. Recall that two random variables X_1 and X_2 are independent if

$$P(X_1 \leq a_1; X_2 \leq a_2) = P(X_1 \leq a_1)P(X_2 \leq a_2) \quad \text{for all} \quad a_1, a_2 \in R.$$

Eventhough the above equation is demanded for real numbers a_i, it holds even if one of the numbers is ∞, that is, for example,

$$P(X_1 < \infty, X_2 \leq a_2) = P(X_1 < \infty)P(X_2 \leq a_2) = P(X_2 \leq a_2);$$

a fact easy to see. Returning to the earlier equaiton, let us fix a number a_2. Now for $-\infty < a_1 \leq b_1 \leq \infty$ write down that equation for a_1 , b_1 and subtract one from the other to get $P(X_1 \in B; X_2 \leq a_2) = P(X_1 \in B)P(X_2 \leq a_2)$ for any $B \in \mathcal{S}$. Recall $\mathcal{S}$ is the collection of intervals $(a, b]$; $-\infty \leq a \leq b \leq \infty$ with the convention that, for example $(4, \infty] = (4, \infty)$. But then the equation remains correct for any set B which is a finite disjoint union of sets in $\mathcal{S}$. This class is a field generating the Borel σ-field $\mathcal{B}$ of R. Now, an application of the monotone class theorem tells us that the equation

$$P(X_1 \in B; X_2 \leq a_2) = P(X_1 \in B)P(X_2 \leq a_2)$$

for all Borel sets. Now fix a Borel set B_1 and repeat similar argument w.r.t. X_2 to see that independence is equivalent to

$$P(X_1 \in B_1; X_2 \in B_2) = P(X_1 \in B_1)P(X_2 \in B_2) \quad \text{for all} \quad B_1, B_2 \in \mathcal{B}.$$

Suppose that we denote by $\mathcal{A}_i$ the class of sets $\{X_i^{-1}(B) : B \in \mathcal{B}\}$. This is called the σ-field generated by X_i because it is the smallest σ-field that makes X_i a random variable and is denoted as $\sigma(X_i)$. To restate the equation we just derived, independence of X_1 and X_2 is equivalent to

$$P(A_1 \cap A_2) = P(A_1)P(A_2) \quad \text{for all} \quad A_1 \in \mathcal{A}_1; A_2 \in \mathcal{A}_2.$$

Suppose we have random variables $X_1, X_2, \cdots, X_n$. Denote $\mathcal{A}_i = \{X_i^{-1}(B) : B \in \mathcal{B}\}$, that is, the σ-field generated by X_i. Then independence of the

random variables $X_1, X_2, \cdots, X_n$ is equivalent to any of the following three conditions.

$$P(X_i \leq a_i; 1 \leq i \leq n) = \prod_1^n P(X_i \leq a_i) \quad \text{for all } \ a_i \in R.$$

$$P(X_i \in B_i; 1 \leq i \leq n) = \prod_1^n P(X_i \in B_i) \quad \text{for all } \ B_i \in \mathcal{B}.$$

$$P(\bigcap_1^n A_i) = \prod_1^n P(A_i) \quad \text{for all } \ A_i \in \mathcal{A}_i; 1 \leq i \leq n.$$

We say that a sequence $(X_i : 1 \geq 1)$ of random variables are independent if for every n, the random variables $\{X_1, X_2, \cdots, X_n\}$ are independent.

Suppose that we have σ-fields $\mathcal{A}_i$; $1 \leq i \leq n$ where for each i, $\mathcal{A}_i \subset \mathcal{A}$. We say that they are independent if $P(\bigcap_1^n A_i) = \prod_1^n P(A_i)$ for any sets $A_i \in \mathcal{A}_i$; $1 \leq i \leq n$. If we have a sequence of σ-fields $(\mathcal{A}_i : i \geq 1)$ with each $\mathcal{A}_i \subset \mathcal{A}$, we say that they are independent if for each n; $(\mathcal{A}_i : 1 \leq i \leq n)$ is so. Here then is the theorem.

Kolmogorov zero-one Law: Let $(\mathcal{A}_i : i \geq 1)$ be an independent sequence of σ-fields with each $\mathcal{A}_i \subset \mathcal{A}$. Let $\mathcal{T}_n = \sigma\{\mathcal{A}_i : i \geq n\}$ and $\mathcal{T} = \bigcap_{n \geq 1} \mathcal{T}_n$. Then for every set $A \in \mathcal{T}$; $P(A) = 0$ or 1.

In other words, if we take a set $A \in \mathcal{T}$ then either almost every sample point is in A or almost no sample point is in A. This is the typical behaviour mentioned above. Incidentally, $\mathcal{T}$ is called the *tail σ-field* and sets in this σ-field are called *tail sets*. Let us see some examples. Suppose we have an independent sequence of random variables. We can take $\mathcal{A}_i = \sigma(X_i)$. Consider the set $A = (1 \leq X_n \leq 2$ for infinitely many $n)$. For any integer $m \geq 1$, we can describe A as the set of points ω such that $1 \leq X_n(\omega) \leq 2$ for infinitely many values of $n \geq m$, This description can be used to show that $A \in \mathcal{T}_m$ for every m and is hence in $\mathcal{T}$. The theorem says that $P(A)$ is either zero or one. In other words either for almost all points ω the set $\{n : 1 \leq X_n(\omega) \leq n)$ is infinite or for almost no point this set is infinite.

Consider the set $A = \{\sum X_n$ converges $)$. Note that, whatever be $m \geq 1$; a series of numbers $\sum a_n$ converges iff $\sum_{n \geq m} a_n$ converges. As above, this can be used to argue that A has probability zero or one. Thus, either for almost every sample point the series $\sum X_n(\omega)$ converges or for almost no point the

series converges. Similarly, the averages $(\sum_1^n X_i)/n$ converge either for almost every sample point or for almost no sample point.

We shall now prove the theorem. To prove the result, we shall show that $\mathcal{T}$ and $\mathcal{T}$ are independent, that is, if you take two sets A and B in $\mathcal{T}$ then $P(A \cap B) = P(A)P(B)$. Then, in particular, if you take a set A in $\mathcal{T}$ then $P(A \cap A) = P(A)P(A)$. That is, $P(A) = [P(A)]^2$, that is, $P(A) = 0$ or 1. Let us denote by $\mathcal{B}_n = \sigma(\mathcal{A}_1, \mathcal{A}_2, \cdots \mathcal{A}_{n-1})$ and $\mathcal{B}_\infty = \sigma(\mathcal{A}_1, \mathcal{A}_2, \cdots)$, that is the σ-field generated by all the sets in all the $\mathcal{A}_n$ put together.

We shall in fact show that $\mathcal{B}_\infty$ and $\mathcal{T}$ are independent. Since $\mathcal{B}_\infty$ which is actually $\mathcal{T}_1$ includes $\mathcal{T}$; independence of $\mathcal{T}$ with itself follows. Fix any n. We show that $\mathcal{B}_n$ is independent of $\mathcal{T}$. Then it follows that if you take any set $T \in \mathcal{T}$ then $P(T \cap B) = P(T)P(B)$ holds for any set $B \in \bigcup \mathcal{B}_n = \mathcal{F}$ (say), the collection of all sets belonging to any one of the $\mathcal{B}_n$. But the class of sets B for which this equality holds is a monotone class. Since the σ-fields $\mathcal{B}_n$ is an increasing family, their union $\mathcal{F}$ is a field and hence by the monotone class theorem, it follows that the class of sets for which the equality holds includes all sets in $\sigma(\mathcal{F}) = \mathcal{B}_\infty$ as desired.

So now fix an n. We shall in fact show that $\mathcal{B}_n$ and $\mathcal{T}_n$ are independent. Since $\mathcal{T} \subset \mathcal{T}_n$ it follows that $\mathcal{B}_n$ and $\mathcal{T}$ are independent. For $m > n$, let $\mathcal{C}_m = \sigma(\mathcal{A}_n, \mathcal{A}_{n+1}, \cdots, \mathcal{A}_m)$. These are again σ-fields increasing in m and as in the earlier para (passing from $\mathcal{B}_n$ to $\mathcal{B}_\infty$), if we can show that for each $m > n$, $\mathcal{B}_n$ and $\mathcal{C}_m$ are independent, then it follows that $\mathcal{B}_n$ and $\mathcal{T}_n$ are independent.

Thus let us fix $m > n$. Let $\mathcal{S} = \{A_1 \cap A_2 \cap \cdots A_{n-1} : A_i \in \mathcal{A}_i; 1 \leq i \leq n-1\}$ and $\mathcal{S}' = \{A_n \cap A_{n+1} \cap \cdots A_m : A_i \in \mathcal{A}_i; n \leq i \leq m\}$. It is easy to see, by independence of the given σ-fields these two families $\mathcal{S}$ and $\mathcal{S}'$ are independent. Both these families satisfy the three conditions for them to form a semifield, namely, (i) contain empty set and whole space; (ii) if two sets are in the family so is their intersection; (iii) if a set is in the family, then its complement is a finte disjoint union of sets in the family. As a consequence, $\mathcal{F}$ = finite disjoint unons of sets in $\mathcal{S}$ and $\mathcal{F}'$ = finite disjoint unions of sets in $\mathcal{S}'$ are fields. Clearly independence of $\mathcal{S}$ and $\mathcal{S}'$ implies that $\mathcal{F}$ nd $\mathcal{F}'$ are also independent. But then by the monotone class theorem, $\sigma(\mathcal{F}) = \mathcal{B}_n$ and $\sigma(\mathcal{F}') = \mathcal{C}_m$ are independent.

This completes the proof.

Lecture 9: Borel's strong law.

Today we shall discuss the following phenomenon.

For any typical number in the interval $[0,1]$, in its expansion (binary or decimal or any) each pattern of digits appears with the right frequency.

Let us start recalling the uniform probability on the interval $[0,1]$. Consider the distribution function F defined as $F(x) = 0$ for $x \leq 0$; $F(x) = x$ for $0 \leq x \leq 1$; and $F(x) = 1$ for $x \geq 1$. Consider the probability λ corresponding to this F on $(R, \mathcal{B})$. The definition $\lambda(a,b] = F(b) - F(a)$ allows us to conclude that the probability of $(-\infty, 0)$ and $(1, \infty)$ ar zero, so that we can restrict λ to $[0,1]$. Accordingly, we restrict to $[0,1]$ in what follows.

