

[SQUEAKING]

PROFESSOR: Last time, we started discussing graph limits. And let me remind you some of the notions and definitions that were involved.

One of the main objects in graph limits is that of a graphon, which are symmetric, measurable functions from the unit squared to the unit interval. So here, symmetric means that $w(x, y) = w(y, x)$.

We define a notion of convergence for a sequence of graphons. And remember, the notion of convergence is that a sequence is convergent if the sequence of homomorphism densities converges as n goes to infinity for every fixed F , every fixed graph.

So this is how we define convergence. So a sequence of graphs or graphons, they converge if all the homomorphism densities-- so you should think of this as subgraph statistics-- if all of these statistics converge. We also say that a sequence converges to a particular limit if these homomorphism densities converge to the corresponding homomorphism density of the limit for every F .

OK. So this is how we define convergence. We also define this notion of a distance. And to do that, we first define the cut norm to be the following quantity defined by taking two subsets, S and T , which are measurable. Everything so far is going to be measurable. And look at what is the maximum possible deviation of the integral of this function on this box, $S \times T$.

And here, w , you should think of it as taking real values, allowing both positive and negative values, because otherwise, you should just take S and T to be the whole interval. OK. And this definition was motivated by our discussion of discrepancy coming from quasi randomness.

Now, if I give you two graphs or graphons and ask you to compare them, you are allowed to permute the vertices in some sense, so to find the best overlay. And that notion is captured in the definition of cut distance, which is defined to be the following quantity, where we consider over all possible measure-preserving bijections from the interval to itself of the difference between these two graphons if I rotate one of them using this measure-preserving bijection. So think of this as permuting the vertices.

So these were the definitions that were involved last time. And at the end of last lecture, I stated three main theorems of graph limit theory. So I forgot to mention what are some of the histories of this theory.

So there were a number of important papers that developed this very idea of graph limits, which is actually somewhat-- if you think about all of combinatorics, we like to deal with discrete objects. And even the idea of taking a limit is rather novel. So this work is due to a number of people.

In particular, Laszlo Lovasz played a very important central role in the development of this theory. And various people came to this theory from different perspectives-- some from more pure perspectives, and some from more applied perspectives. And this theory is now getting used in more and more places, including statistics, machine learning, and so on. And I'll explain where that comes up just a little bit.

At the end of last lecture, I stated three main theorems. And what I want to do today is develop some tools so that we can prove those theorems in the next lecture. OK. So I want to develop some tools.

In particular, you'll see some of the things that we've talked about in the chapter on Szemerédi's regularity lemma come up again in a slightly different language. So much of what I will say today hopefully should already be familiar to you, but you will see it again from the perspective of graph limits.

But first, before telling you about the tools, I want to give you some more examples. So one of the ways that I motivated graph limits last time is this example of an Erdos-Renyi random graph or a sequence of quasi-random graphs converging to a constant. The constant graphon is the limit.

But what about generalizations? What about generalizations of that construction when your limit is not the constant? So this leads to this idea of a w random graph, which generalizes that of an Erdos-Renyi random graph.

So in Erdos-Renyi, we're looking at every edge occurring with the same probability, p , uniform throughout the graph. But what I want to do now is allow you to change the edge probability somewhat. OK.

So before giving you the more general definition, a special case of this is an important model

of random graphs known as the stochastic block model. And in particular, a two-block model consists of the following data where I am looking at two types of vertices-- let's call them red and blue-- where the vertices are assigned to colors at random-- for example, 50/50. But any other probability is fine.

And now I put down the edges according to which colors the two endpoints are. So two red vertices are joined with edge probability P_{rr} . If I have a red and a blue, then I may have a different probability joining them, and likewise with blue-blue, like that. So in other words, I can encode this probability information in the matrix, like that. So it's symmetric across the diagonal.

So this is a slightly more general version of an Erdos-Renyi random graph where now I have potentially different types of vertices. And you can imagine these kinds of models are very important in applied mathematics for modeling certain situations such as, for example, if you have people with different political party affiliations. How likely are they to talk to each other?

So you can imagine some of these numbers might be bigger than others. And there's an important statistical problem. If I give you a graph, can you cluster or classify the vertices according to their types if I do not show you in advance what the colors are but show you what the output graph is? So these are important statistical questions with lots of applications.

This is an example of if you have only two blocks. But of course, you can have more than two blocks. And the graphon context tells us that we should not limit ourselves to just blocks. If I give you any graphon w , I can also construct a random graph.