Recall that any number x in this interval has a binary expansion, that is, $x = \sum_{i \geq 1} (x_i/2^i)$ where each x_i is either zero or one. Indeed take $x_1 = 0$ or 1 according as the point is in $[0, 1/2]$ or $(1/2, 1]$. Then clearly $|x - (x_1/2)| \leq 1/2$. Divide each of these intervals iinto two halfs, take x_2 to be zero or one according as the point is in the left half or the right half. Then clearly $|x - (x_1/2) - (x_2/2^2)| \leq 1/2^2$ and so on. Of course, there are two such representations for some numbers. Such numbers, having two expansions are countable and hence pose no problem in calculating measures. This expansion is called *binary expansion*.

We define random variables on $([0,1], \mathcal{B}, \lambda)$ for $i \geq 1$ by putting X_i to be the i-th digit in this expansion. It is easy to see that these are indeed random variables. Each of them takes two values zero and one with probabilitiy $1/2$. This is an independent sequence. To verify the definition of independence of a sequence, we can fix n and show $\{X_1, \cdots, X_n\}$ are independent. For example, the event $(X_1 = 1, X_2 = 0, X_3 = 0)$ is the interval

$$\left(\frac{1}{2} + \frac{0}{2^2} + \frac{0}{2^3},\ \frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3}\right] = \left(\frac{4}{8}, \frac{5}{8}\right]$$

whose λ value is $1/2^3$ which is the product of the corresponding probabilities. General case is similar. Let $Y_n = (\sum_1^n X_i)/n$. Thus Y_n gives the proportion of ones in the first n binary digits of the sample point. We shall show that $Y_n \to 1/2$ a.e. Thus for almost every sample point the proportion of the digit one (and hence proportion of zeros) is $1/2$. Of course, we are interpreting proportion in an infinite sequence to be the limit of the usual proportion in the first n places. This limit need not exist for all sample points, that it

exists for almost every point is also part of the conclusion.

To prove the statement, we show that $\sum \lambda(|Y_n - 1/2| > \epsilon) < \infty$ for every $\epsilon > 0$. This will then complete the proof as follows. If limsup of the events that appear in braces is denoted by N_ϵ, then Borel-Cantelli tells us that $\lambda(N_\epsilon) = 0$. Let $N = \bigcup_{k\geq 1} N_{1/k}$. Then $\lambda(N) = 0$. If $x \notin N$, then it does not belong to $N_{1/k}$, that is, $|Y_n(x) - 1/2| > 1/k$ only for finitely many n. This being true for every integer $k \geq 1$, we conclude that $Y_n(x)$ does indeed converge to $1/2$. So fix $\epsilon > 0$. By Markov,

$$\lambda(|Y_n - 1/2| > \epsilon) \leq \frac{E(Y_n - 1/2)^4}{\epsilon^4} = \frac{1}{n^4\epsilon^4} E\left(\sum_1^n (X_i - 1/2)\right)^4.$$

Towards understanding the right side, first note that each $|X_n - 1/2|$ is at most one because values of X_n are zero and one. The expansion on the right side consists of terms $(X_i - 1/2)(X_j - 1/2)(X_k - 1/2)(X_l - 1/2)$. If one of the indices appears only once in this product then the expectation is zero, use independence. Thus the only terms whose expectation is non-zero are of the form $(X_i - 1/2)^4$ and $(X_i - 1/2)^2(X_j - 1/2)^2$ $(i \neq j)$. There are n terms of the first kind and at most $6n^2$ of the second kind. Since each term is atmost one, the expectation is at most $7n^2$ and thus returning to the above inequality, the right side is at most a number times $1/n^2$ and hence their sum is finite. This completes the proof.

Let us fix a pattern of length three, say $\langle 010\rangle$. Let us put $Z_i = 1$ if $\langle X_i X_{i+1} X_{i+2}\rangle = \langle 010\rangle$. Thus Z_i takes zero-one values and has expectation $1/8$. Put as earlier, $Y_n = (\sum_1^n Z_i)/n$, that is, the proportion of the pattern in the first n places (to know this, you should look at the first $n + 2$ digits in the expansion). We now show $Y_n \to 1/8$ a.e. Thus the pattern appears with proportion $1/8$. Proceed as earlier to get,

$$\lambda(|Y_n - 1/8| > \epsilon) \leq \frac{E(Y_n - 1/8)^4}{\epsilon^4} = \frac{1}{n^4\epsilon^4} E\left(\sum_1^n (Z_i - 1/8)\right)^4.$$

We use the same method of estimating as earlier. Because the (Z_i) are not independent; not all the terms which were zero earlier become zero now. We have only to show that most of them are still zero. This is made possible by the fact that though the sequence (Z_i) is not independent, it is clear that Z_i depends on $Z_{i\pm 2}, Z_{i\pm 1}$ but independent of others. Let us a take term of the type $(Z_i - 1/8)^3(Z_j - 1/8)$. Its expectation is zero unless j is $i \pm 2$, $i \pm 1$.

So at most $4n$ such terms will be non-zero. Let us take a term of the type $(Z_i - 1/8)(Z_j - 1/8)(Z_k - 1/8)(Z_l - 1/8)$ with $(i < j < k < l)$. Unless $j = i + 1, i + 2$ and $k = l - 1, l - 2$ its expectation is zero and thus there are at most $4n^2$ such terms whose expectation is non-zero. Arguing this way for all the terms, we see that the expectation on the right side of the inequality above is at most a constant times n^2 so that the left side is summable to conplete the proof of the claim made.

We can show, in the same way the following. Fix any pattern of length k, that is a sequence of zeros and ones of length k. Put Z_i to be one or zero according as $\langle X_i X_{i+1} \cdots X_{i+k-1} \rangle$ equals that pattern or not and put $Y_n = (\sum_1^n Z_i)/n$. Then $Y_n \to 1/2^k$ a.e. We can make a better statement. Almost surely every number has the following property: in its binary expansion, the proportion of any pattern exists and equals $1/2^k$ where k is length of the pattern. The difference between the present statement and the previous one is this. Earlier, given a pattern, there is a null set (a set of probability zero) such that for points outside that null set the pattern occurs with right frequency. Now we are saying that there is a null set such that if you take any sample point outside this null set and any pattern, it appears with the right frequency. The null set does not depend on the pattern. This is because, there are only countably many possible patterns and countable union of null sets is null.

There is nothing special about binary expansion. Let $r \geq 2$ be an integer. Then any number $x \in [0, 1]$ can be expressed as $\sum_{n \geq 1} (x_n / r^n)$ where each of the numbers x_n is in the set $\{0, 1, \cdots, r - 1\}$. Further, such an expansion is unique, except for countably many numbers x. In fact, if there are two expansions with (x_n) and (y_n) for a number, then there exists an $n \geq 1$ such that $x_i = y_i$ for $i < n$, $x_n = y_n \pm 1$ (take $r - 1 + 1 = 0$); for $i > n$ one of the sequences consists of all zeros and the other consists of all $r - 1$. The expansion consisting of all zeros after n is called the terminating expansion and the one consisting of all $r - 1$ is called the non-terminating expansion. However, we do not need all this. The digit x_i is called the i-th digit in the expansion of x to the *base* r. Define, for $i \geq 1$, random variable X_i to be the i-th digit of the sample point. It is easy to see that for each i, X_i takes values $0, 1, \cdots, r-1$ each with probability $1/r$. A pattern now is a finite sequence of digits from $\{0, 1, 2, \cdots, r - 1\}$. As earlier we can show the following. Almost every sample point has the following property: any given pattern appears with frequency $1/r^k$ where k is length of the pattern.

We can make a better statement. Almost surely the following happens: Take a number x. Take a base r and an r-pattern. Then this pattern appears in the r-expansion of x with frequency $1/r^k$ where k is length of the pattern. In other words, there is one set $N \subset [0,1]$ with $\lambda(N) = 0$ such that for points $x \notin N$ the above happens. This null set depends neither on the base r nor on the pattern. This is simply because, there are only countably many bases r possible and countable union of null sets is again null.

Numbers which satisfy the property we have been discussing, namely appearance of patterns with given proportion, w.r.t. a fixed base are called *normal* to that base. If they satisfy the condition w.r.t. all bases, then they are called *absolutely normal* or *strongly normal*. Thus almost every number is absolutely normal. All these considerations are due to Emile Borel.

Some numbers we all know very well, like π and e; are they absolutely normal; or at least normal to the base ten? We do not yet know the answer. Is every algebraic number which is not rational, normal to the base ten? This is Borel's question. We do not yet know the answer, though there is some work on the complexity of such numbers. However there are numbers that arise from complexity theory, like Chaitin's constant (it is, roughly speaking, the probability that a universal prefix free computer halts) is known to be absolutely normal. Erdos proved that certain numbers are normal to the base ten. The first known normal number appears to be the Champernowne's number which consists of an enumeration of all the positive integers after the decimal place, that is,

$$0.12345678910111213141516 17 \cdots .$$

Lecture 10/11: Three series theorem and SLLN.

We shall now discuss matters related to the following phenomenon.

You know that the series $1+\frac{1}{2}+\frac{1}{3}+\cdots+\cdots$ diverges. On the other hand the series $1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\cdots\pm\cdots$ converges. What if you tossed a fair coin to decide the sign of each term? Of course, you might say that the answer depends on the outcome of heads and tails. Yes, true, but remember there is a collective behaviour of sequence of independent random variables. We would like to know it.

Kolmogorov's one series theorem: Let (X_n) be a sequence of independent random variables, uniformly bounded, having mean zero and finite variances (σ_n^2). Then the series $\sum X_n$ converges almost surely iff the series of numbers $\sum \sigma_n^2$ converges. Even if the variables are not bounded, convergence of $\sum \sigma_n^2$ implies the convergence of $\sum X_n$ almost surely.

First let us see some consequences. Suppose that (ϵ_i) is a sequence of independent random variables each taking values ± 1 with equal probability. Then $X_n = \epsilon_n/n$ falls under the theorem, leading to the conclusion that the series $\sum X_n$ converges almost surely, because $\sum(1/n^2)$ converges. On the other hand if $Z_n = \epsilon_n/\sqrt{n}$, then the series $\sum Z_n$ converges almost nowhere. Here is another consequence, called, the *Strong Law of Large Numbers*.

Theorem (SLLN): Let (X_n) be a sequence of independent random variables with zero means and finite variances (σ_n^2) such that the series of numbers $\sum \frac{1}{n^2}\sigma_n^2$ converges. Then the averages $\frac{1}{n}\sum_1^n X_i$ converge to zero a.e. In particular, if we have a sequence of i.i.d. random variables with finite mean and variance, then their averages converge to the mean almost surely.

In statistics this has the following interpretation. If $X_1, X_2, \cdots, X_n$ are i.i.d. random variables having the same distribution as that of a random variable X, one says that these X_i are a *sample* of size n from X. The quantity $E(X) = \int X dP$ is called the *population average*. The quantity $\frac{1}{n}\sum_1^n X_i(\omega)$ is called *sample average* because it is average of the observed sample values $X_1(\omega), \cdots, X_n(\omega)$. Thus what the SLLN tells us is that the sample averages converge to population average. This has the consequence that if we did not know the population average, then the sample average is a good estimate.

In physics, this has the following interpretation. There is a phase space Ω which comes equipped with a volume element, P. A particle is moving in the phase space, described by a transformation $\omega \mapsto T(\omega)$. If the particle is at ω now, it will be at $T(\omega)$ at the next minute. There is a real valued function f defined on the space Ω, a measurement on the particle, like velocity or temperature. The quantity $\int f dP$ is called the spatial average of f, simply because we are taking average of f w.r.t. the spatial parameter ω, namely, $\int f(\omega) dP(\omega)$; recall, integration on a probability space is like a weighted average. Let $X_i = f(T^i(\omega))$. The quantity $\frac{1}{n}\sum_1^n X_i(\omega)$ is called time average of X because we are taking averages of the measurements on the particle at successive time points, namely, over time $t = 1$ to $t = n$. Thus what the SLLN tells is that the time averages converge to space average. This was what was done in the ergodic theorem during the lecture of MGN. Of course, in this context, generally, (X_i) are not independent, a very important issue. On the other hand, the present version of SLLN can be deduced from the ergodic theorem.

Here is a proof of SLLN. We start with two observations. Suppose a sequence (x_n) of numbers converges to a number x. Then the averages, $a_n = \frac{1}{n}\sum_1^n x_i$ also converge to x. To see this, no loss to assume $x = 0$; if necessary replace x_i by $x_i - x$. Fix $\epsilon > 0$. Pick n_0 so large that $|x_n| < \epsilon/2$ for $n \geq n_0$. Now that n_0 is fixed, pick $N \geq n_0$ so large that $|\frac{1}{N}\sum_1^{n_0} x_i| \leq \epsilon/2$. If $n \geq N$,

$$\left|\frac{1}{n}\sum_1^n x_i\right| = \frac{1}{n}\left|\sum_1^{n_0} x_i\right| + \frac{1}{n}\sum_{n_0+1}^n |x_i| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Here is another fact we need. Suppose the series of numbers $\sum_{n\geq 1}\frac{1}{n}x_n$ converges. Then the averages $\frac{1}{n}\sum_1^n x_i$ converge to zero. To see this, put $s_n = \sum_{j=1}^n \frac{1}{j}x_j$ for $n \geq 1$, so that $s_n \to c$, say. Then $x_m = m(s_m - s_{m-1})$ for $m \geq 1$, with the understanding $s_0 = 0$.

$$\frac{1}{n}\sum_{m=1}^n x_m = \frac{1}{n}\sum(ms_m - ms_{m-1}) = s_n - \frac{1}{n}\sum_1^n s_m.$$

The first term converges to c by hypothesis and the second term also converges to c by earlier observation. This completes the proof.

Returning to the proof of SLLN, note that X_n/n has mean zero and variance σ_n^2/n^2. The one series theorem implies almost sure convergence of

the series $\sum \frac{1}{n} X_n$. For any sample point ω for which this series converges, we conclude from above discussion that the averages $\frac{1}{n} \sum_1^n X_i(\omega)$ converge to zero.

If the random variables are i.i.d. then $\sigma_n^2 = \sigma^2$, the common variance, for each n and hence the series $\sum(\sigma_n^2/n^2) = \sigma^2 \sum(1/n^2) < \infty$. So the conclusion holds. We only have to apply the previous result to the sequence $\{(X_n - \mu)\}$ where μ is the common mean.

In passing, let us note that the last part of SLLN regarding i.i.d. random variables is about the observed averages converging to the mean. It is natural to ask whether variance need be finite. Indeed it is true without any asuumption on the variance. This is obtained by truncation arguments. We shall now return to the one series theorem. But before proving it, we shall improve it. We shall remove the restriction that the variables be centered, that is, expectation zero.

Kolmogorov's two series theorem: Let (X_n) be a sequence of independent random variables, uniformly bounded and having finite means (μ_n) and finite variances (σ_n^2). Then the series $\sum X_n$ converges almost surely iff the two series of numbers $\sum \mu_n$ and $\sum \sigma_n^2$ converge.

This can be deduced from the one series theorem as follows. Suppose the two series of numbers converge. Suppose that for each n, $|X_n| \leq c$. Then for each n, $|\mu_n| \leq c$. Thus if $Y_n = X_n - \mu_n$, then for each n, $|Y_n| \leq 2c$ and has mean zero. So one series theorem implies that $\sum Y_n$ converges a.e. Since $\sum \mu_n$ converges, we conclude that $\sum X_n$ also converges a.s.

Conversely, suppose that the series $\sum X_n$ converges a.s. Let A be a set with $P(A) = 1$ such that the series $\sum X_n(\omega)$ converges for all points $\omega \in A$. Consider the product space $(\Omega \times \Omega, \mathcal{A} \otimes \mathcal{A}, P \otimes P)$. Define on the product space $Y_n(\omega, \eta) = X_n(\omega)$ and $Z_n(\omega, \eta) = X_n(\eta)$. Using the fact that we have a product space, it is easy to verify that the sequence of random variables $(Y_n - Z_n; n \geq 1)$ are independent. Moreover Y_n and Z_n have the same expectation (integrate and see) so that $Y_n - Z_n$ has mean zero. They are uniformly bounded. Indeed if (X_n) are bounded by c in modulus, then $(Y_n - Z_n)$ are bounded by $2c$. Further the series $\sum(Y_n - Z_n)$ converges for all points in $A \times A$ and $P \otimes P(A \times A) = 1$. Since variance of $Y_n - Z_n$ is $2\sigma_n^2$, the one series theorem applied to the sequence $(Y_n - Z_n)$ implies that the series $\sum 2\sigma_n^2$ converges a.e. But then the same theorem implies that $\sum(X_n - \mu_n)$ converges a.e. This is because these variables are independent, are bounded by $2c$ and

from what was deduced just now, sum of their variances converges. Thus $\sum X_n$ as well as $\sum(X_n - \mu_n)$ converge a.e.; first one by hypothesis, and the second one is shown just now. So both these happen on a set of probability one. But then if you take one ω such that both the series $\sum X_n(\omega)$ and $\sum(X_n(\omega) - \mu_n)$ converge; you immediately deduce that $\sum \mu_n$ converges.

The above theorem, in turn, can be used to answer in full generality, the question of convergence of a series of independent random variables.

Kolmogorov's three series theorem: Let (X_n) be a sequence of independent random variables having means (μ_n) and variances (σ_n^2). If the series $\sum X_n$ converges almost surely then the following three series of numbers converge for any $c > 0$.

$$\sum P(|X_n| > c); \qquad \sum \int_{|X_n| \le c} X_n; \qquad \sum \left[\int_{|X_n| \le c} X_n^2 - \left(\int_{|X_n| \le c} X_n \right)^2 \right].$$

Conversely if the three series of numbers above converge for some $c > 0$, then $\sum X_n$ converges almost surely.

Here is a proof. Suppose first that $c > 0$ is given and that the above three series converge for this value of c. Put $Z_n = X_n$ if $|X_n| \le c$, and put $Z_n = 0$ otherwise. Since Z_n depends on X_n only, we conclude that (Z_n) are independent random variables. Since $\sum P(X_n \ne Z_n) = \sum P(|X_n| > c) < \infty$; the Borel-Cantelli tells us that for almost every sample point ω, we have $X_n(\omega) = Z_n(\omega)$ after some stage. But for such points ω, $\sum X_n(\omega)$ converges iff $\sum Z_n(\omega)$ converges. Thus we only need to show that the series $\sum Z_n$ converges a.e. But these are independent and bounded by c. The two series theorem and hypothesis imply that $\sum Z_n$ converges a.e.

Conversely, assume that the series $\sum X_n$ converges a.e. Fix any number $c > 0$. Define Z_n as in the earlier para. For all those sample points ω for which $\sum X_n(\omega)$ converges, we must have $|X_n(\omega)| \le c$ after some stage. Firstly, this implies limsup of the sequence of events (A_n) defined by $A_n = (X_n \ne Z_n) = (|X_n| > c)$ is null. Thus $\sum P(|X_n| > c) < \infty$. Indeed, if the sum were infinity, independence of the events and Borel-Cantelli show that their limsup must have probability one. Secondly, the series $\sum Z_n$ also converges a.e. But since these are bounded by c, the two series theorem tells that the other two series of this theorem also converge. This completes the proof.

Before proving the one series theorem, we shall first observe *Kolmogorov inequality*. Let $X_1, \cdots, X_n$ be independent mean zero random variables. Denote $S_0 = 0$ and for $1 \leq m \leq n$, $S_m = \sum_1^m X_j$. Then for any number $a > 0$,

$$P\{\max_{1\leq m\leq n} |S_m| \geq a\} \leq \frac{1}{a^2} \sum_1^n \sigma_m^2.$$

This is a special case of Doob's maximal inequality proved by BR for martingales. Here is a quick proof. Let for $1 \leq m \leq n$,

$$A_m = \{|S_i| < a, \text{ for } i < m; |S_m| \geq a\}, \quad A = \cup A_m.$$

Since variance adds up for independent variables,

$$\sum \sigma_m^2 = E(S_n^2) \geq E(S_n^2 I_A) = \sum_k E(S_n^2 I_{A_k}).$$

If $T_k = \sum_{k+1}^n X_k$, then

$$S_n^2 I_{A_k} = (S_k + T_k)^2 I_{A_k} = S_k^2 I_{A_k} + T_k^2 I_{A_k} + 2 S_k T_k I_{A_k}.$$

since $|S_k| \geq a$ on A_k, expectation of the first term is at least $a^2 P(A_k)$. Expectation of the second term is non-negative. Since $E(T_k) = 0$ and S_k, A_k depend only on $X_i, i \leq k$, we see that $E(S_k T_k I_{A_k}) = E(T_k) E(S_k I_{A_k}) = 0$. Thus

$$\sum \sigma_m^2 \geq \sum_k a^2 P(A_k) = a^2 P(A).$$

This completes proof of the inequality.

Returning to the proof of one series theorem, first assume that the series $\sum \sigma_n^2$ converges. We shall show that the partisl sums $\{S_n(\omega)\}$ form a Cauchy sequence for a.e. sample point ω. Recall that a sequence (x_n) of numbers is Cauchy if for every $\epsilon > 0$ there is an N such that $|x_n - x_m| \leq \epsilon$ for all $n, m \geq N$. This is same as saying that for all $\epsilon > 0$ there is an N such that $|x_N - x_n| \leq \epsilon$ for all $n \geq N$. If you choose N to satisfy the second sentence for $\epsilon/2$, then it satisfies the first sentence for ϵ. In fact, it is enough to do this not for all $\epsilon > 0$ but for $\epsilon = 1/2, 1/2^2, \cdots$.

First observe the following, for any $\epsilon > 0$

$$P(\sup_{n\geq N} |S_n - S_N| > \epsilon\} \leq \frac{1}{\epsilon^2} \sum_{N+1}^{\infty} \sigma_m^2. \qquad (*)$$

This is because if you denote

$$A_k = \{\max(|S_{N+1} - S_N|, \cdots, |S_{N+k} - S_N|) > \epsilon\},$$

then Kolmogorov's inequality says

$$P(A_k) \leq \sum_{N+1}^{N+k} \sigma_i^2 \leq \sum_{n>N} \sigma_i^2.$$

Since the events A_k increase to the event described in $(*)$, that inequality follows. Since the right side of that inequality is the tail sum of a convergent series, we can choose N so that right side is as small as we please. Thus for every integer $k \geq 1$, we can use $(*)$ with $\epsilon = 1/2^k$ and choose N_k so that the right side of $(*)$ is smaller than $1/2^k$. In other words, we can get increasing sequence of integers $N_1, N_2, \cdots, N_k, \cdots$ such that for each $k \geq 1$

$$P\left(\sup_{n \geq N_k} |S_n - S_{N_k}| > \frac{1}{2^k}\right) \leq \frac{1}{2^k} \qquad (*)$$

Thus if $A_k = \{\sup_{n \geq N_k} |S_n - S_{N_k}| > \frac{1}{2^k}\}$, then $\sum P(A_k)$ is finite. If we denote $\limsup A_k$ by A then Borel-Cantelli tells $P(A) = 0$. We now complete the proof by arguing that if $\omega \notin A$, then the sequenc $\{S_n(\omega)\}$ is a Cauchy sequence. Since $\omega \notin A$, fix m such that $\omega \notin A_k$ for all $k \geq m$. But then this means, if you give $\epsilon = 1/2^k$ where $k \geq m$, then we will have N_k to verify the Cauchy criterion mentioned at the beginning of the proof.

Now we shall prove the converse. So assume that $\sum X_n$ converges, that is, $\{S_n\}$ converges a.e., say to S, a random variable. We show that the series $\sum \sigma_n^2$ converges. Assume that the variables X_n are bounded by c. We first argue that there is a number d such that the set $A = \{|S_n| \leq d \text{ for all } n\}$ has positive probability. Indeed, since $(|S| < d_1) \uparrow \Omega$ as $d_1 \uparrow \infty$, first get d_1 such that this set has probability larger than 15/16. Next, $S_n \to S$ tells us that the sets $(|S_n| < d_1 \text{ for all } n \geq N)$ increases to a set as $N \uparrow \infty$ which includes the previous set. Thus get N such that $(|S_n| < d_1 \quad \text{for all} \quad n \geq N)$ has probability larger than 15/16. Since $(|S_n| < d_2 \text{ for all } n \leq N) \uparrow \Omega$, get d_2 such that probability of this set is larger than 15/16. If we take $d = \max(d_1, d_2)$ we have what we wanted. Set now $A_n = \{|S_m| \leq d \text{ for all } m \leq n\}$ and $A = \cap A_n$. Thus $A_n \downarrow A$ and $P(A) > 0$.

We shall now obtain bounds for $E(S_n^2 I_{A_n})$ leading to an inequality for the partial sums of $\sum \sigma_n^2$. On one hand, using the fact $|S_n| \leq d$ on A_n, we have

$$E(S_k^2 I_{A_k}) \leq d^2 P(A_k) \leq d^2. \qquad (**)$$

$$E(S_n^2 I_{A_n}) - E(S_{n-1}^2 I_{A_{n-1}}) = E(S_n^2 I_{A_{n-1}}) - E(S_n^2 I_{A_{n-1}-A_n}) - E(S_{n-1}^2 I_{A_{n-1}})$$
$$= E(X_n^2 I_{A_{n-1}}) + 2E(X_n S_{n-1} I_{A_{n-1}}) - E(S_n^2 I_{A_{n-1}-A_n}).$$

Here for the first equality, we expressed A_n as the difference of A_{n-1} and $A_{n-1} - A_n$. For the second equality, we expressed $S_n = S_{n-1} + X_n$.

Since A_{n-1} depends on $X_i, i \leq n-1$ we conclude the first term equals $P(A_{n-1})\sigma_n^2 \geq P(A)\sigma_n^2$.

Since $S_{n-1} I_{A_{n-1}}$ depends on $X_i, i \leq n-1$ and $E(X_n) = 0$, the second term is zero.

On $A_{n-1} - A_n$ we have $|S_n| = |S_{n-1} + X_n| \leq c + d$ so that

$$-E(S_n^2 I_{A_{n-1}-A_n}) \geq -(c+d)^2 P(A_{n-1} - A_n).$$

Thus

$$E(S_n^2 I_{A_n}) - E(S_{n-1}^2 I_{A_{n-1}}) \geq P(A)\sigma_n^2 - (c+d)^2 P(A_{n-1} - A_n).$$

Adding this over $n = 1, \cdots, k$ we get

$$E(S_k^2 I_{A_k}) \geq P(A) \sum_1^k \sigma_n^2 - (c+d)^2 \qquad (***).$$

From $(**)$ and $(***)$, we get

$$d^2 \geq P(A) \sum_1^k \sigma_n^2 - (c+d)^2 \text{ or } \sum_1^k \sigma_n^2 \leq \frac{d^2 + (c+d)^2}{P(A)}.$$

This inequality, which is true for all k shows that the series $\sum \sigma_n^2$ converges. This completes proof of the theorem.

We have followed parts of W. Rudin 'Real and Complex Analysis', P. R. Halmos 'Measure Theory' in this exposition. Best source is the book by S R S Varadhan 'Probability Theory'. He has another book 'Stochastic Processes'. Both are in the Courant-AMS lecture notes series.

Lecture 12: Brownian Heuristics

In the symmetric random walk, you start at zero and move, every unit of time, as follows. Toss a fair coin; Heads up move one step forward, Tails up, move one step backward. What if we move 'continuously'? What is the meaning to be assigned to this continuous motion? One possibility is to move every half unit/quarter unit and so on with smaller and smaller jump sizes and see if there is a limit in any sense. Let us first discuss a simpler problem.

Suppose you put one rupee in a bank. Suppose the interest is calculated yearly and the rate is r Rs. per rupee per year. Thus if the interest rate, in the customary sense, is 6% per annum, then $r = 0.06$. Thus at the end of the year you get $(1 + r)$ Rs. Suppose the interest is calculated half-yearly and the rate is $r/2$ Rs. per rupee per half year. Then after a half year you have Rs. $(1 + \frac{r}{2})$ and at the end of the year you get Rs. $(1 + \frac{r}{2})^2$. More generally, if the interest is calculated $(1/n)$-yearly and the rate is r/n Rs. per Rupee per $(1/n)$-year, then at the end of the year you get $(1 + \frac{r}{n})^n$ Rs. If the bank says it updates continuously, then what should it mean? A reasonable meaning is that you take limit of these numbers, which, luckily exists giving Rs. e^r at the end of one year. As you would notice, we did not say that the interest is Rs, r per year and hence it is r/n per $1/n$-year. We outright said it is r/n Rs per $1/n$-year. The subtle point is that the bank has the option of announcing, a possibly different interest rate, at the beginning of each segment. However, we shall not continue on this aspect further.

With continuous updating, can the bank afford to announce rate of Rs. $1/\sqrt{n}$ per rupee per $1/n$ year? No. because then at the end of the year they should pay you Rs $\lim(1 + \frac{1}{\sqrt{n}})^n$ which is ∞! Indeed

$$\left(1 + \frac{1}{\sqrt{n}}\right)^n \geq 1 + n.\frac{1}{\sqrt{n}} > \sqrt{n} \to \infty.$$

On the other hand, suppose the bank announces a rate of Rs. $1/n^2$ per rupee per $1/n$-year, will you accept? No. Because then you will receive no interest at all; $\lim(1 + \frac{1}{n^2})^n = 1$. Indeed, given any $\epsilon > 0$, for all sufficiently large n, we have $\frac{1}{n} < \epsilon$ so that

$$\limsup \left(1 + \frac{1}{n^2}\right)^n \leq \limsup \left(1 + \frac{\epsilon}{n}\right)^n = e^\epsilon.$$

This being true for every $\epsilon > 0$ we conclude that the lim sup is at most one. But of course all these quantities are larger than one. Thus the limit exists

and equals one.

Thus there must be a match between the time units and the rates. The square root is unimportant, bank can not afford to have any power of $1/n$ which is smaller than one. Similarly any power of $1/n$ larger than one is not acceptable to you. in other words, when such a thing happens the limit we are seeking turns out to be either zero or infinity. Similarly, in the random walk case we can interpret continuous motion as limit of discrete motions. What is the correct amount of jump if we choose to move every $1/n$ unit of time? The central limit theorem proved by JM gives the answer: it should be of the order of $1/\sqrt{n}$. We shall now describe more precisely.

To simplify matters, let us not look at the entire infinite time axis. Let us continue the random walk upto time one, starting at time zero. To describe various of these motions in a uniform way, let us also fix once and for all a sequence of i.i.d. random variables $\eta_1, \eta_2, \cdots$ ecah taking values ± 1 with equal probability, namely, $1/2$. Thus these have mean zero and variance one. Let us denote their partial sums by $S_n = \sum_1^n \eta_i$ with $S_0 = 0$. Thus S_n has mean zero and variance n. At time zero we are always at $S_0 = 0$. If we play at intervals of one time unit then we have only one play, namely, at time one we are at $S_1 = \eta_1$. If we play at every $1/2$ time unit, our state at time $1/2$ is $S_1/\sqrt{2}$ and at time one it is $S_2/\sqrt{2}$. More generally, if we play at each $1/n$ time unit with jump size $\pm 1/\sqrt{n}$, then our state at time k/n is $S_k/\sqrt{n}$ for $k = 0, 1, \cdots, n$.

It is awkward to have the n-th motion defined at a certain discrete set of points depending on n. To rectify this and to make matters more visually appealing, let us define the n-th process $X_n(t)$ for $0 \leq t \leq 1$ as follows. As earlier, for $t = k/n$ the random variable is $S_k/\sqrt{n}$; $k = 0, 1, \cdots, n$ and for t between k/n and $(k+1)/n$ the value of $X_n(t)$ is obtained by joining the values at these two points by a straight line. More precisely, though not important for us,

$$X_n(t) = (k + 1 - nt)\frac{S_k}{\sqrt{n}} + (nt - k)\frac{S_{k+1}}{\sqrt{n}} \quad \text{for } \frac{k}{n} \leq t \leq \frac{k+1}{n}.$$

Thus for each integer $n \geq 1$, we have a process $X_n(t)$ defined for $0 \leq t \leq 1$. There is, of course, another possibility for defining the n-th process. Do not move continuously from $S_k/\sqrt{n}$ to $S_{k+1}/\sqrt{n}$ as time increases from k/n to $(k+1)/n$. Just stay put at $S_k/\sqrt{n}$ during the time $(k/n) \leq t < (k+1)/n$ and at time $(k+1)/n$ move to $S_{k+1}/\sqrt{n}$. The paths are no longer continuous, but

are made up of flat parts. We shall not discuss this possibility. In any case there is only a difference of $1/\sqrt{n}$ between the earlier path and the present one at any time instant and one can argue that both have the same properties in the limit. Returning to our original processes $\{X_n(t) : 0 \leq t \leq 1\}$, we raise the question: are these converging to something in some sense?

You notice immediately, using CLT, that $X_n(1) = S_n/\sqrt{n}$ converges to standard normal in the sense of distribution, or in the sense of weak convergence. So we are advised to consider weak convergence. This is what convergence means in what follows..

Let us fix a time point $0 < t < 1$ then the value of the n-th walk at this time is $S_{[nt]}/\sqrt{n}$ plus an error ϵ_n of magnitude at most $1/\sqrt{n}$. Here $[x]$ is as usual, the largest integer not exceeding x. But as $n \to \infty$ we know $[nt] \to \infty$ and hence by CLT we conclude that, $S_{[nt]}/\sqrt{[nt]} \to N(0,1)$, standard normal variable in the sense described above. As a consequence

$$X_n(t) = \frac{\sqrt{[nt]}}{\sqrt{n}} \cdot \frac{S_{[nt]}}{\sqrt{[nt]}} + \epsilon_n \to N(0,t).$$

Here we have used some results. For instance the term ϵ_n, the small random error, can be ignored since it goes to zero a.e. Regarding the first expression on the right side, the first factor is deterministic and converges to $\sqrt{t}$ and the second factor converges to standard normal. So the product converges to centered normal with variance t. These are consequences of what is known as Slutsky's Theorem.

Let us now take two time points $s < t$. Let us consider the pair of random variables $(X_n(s), X_n(t) - X_n(s))$. The first is sum of the first $[ns]$ many η_i and the other is sum of the next $[nt]-[ns]$ many η_i. Of course, all this is upto an error of $2/\sqrt{n}$. More important is the fact that the η_i that participate in these sums have disjoint indices. Thus the limiting distribution of this pair $(X_n(s), X_n(t) - X_n(s))$ is bivariate normal with means zero; variances s and $t - s$; covariance zero. A slight extension of the one dimensional CLT proved by JM is needed here.

This argument can be generalized to any finite number of time points. If $0 < t_1 < \cdots < t_k$ are time points, then the distribution of $(X_n(t_1), X_n(t_2) - X_n(t_1) \cdots X_n(t_k) - X_n(t_{k-1}))$ converges to the multivariate normal with mean vector zero, variances $t_1, t_2 - t_1, \cdots, t_k - t_{k-1}$ respectively and other covariances zero. In other words these increments, in the limit, are independent.

Thus if there is a limiting process $(B_t)_{t\geq 0}$ then we have (i) $B_0 \equiv 0$, (ii) for any finitely many time points $0 \leq t_1 < \cdots < t_k \leq 1$, the variables $B_{t_1}, B_{t_2} - B_{t_1}, \cdots, B_{t_k} - B_{t_{k-1}}$ are independent, (iii) for any $0 \leq s < t \leq 1$, $B_t - B_s \sim N(0, t-s)$. Finally (iv) the process has continuous paths, that is, for any sample point ω, the trajectory $t \mapsto B_t(\omega)$ is a continuous function on $[0, \infty)$. Such a process is called Standard Brownian Motion. Of course condition (iv), the continuity of paths, is not explained by the above argument; simply because we were only taking limits in the weak sense. However, since we are trying to describe the motion of a particle, it is natural to postulate that the particle moves continuously.

How does one show that such a process exists? There are several ways. Here are two ways. One constructs a probability on the uncountable product space $R^{[0,1]}$ with the coordinate process (X_t) having the required finite dimensional distributions. Then restricts the probability to the set of sample points ω for which the map $t \mapsto X_t(\omega)$ is continuous. Of course, this last set is not measurable, so one needs to do with care.

Here is another way. Let us go back to the n-th walk. For each sample point ω we associated a continuous function,with value at k/n being $S_k(\omega)/\sqrt{n}$ for $k \geq 0$ and in between joined by straight line. In other words, if we denote by $C = C[0,1]$ the collection of real valued continuous functions on $[0,1]$, then we have a map T_n from our probability space to the space C. The sample sdpace for the n-th walk has 2^n points, corresponding to the two values of each of the η_i, $1 \leq i \leq n$. Thus the probability on C induced by T_n is concentrated on a set of 2^n points in C. Let us denote this probability by μ_n. The space C has a natural metric and hence the space of probabilities has a natural notion of convergence. Suppose that (P_n) is a sequence of probabilities on C and P is a probability on C. To imitate what JM did on R, say $P_n \Rightarrow P$ if $\int f dP_n \to \int f dP$ for every bounded real continuous function f on C. Then one shows that the special sequence (μ_n) constructed above does indeed converges to some probability μ on C. If you consider the space C with this limiting probability μ, then the the coordinate process on C $--$ defined as $B_t(f) = f(t)$ for $0 \leq t \leq 1$ $--$ satisfies all the conditions stated above. In other words, this is Brownian Motion.

Lecture 13: elementary probability Models

Today we are just going to distribute balls into boxes following prescribed rules, no more than that. Just as in music you can weave intricate musical patterns with a few basic notes, so is it in mathematics. With the help of a few basic ingredients we can weave intricate mathematical patterns.

1.

We are given strictly positive integers b, r. To start with, we have an urn containing b black and r red balls. Here is the scheme: Select a ball at random, note its colour, put it back and add one ball of that colour to the urn. Repeat.

We begin evaluating probabilities of some simple events. Let R_n be the event that the nth ball drawn is red and B_n be the event that the nth ball drawn is black. Clearly $P(R_1) = r/(r+b)$. $P(R_2) = P(R_2R_1) + P(R_2B_1) = P(R_1)P(R_2|R_1) + P(B_1)P(R_2|B_1)$ which equals $r/(r+b)$ after simplification. Using induction on n (condition on the first draw) we can easily show that

$$P(R_n) = \frac{r}{b+r} \quad \text{and} \quad P(B_n) = \frac{b}{b+r}$$

We further observe the following regarding conditional probabilities:

$$P(R_2\|R_1) = \frac{r+1}{r+b+1} > \frac{r}{r+b} = P(R_1)$$

$$P(R_2\|B_1) = \frac{r}{r+b+1} < \frac{r}{r+b} = P(R_1)$$

Let us see what this means. The unconditional chances of seeing a red ball at any stage remains the same, namely, $r/(r+b)$. However if the first ball seen was red, then the conditional chances of seeing a red ball at the second stage is $(r+1)/(r+b+1)$ which is larger than $r/(r+b)$. If the second ball seen was also red, then the conditional chances of seeing a red ball at the third draw are further increased. On the other hand, if the first ball seen was black then the conditional chances of seeing a red ball at the second stage is reduced to $r/(r+b+1)$ from $r/(r+b)$. This is very interesting and reflects a phenomenon that we observe in the context of contagious diseases.

Imagine that you are going to Trivandrum. Before you leave, a friend of yours tells you 'be careful, there is flu in Tr.'. Anyway you do reach Tr., a little worried. If now you meet a person with flu what would you think? Oh,

I have already met a person with flu, so there is indeed flu here. If you meet another person with flu what would you think? My God, I have already met two persons with flu, so flu is really rampant here. On the other hand, if you first met a person without flu what would be your thoughts? Well, I met this person who did not have flu, perhaps there is not that much flu afterall. And so on. The point is that, depending on *your* experiences you revise your opinion. But keep in mind, your revision of opinion has no influence on the realities. The reality remains the same. Do not get confused by thinking of new infections accruing etc., we are talking about the situation in a very small time interval after you reached Tr.. This is precisely what our model is telling us.

One can do a good amount of mathematics with this model, but we shall not. This scheme was proposed by G.Polya and F.Eggenberger in 1923 and is known as Polya Urn Model.

2.

Consider the same urn as earlier with r red balls and b black balls. Consider the same scheme but with the following change. Instead of adding one ball of the colour seen, we add one ball of the opposite colour. Of course, here the chance of red ball at the nth draw changes with time and is not constant. We can calculate several probabilities in this model too.

This was proposed as a model for safety campaign. If you interpret red balls as accidents and black balls as safety measures, what we are doing reflects the following. In practice, if you see an increase in the number of accidents, then you tend to increase the safety measures. As the safety measures take over and the accidents decrease, we tend to be lax and we see an increase in the number of accidents. This was invented by Bernard Friedman in 1949 and is known as Friedman Urn model.

3.

We have k Urns. The i-th Urn contains r_i red balls and b_i black balls for $1 \leq i \leq k$, all these numbers are strictly positive integers. We pick one of the urns, say Urn i is chosen with probability p_i, and draw balls from the selected Urn, with replacement.

The chances of red ball in any draw remains the same, namely, $\sum p_i \alpha_i$ where $\alpha_i = \frac{r_i}{r_i + b_i}$. Note that we pick one Urn to start with and balls are drawn from that Urn. What are the conditional chances that the second ball is red given first is red? Following earlier notation, $P(R_1\ R_2) = \sum p_i \alpha_i^2$, so

that

$$P(R_2|R_1) = (\sum p_i\alpha_i^2)/(\sum p_i\alpha_i) \geq \sum p_i\alpha_i = P(R_1) = p(R_2),$$

where the inequality is from Cauchy-Schwarz.

In other words the chances of red ball at second draw increase if the first ball was red. This is rather curious. In the earlier two models there were after-effects. For example, in the Polya Urn if we see red ball, there was an after-effect, namely, the chances of red ball in the next draw is increased. In the Friedman Urn, if we see a red ball, there was an after-effect, namely, the chances of red ball in the next draw is decreased. However, in the present model there is no apparent after-effect, we are drawing balls from the same urn with replacement. Yet the conditional chances of second ball being red are larger (than the unconditional chances) if the first ball was red.

Feller invented this (perhaps around 1950) to illustrate that sampling effect should not be confused with contagion. Each person (or profession) is liable to accidents and this occurrence is determined by random drawings from an Urn. We can imagine each person (or profession) having his own Urn. We assume that there is no after-effect so that the composition of the Urn remains unchanged throughout the process. Chance of an accident or proneness to accidents may vary from person to person. The phenomenon observed is an effect of sampling rather than contagion.

4.

This is a static experiment, in the sense, there is no dynamics involved. We only want to distribute the balls and calculate certain probabilities.

We have two Urns and r balls. We want to distribute the balls into the urns. We take a coin with chance of heads p in each toss. We take a ball, toss the coin and place the ball in the Urn 1 or Urn 2 according as we get Heads or Tails. Do this for each ball. We want to know how many balls are in urn 1. More precisely, let p_k be the probability that there are k balls in Urn 1. Clearly, this is same as the chances of getting k heads in r tosses and hence $p_k = \binom{r}{k}p^k(1-p)^{r-k}$.

We shall now modify the experiment. I have three coins with chances of heads p_1, p_2, p_3 respectively. I choose one of the coins at random and do as earlier with the selected coin. What is p_k now? If you denote the earlier value by $f(p)$, then in the modified experiment we have $p_k = [f(p_1)+f(p_2)+f(p_3)]/3$. This is because, the event (k heads) is the unon of disjoint subevents ($k$heads

and coin i is chosen); $1 \leq i \leq 3$. The probability of the i-th event equals $P(\text{coin k is chosen})P(k \text{ Heads}|\text{coin } k)$.

We shall now modify the experiment further. Suppose I have one coin for each number $0 < p < 1$, such that the chance of heads for that coin is p. I pick a coin at random and do as earlier with the selected coin. What is p_k now? Earler we had expressed the event (k Heads) as disjoint union of three subevents. Now it will be union of uncountably many subevents. It would still be average, but instead of sum of terms divided by the total number, we need to take integral of the function divided by the length of the interval. This can all be justified invoking rigorous development of conditional probabilities; we do not pause to do it now. Thus $p_k = \int_0^1 f(p)dp = 1/(r+1)$; integrate by parts. Thus all values $k = 0, 1 \cdots, r+1$ are equally likely.

The same answer can be obtained in a different way. assume that the balls are all looking alike. How many possible arrangements can our eye distinguish? Since the balls look alike, it makes no sense to ask which ball went into urn one. You can distinguish two arrangements only by their occupancy numbers. Thus there are exactly $r+1$ arrangements. What we saw above is that these arrangements are equally likely. So instead of saying that the distribution of the balls is achieved by picking a coin at random, we could have said that the balls all look alike and the distinguishable arrangements are equally likely. This was how Satyen Bose proposed in 1924.

Well, imagine for a moment that the balls are photons and urns are different energy levels. We are not distributing balls, but the photons are organizing themselves into the available energy levels. How do they place themselves? The answer is as follows. All possible occupancies are equally likely. This is a discovery of the famous physicist Satyendranath Bose. He was explaining Planck's formula for the distribution of energy in the radiation of a black body. Though this law is quantum mechanical in nature, all its derivations at that time were based on classical physics, an unhappy situation. Bose proposed this postulate in 1924. The interpretation in terms of random coin is due to the statistician Sudhakar Kunte in 1977. Incidentally, we can do the same experiment when there are n urns (or energy levels) instead of two. We can assume that the balls (photons) all look alike and the distinguishable arrangements are equally likely. In this case there are $\binom{n+r-1}{n-1}$ distinguishable arrangements and thus they are equally likely. We can use a random n faced die to distribute. That is consider the set of

all n faced dice and pick one at random (in an appropriate sense) and use it to distribute each ball into the n boxes. Both lead to the same answer.

The reason why the proposal was revolutionary is the following. If I had an urn with two balls; one red and one green and I pick a ball at random; you agree that the chances of a red ball is 1/2. Suppose that the urn has 1 red ball and 999 green balls and I pick a ball at random. I can not persuade you to accept the incorrect statement that the chances of red ball is 1/2 by simply arguing that the green balls all look alike and there are only two distinguishable outcomes 'red ball' and 'green ball'.

5.

Now we consider only one urn. We have an unlimited supply of balls with us. We have a coin whose chance of heads in a single toss is p and we have a number $\lambda > 0$.

Here is the scheme. Suppose we see j balls in the Urn this morning. Take a ball, toss the coin, Heads up remove the ball, Tails up keep the ball. We do this for each ball in the urn. Since the chance of keeping a ball is $(1-p)$, the chances that we keep i balls is $\binom{j}{i}(1-p)^i p^{j-i}$. Independently, we shall add r balls with probability $e^{-\lambda}\lambda^r/r!$, where r could take any non negative integer value. These are the Poisson probabilities, which as we all know, arise as appropriate limits of binomial probabilities. Thus the chances that we add zero balls equals $e^{-\lambda}$; the chances that we add 29 balls is $e^{-\lambda}\lambda^{29}/(29!)$ and so on. Even though, there is no *a priori* limit on the number of balls that we add, please do keep in mind that we add only a finite number of balls. Thus the number of balls in the urn on any day is a finite number, perhaps zero, but never infinity.

What happens in the long run? To make this question precise, denote by $p_n(k)$ the probability of having k balls in the urn on day n. Of course, this probability depends also on how we started the whole game. But this dependence is not shown in the notation. Problem is to see if the limit, $\lim_n p_n(k)$ exists and to find when it exists.

Let us start with the simplest initial condition, namely, on day 0, there are no balls in the urn. This does not make the problem any uninteresting. The balls in the urn on day 1 are just the balls we have put into the urn. Thus, by the specification of the mechanism it is clear that

$$p_1(k) = e^{-\lambda}\lambda^k/k!.$$

If A is the event that there are k balls on day two, then $p_2(k) = P(A)$. Let A_j, for $j \geq 0$, be the event that 'there are k balls on day two and j balls on day one'. Clearly, these events are disjoint and their union is A. Thus

$$P(A) = \sum_{j \geq 0} P(A_j).$$

Now

$$P(A_j) = P(j \text{ balls on day one}) \cdot P(k \text{ balls on day two} \mid j \text{ balls on day one}).$$

The first quantity on the right side is just $e^{-\lambda}\lambda^j/j!$ as seen above. To calculate the second quantity we proceed as follows. Given that there are j balls, the only way to end up with k balls is to keep i balls out of the j and then add $k-i$ balls. Of course i can not be larger than k (you can not keep more than what you have) and it can not be larger than j either (you can not keep more than what you want to endup with). Since the events of removing balls and adding balls are independent, the second term on right side is

$$\sum_{i=0}^{j \wedge k} \binom{j}{i} q^i p^{j-i}\ e^{-\lambda}\lambda^{k-i}/(k-i)!,$$

where we denoted $1-p$ by q. Thus,

$$P(A) = \sum_{j \geq 0} e^{-\lambda}\ \lambda^j\ \frac{1}{j!}\ \ \sum_{i=0}^{j \wedge k} \binom{j}{i}\ q^i\ p^{j-i}\ e^{-\lambda}\ \lambda^{k-i}\ \frac{1}{(k-i)!}.$$

A careful interchange of the order of summation shows,

$$p_2(k) = e^{-[\lambda(1+q)]}\ \frac{[\lambda(1+q)]^k}{k!}.$$

You can now easily guess the value of $p_n(k)$ and prove that your guess is right. For $n \geq 0$,

$$p_{n+1}(k) = e^{-[\lambda(1+q+\cdots+q^n)]}\ \frac{[\lambda(1+q+\cdots+q^n)]^k}{k!}.$$

Thus the limit we are looking for, exists and indeed,

$$\lim_n p_n(k) \quad = \quad e^{-\lambda/p}\ [\lambda/p]^k\ \frac{1}{k!}.$$

Of course, all this calculation is done with the condition that there were no balls in the urn initially, that is, to start with the urn is empty. Actually the

answer does not depend on the initial condition. What ever be the initial number of balls in the urn, the limit of $p_n(k)$ exists and is the same as above.

Actually we do not have balls, instead, we have a solution containing Brownian particles (under diffusive equilibrium, but ignore this phrase). We do not have urn, instead, we have a geometrically well defined volume V in the solution. During each time period, certain particles emerge from this volume V and certain particles will have entered this volume V. Time period is not one day, but is of the order of few hundredths of a second. This is a model of the famous Astrophysicist Subrahmanyan Chandrasekhar proposed in 1943. Chandrasekhar was explaining the inner relationships that exist between the phenomena of Brownian motion, diffusion, and fluctuations in molecular concentrations.

6.

We conclude with a brief description of another illuminating Urn model, possibly the first Urn model. Consider an urn divided into two compartments and they together have $2N$ balls, numbered $1, 2, \cdots, 2N$. Choose one of these numbers at random, change the ball of that number from its compartment to the other. Repeat. What can be expected in the long run? By state of the system let us mean the number of balls in compartment I. Thus there are $2N+1$ possible states, namely, 0,1,$\cdots$, $2N$. The following can be shown: no matter what state k is prescribed, $0 \leq k \leq 2N$, and no matter how much time elapsed you are sure to see k balls in compartment I at some future time. Thus, the state keeps on changing and we are sure to visit each state infinitely many times. In a sense the state of the system appears to move in a *chaotic* way.

One can also show (in a sense that can be made precise) the following. In the long run, the state of the system is k with probability $\binom{2N}{k} \big/ 2^{2N}$ – as if each of the $2N$ balls are placed in the compartments at random. In particular, a *steady state* is approached by the ensemble. Did some one say that there must be an error because the motion can not be *chaotic* and yet approach a *steady state*? Well, there is no contradiction and it is precisely to prove this point the physicists Paul and Tatiana Ehrenfests invented this model in 1907. This is called the Ehrenfests model of heat exchange. Let us also note, in passing, that in the steady state, the expected number of balls in each compartment is N, expected value of binomial variable with parameters $2N$ and $1/2$.

According to the Boltzman's formalism of Thermodynamics, one explains heat as a reflection of molecular motion (speed) and exchange of heat as reflection of molecular collisions and the resulting change of momentum. If you mix hot milk and cold water, then over a period of time a steady state is reached, as we observe in practice (of course, in practice, the environment also plays a role). Let us discuss a closed system, insulated from the environment etc. (isolated system, to use a technical phrase). The explanation is that milk being hot, those molecules are moving with higher speeds. On the other hand, water being cold, those molecules are moving slower. When you bring them together the molecules collide with each other exchanging speeds. But this is not an orderly process. It is not as simple as saying that the milk molecules, on impinging with water molecules, pass on speed, as if it is a one way process. For example, it could very well happen that during the process a water molecule might have gained speed after colliding with a fast moving milk molecule. Later it may collide with one milk molecule which lost its speed earlier and exchange takes place.

Thus the two compartments can be thought of as milk and water. Temperature is signified by the number of balls. If a ball is moved from the water compartment to the milk compartment, it signifies that a fast moving water molecule collided with a slow moving milk molecule increasing the speed of the milk molecule. Thermodynamics provides a time arrow, it is in the direction of increasing entropy. Ultimately the system reaches its maximum entropy state.

However in a closed system, under certain conditions, every state is visited infinitely often. This is Poincare recurrence theorem presented in the course by MGN. In the context of Markov chains it takes the following form: that the chain is recurrent, in the sense IKR explained. Obviously, there were serious objections to the thermodynamic formalism of Boltzman. Apart from rejecting the molecular nature, one serious objection was that the phenomenon of recurrence and reaching a steady state appeared contradictory. It is in this context the present model was put forward by the Ehrenfests. Since the number of molecules ($2N$) is very very large and the recurrence time is reciprocal of the steady state probability, the recurrence times are very very long. Thus once the process is nearly in the state of having *approximately* N balls in each compartment, it takes enormously long time to reach the state where 2 balls are in one compartment and the remaining $2N-2$ in the other compartment (several billions of years). Thus the process *appears* irreversible.

Lecture 14: Loose ends

We shall discuss proofs of a few theorems which were used in the course but were not proved.

Holder's inequality:
Let $(\Omega, \mathcal{A}, P)$ be a probability space. Fix $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. For example $p = q = 2$ or $p = 3; q = 3/2$. In fact if you fix any $p > 1$, then you can define q by the formula $\frac{1}{q} = 1 - \frac{1}{p}$. Here then is the inequality. For any two nonnegative random variables X and Y;

$$E(XY) \leq (EX^p)^{1/p}(EY^q)^{1/q}.$$

Using the notation, $||X||_p = \{E(|X|^p)\}^{1/p}$, this inequality takes the form $||XY||_1 \leq ||X||_p||Y||_q$. In the special case $p = q = 2$, this says $E(XY) \leq \sqrt{E(X^2)}\sqrt{E(Y^2)}$. This is known as Cauchy-Schwarz inequality. Proof of the inequality is simple. First note that for any two numbers $a, b > 0$, we have $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$. Indeed, concavity of log function tells that

$$\log\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right) \geq \frac{1}{p}\log a^p + \frac{1}{q}\log b^q = \log(ab).$$

Of course the above inequality trivially holds even if one of a and b is zero.

First assume that $E(X^p) = 1 = E(Y^q)$. Using the earlier observation, $X(\omega)Y(\omega) \leq \frac{1}{p}X^p(\omega) + \frac{1}{q}Y^q(\omega)$ and taking expectations, we get

$$E(XY) \leq \frac{1}{p}EX^p + \frac{1}{q}EY^q = \frac{1}{p} + \frac{1}{q} = 1 = ||X||_p||Y||_q.$$

If $E(X^p) = a^p > 0$ and $E(Y^q) = b^q > 0$; apply the inequality proved above to $X_1 = X/a$; $Y_1 = Y/b$ and simplify to get the desired inequality for X, Y . Finally, assume that one of the expectations is zero, say, $E(X^p) = 0$. Then the integrand being non-negative, $X^p = 0$ a.e., that is, $X = 0$ a.e. and hence so is XY. the required inequality follows. Note that in case, one of $||X||_p$ and $||Y||_q$ are infinite, then the inequality holds obviously. This inequality leads to the following.

Minkowski inequality: For any two random variables X and Y,

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

No loss to assume that both the quantities on the right side are finite. Use Holder with $|X+Y|^{p-1}$ and $|X|$ to get

$$E\left(|X+Y|^{p-1}|X|\right) \leq (E|X|^p)^{1/p}(E|X+Y|^{(p-1)q})^{1/q} = ||X||_p(E|X+Y|^p)^{1/q}.$$

Similarly

$$E\left(|X+Y|^{p-1}|Y|\right) \leq (E|Y|^p)^{1/p}(E|X+Y|^{(p-1)q})^{1/q} = ||Y||_p(E|X+Y|^p)^{1/q}.$$

Adding these two rquations we get,

$$E|X+Y|^p \leq E\{|X+Y|^{p-1}(|X|+|Y|)\} \leq (||X||_p + ||Y||_p)(E|X+Y|^p)^{1/q},$$

leading to the required inequality. This inequality will, for example show, that $d(X,Y) = ||X-Y||_p$ is indeed a metric on the space L_p. In fact with this metric, the space L_p is a complete metric space.

Jensen's inequality:
Let $(\Omega, \mathcal{A}, P)$ be a probability space and X an integrable random variable. Let φ be a convex function on an open interval containing the range of X, then $\varphi(EX) \leq E(\varphi oX)$. Here φoX is the random variable whose value at ω is $\varphi(X(\omega))$.

Note that the inequality makes sense. Since $X > a$ we conclude that $E(X) > a$. Similarly $E(X) < b$. Thus we can evaluate φ at $E(X)$. Regarding the right side, either you can assume that φoX is integrable and proceed, or, deduce from the proof that the integral exists, may be $+\infty$.

We need to understand convex functions. First let us observe that the function f defined on $[3,4]$ by $f(x) = 0$ for $3 < x < 4$; $f(3) = 1$; $f(4) = 2$ is a convex function. This is clearly not continuous. We show that a convex function defined on an open interval must be continuous. So let (a,b) be an open interval, finite or infinite, and φ be a convex function on this interval. If $x_1 < x_2 < x_3$ are points in the interval, then x_2 is a convex combination of x_1 and x_3. More precisely

$$x_2 = \frac{x_3 - x_2}{x_3 - x_1}x_1 \; + \; \frac{x_2 - x_1}{x_3 - x_1}x_3,$$

so that we have

$$\varphi(x_2) \leq \frac{x_3 - x_2}{x_3 - x_1}\varphi(x_1) \; + \; \frac{x_2 - x_1}{x_3 - x_1}\varphi(x_3).$$

A slight rearrangement will give us

$$\frac{\varphi(x_2) - \varphi(x_1)}{x_2 - x_1} \leq \frac{\varphi(x_3) - \varphi(x_2)}{x_3 - x_2}.$$

If we take four points $x_1 < x_2 < x_3 < x_4 \in (a, b)$, then applying the above inequality to the points $\{x_1, x_2, x_3\}$ and $\{x_2, x_3, x_4\}$, we get

$$\frac{\varphi(x_2) - \varphi(x_1)}{x_2 - x_1} \leq \frac{\varphi(x_4) - \varphi(x_3)}{x_4 - x_3}.$$

This inequality has two interesting consequences. First it shows continuity of φ. Indeed, if $z \in (a, b)$, then fix small $c > 0$ such that $[z - c, z + c] \subset (a, b)$. Fix two points u, v such that $z + c \leq u < v < b$. If we denote $M = [\varphi(v) - \varphi(u)]/(v - u)$ then for any two points $x, y \in (z - c, z + c)$, the above inequality shows that $|\varphi(y) - \varphi(x)| \leq M|y - x|$. Since this holds for any two points in the interval $(z - c, z + c)$, we conclude that the function φ is continuous on the interval $(z - c, z + c)$ and hence continuous at z. This being true for every z, we conclude that φ is continuous on (a, b).

Second use of the above inequality is to show that there is a supporting tangent at every point on the graph of φ. More precisely, let $z \in (a, b)$. Then there is a straightline $L(x) = mx + c$ such that for every x, $L(x) \leq \varphi(x)$ and $L(z) = \varphi(z)$. Incidentally, when we say for every point x, we only mean points in the interval (a, b). To see this, observe that the above inequality leads to the following

$$\sup\left\{\frac{\varphi(y) - \varphi(x)}{y - x} : x < y \leq z\right\} \leq \inf\left\{\frac{\varphi(v) - \varphi(u)}{v - u} : z \leq u < v\right\}.$$

Let m be any number in between these two quantities. Of course, if both sides are equal, there is no choice, you take the common value as m. In particular, if $x < z$ then

$$\varphi(z) - \varphi(x) \leq m(z - x); \quad \text{that is,} \quad \varphi(x) \geq mx + [\varphi(z) - mz].$$

If $x > z$ then

$$\varphi(x) - \varphi(z) \geq m(x - z); \quad \text{that is,} \quad \varphi(x) \geq mx + [\varphi(z) - mz].$$

Thus if, $L(x) = mx + [\varphi(z) - mz]$, then for all $x \neq z$ we have $\varphi(x) \geq L(x)$. Clearly, $L(z) = \varphi(z)$.

This has an interesting consequence. There are countably many straight-line functions, $L_n(x) = m_n x + c_n$, $n = 1, 2, \cdots$ such that $\varphi(x) = \sup_{n \geq 1} L_n(x)$. This is because, for each rational $z \in (a, b)$, get a line as above. Since there are countably many rationals, enumerate the lines so obtained as a sequence

(L_n). Put $\psi(x) = \sup_{n\geq 1} L_n(x)$. Note that ψ is convex. Indeed, for any u, v and $0 \leq \lambda \leq 1$,

$$L_n(\lambda u + [1-\lambda]v) = \lambda L_n(u) + [1-\lambda]L_n(v) \leq \lambda\psi(u) + [1-\lambda]\psi(v).$$

Here the equality is because of the affine nature of L_n and the inequality is because of the definition of ψ. Now take sup over n to see the convexity of ψ. So from what has been discussed above, ψ is continuous on (a, b). But by construction of the L_n functions, we see that φ and ψ agree at all rational points and hence they must agree everywhere in (a, b) completing the proof.

Now let us return to Jensen. Take convex φ and get L_n as above. Then for any n, $L_n(EX) = E(L_n o X) \leq E(\varphi o X)$. Here the equality is from linearity of expectation and the inequality is from monotonicity of expectation. Now take sup over n to get the desired inequality. The same argument, keeping track of null sets shows Jensen's inequality for conditional expectation.

Monotone Class theorem:

Several times in the course, the monotone class theorem has been used. Since we did not prove it so far, let us do it now. First recall the theorem. Let Ω be a set and $\mathcal{F}$ be a field of subsets of Ω. Let $\sigma(\mathcal{F})$ be the smallest σ-field which includes sets in $\mathcal{F}$. Let $\mathcal{M}(\mathcal{F})$ be the smallest monotone class which includes sets in $\mathcal{F}$. The $\sigma(\mathcal{F}) = \mathcal{M}(\mathcal{F})$.

Here is a proof. A σ-field is a monotone class — in particular $\sigma(\mathcal{F})$ is a monotone class. Since $\mathcal{M}(\mathcal{F})$ is the smallest monotone class that includes $\mathcal{F}$ we conclude that $\mathcal{M}(\mathcal{F}) \subset \sigma(\mathcal{F})$, that is, every set in the collection $\mathcal{M}(\mathcal{F})$ is in the collection $\sigma(\mathcal{F})$. It remains to show that $\mathcal{M}(\mathcal{F}) \supset \sigma(\mathcal{F})$.

We now show that $\mathcal{M}(\mathcal{F})$ is a field. What does this achieve? Since a field which is a monotone class is a σ-field, this shows that $\mathcal{M}(\mathcal{F})$ is indeed a σ-field. It then follows that $\mathcal{M}(\mathcal{F})$ is a σ-field that includes the class $\mathcal{F}$, so must include the smallest such, namely, $\sigma(\mathcal{F})$.

First, we show that whenever $A \in \mathcal{M}(\mathcal{F})$ then so is A^c. For this, consider the class $\mathcal{M}_0 = \{A \in \mathcal{M}(\mathcal{F}) : A^c \in \mathcal{M}(\mathcal{F})\}$. If $A \in \mathcal{F}$, then $\mathcal{F}$ being a field we have $A^c \in \mathcal{F}$ which is included in $\mathcal{M}(\mathcal{F})$, it follows that $\mathcal{F} \subset \mathcal{M}_0$. If each A_n is in $\mathcal{M}_0$ and $A_n \uparrow A$, then firstly A_n as well as A_n^c are in $\mathcal{M}(\mathcal{F})$. Secondly, $A_n \uparrow A$ as well as $A_n^c \downarrow A^c$ so that both A and A^c are in $\mathcal{M}(\mathcal{F})$. Thus $A \in \mathcal{M}_0$. Similarly, if each A_n is in $\mathcal{M}_0$ and $A_n \downarrow A$ then $A \in \mathcal{M}_0$. In other words $\mathcal{M}_0$ is a monotone class that includes $\mathcal{F}$. But $\mathcal{M}(\mathcal{F})$ is the

smallest such. This means that every set in $\mathcal{M}(\mathcal{F})$ must be in $\mathcal{M}_0$. Thus, whenever $A \in \mathcal{M}(\mathcal{F})$ so is A^c.

Second, we show that if A and B are in $\mathcal{M}(\mathcal{F})$ then so is their intersection $A \cap B$. To start with, observe that if both the sets A and B are in $\mathcal{F}$ then their intersection is already in $\mathcal{F}$ and hence in $\mathcal{M}(\mathcal{F})$. Fix $A \in \mathcal{F}$. Consider $\mathcal{M}_0 = \{B \in \mathcal{M}(\mathcal{F}) : A \cap B \in \mathcal{M}(\mathcal{F})\}$. Beacuse of the earlier sentence, this family includes $\mathcal{F}$. As in the earlier paragraph, $\mathcal{M}_0$ is a monotone class and hence must be $\mathcal{M}(\mathcal{F})$. In other words, for every $B \in \mathcal{M}(\mathcal{F})$ we have $A \cap B \in \mathcal{M}(\mathcal{F})$. Now fix any $B \in \mathcal{M}(\mathcal{F})$ and consider $\mathcal{M}_0 = \{A \in \mathcal{M}(\mathcal{F}) : A \cap B \in \mathcal{M}(\mathcal{F})\}$. From what we proved just now, this family includes $\mathcal{F}$. This is again a monotone class and hence must be $\mathcal{M}(\mathcal{F})$. Thus intersection of two sets in $\mathcal{M}(\mathcal{F})$ is again in $\mathcal{M}(\mathcal{F})$.
This completes the proof.

Independence:

Several times during the course, we used arguments like the following. X_1, X_2 are independent; Z_1 depends only on X_1 and Z_2 depends only on X_2; hence Z_1 and Z_2 are also independent. Here the word 'depends' is used in the sense of English word, whereas the word 'independent' is used in our technical sense. We need to understand this.

Recall that, while discussing the zero-one law we have proved the following. Random variables $X_1, X_2, \cdots, X_n$ are independent iff $\mathcal{A}_1, \mathcal{A}_2, \cdots \mathcal{A}_n$ are independent; where $\mathcal{A}_i = \sigma(X_i)$. In such a case it is obvious that if we have random variables $Z_1, Z_2, \cdots Z_n$ such that $\sigma(Z_i) \subset \sigma(X_i)$, then $Z_1, Z_2, \cdots, Z_n$ are also independent. Thus in the cases where we said Z_i depends on X_i only, this was what was happening and hence the statement of independence holds. In fact $\sigma(Z) \subset \sigma(X)$ can be taken as the meaning of the statement 'Z depends only on X.

Let $(\Omega, \mathcal{A}, P) = (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, P_1 \otimes P_2)$ be a product space. Let us define $\mathcal{B}_1 = \{A_1 \times \Omega_2 : A_1 \in \mathcal{A}_1\}$; $\mathcal{B}_2 = \{\Omega_1 \times A_2 : A_2 \in \mathcal{A}_2\}$. Then the definition of product probability tells us that $\mathcal{B}_1 \subset \mathcal{A}$ and $\mathcal{B}_2 \subset \mathcal{A}$ are independent. In particular, suppose we have independent σ-fields $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_n$ all contained in $\mathcal{A}_1$ (here independence is under P_1) and independent σ-fields $\mathcal{D}_1, \mathcal{D}_2, \cdots \mathcal{D}_m$ all contained in $\mathcal{A}_2$ (here independence is under P_2). Put $\mathcal{G}_i = \{A_1 \times \Omega_2 : A_1 \in \mathcal{C}_i\}; 1 \leq i \leq n$; $\mathcal{H}_j = \{\Omega_1 \times A_2 : A_2 \in \mathcal{D}_j\}; 1 \leq j \leq m$. Then the family $\{\mathcal{G}_1, \cdots, \mathcal{G}_n, \mathcal{H}_1, \cdots \mathcal{H}_m\}$ is independent (here independence is under $P = P_1 \otimes P_2$). Such considerations were needed in the proof of the two series theorem.

We shall conclude with the proof of one more simple fact that was used.

For independent random variables $(X_i : 1 \leq i \leq k)$, variance adds up, that is, $var(\sum X_i) = \sum var(X_i)$. A simple algebra shows that this will follow if we can show that for two independent integrable random variable X and Y; $E(XY) = E(X)E(Y)$. This can be seen in several ways. We denote $\mathcal{A} = \sigma(X)$ and $\mathcal{B} = \sigma(Y)$. Then independence of these σ-fields — which is a consequence of independence of X and Y — will yield that for every $A \in \mathcal{A}$ and $B \in \mathcal{B}$ we have $E(I_A I_B) = E(I_A)E(I_B)$. Hence for simple random variables $\mathcal{A}$-measurable s and $\mathcal{B}$-measurable t, we have $E(st) = E(s)E(t)$. By monotone convergence theorem similar statement holds for non-negative random variables. Finally, for any integrable random variables $\mathcal{A}$-measurable U and $\mathcal{B}$-measurable V, we have $E(UV) = E(U)E(V)$. This does it.

Another way of looking at it is the following. Define the map $\omega \mapsto (X(\omega), Y(\omega))$ from Ω to $R \times R = R^2$. This is a measurable map. Here $R \times R$ is equipped with the product σ-field $\mathcal{B} \otimes \mathcal{B}$ where $\mathcal{B}$ is the Borel σ-field of R. Let μ be the induced probability on R^2. Then independence of X and Y translates to the fact $\mu = \mu_1 \otimes \mu_2$, product probability. Here μ_1 and μ_2 are respectively the distributions of X and Y. By the change of variable formula, integrating XY on the original probability space is same as integrating the product of coordinate functions on R^2 w.r.t. μ. Fubini will complete the proof.