So what I would like to do is to consider the following construction where-- OK, so let's just call it w random graph denoted by g and w -- where I form the graph using the following process. First, the vertex set is labeled by 1 through n . And let me draw the vertex types by taking uniform random x_1 through x_n -- OK, so uniform iid.

So you think of them as the vertex colors, the vertex types. And I put an edge between i and j with probability exactly w of x_i, x_j , so for all i less than j independently. That's the definition of a w random graph.

And the two-block stochastic model is a special case of this w random graph for the graphon, which corresponds to this red-blue picture here. So the generation process would be I give you some x_1, x_2, x_3 , and then, likewise, x_1, x_3, x_2 . And then I evaluate, what is the value of

this graphon at these points?

And those are my edge probabilities. So what I described is a special case of this general w random graph. Any questions?

So like before, an important statistical question is if I show you the graph, can you tell me a good model for where this graph came from? So that's one of the reasons why people in applied math might care about these types of constructions.

Let me talk about some theorems. I've told you that the sequence of Erdos-Renyi random graphs converges to the constant graphon p . So instead of taking a constant graphon p , now I start with w random graph. And you should expect, and it is indeed true, that this sequence converges to w as their limit.

So let w be a graphon. So let w be a graphon. And for each n , let me draw this graph $G_{sub n}$ using the w random graph model independently. Then with probability 1, the sequence converges to the graphon w .

So in the sense that I've shown above, described above. So this statement tells us a couple of things-- one, that w random graphs converge to the limit w , as you should expect; and two, that every graphon w is the limit point of some sequence of graphs. So this is something that we never quite explicitly stated before. So let me make this remark.

So in particular, every w is the limit of some sequence of graphs, just like every real number, in analogy to what we said last time. Every real number is the limit of a sequence of rational numbers through rational approximation. And this is some form of approximation of a graphon by a sequence of graphs.

OK. So I'm not going to prove this theorem. The proof is not difficult. So using that definition of subgraph convergence, the proof uses what's known as Azuma's inequality. So by an appropriate application of Azuma's inequality on the concentration of martingales, one can prove this theorem here by estimating the probability that-- to show that the probability that the F density in G_n , it is very close to the F density in w with high probability.

OK. Any questions so far? So this is an important example of one of the motivations of graph limits. But now, let's get back to what I said earlier. I would like to develop a sequence of tools that will allow us to prove the main theorem stated at the end of the last lecture.

And this will sound very familiar, because we're going to write down some lemmas that we did back in the chapter of Szemerédi's regularity lemma but now in the language of graphons. So the first is a counting lemma.

The goal of the counting lemma is to show that if you have two graphons which are close to each other in the sense of cut distance, then their F densities are similar to each other. So here's a statement. So if w and u are graphons and F is a graph, then the F density of w minus the F density of u , their difference is no more than a constant-- so number of edges of F times the cut distance between u and w .

So maybe some of you already see how to do this from our discussion on Szemerédi's regularity lemma. In any case, I want to just rewrite the proof again in the language of graphons. And this will hopefully-- so we did two proofs of the triangle counting lemma.

One was hopefully more intuitive for you, which is you pick a typical vertex that has lots of neighbors on both sides and therefore lots of edges between. And then there was a second proof, which I said was a more analytic proof, where you took out one edge at a time. And that proof, I think it's technically easier to implement, especially for general H .

But the first time you see it, you might not quite see what the calculation was about. So I want to do this exact same calculation again in the language of graphons. And hopefully, it should be clear this time.

So this is the same as the counting lemma over epsilon-regular pairs. So it suffices to prove the inequality where the right-hand side is replaced not by the cut distance but by the cut norm. And the reason is that once you have the second inequality by taking an infimum over all measure-preserving bijections ϕ -- and notice that that change does not affect the F density. By taking an infimum over ϕ , you recover the first inequality.

I want to give you a small reformulation of the cut norm that will be useful for thinking about this counting lemma. Here's a reformulation of the cut norm-- namely, that I can define the cut norm. So here, w is taking real values, so not necessarily non-negative.

So the cut norm we saw earlier is defined to be the supremum over all measurable subsets of the $[0, 1]$ interval of this integral in absolute value. But it turns out I can rewrite this supremum over a slightly larger set of objects. Instead of just looking over measurable subsets of the interval, let me now look at measurable functions.

Little u . So OK, let me look at functions. So u and v from $0, 1$ to $0, 1$ -- and as always, everything is measurable-- of the following integral.

So I claim this is true. So I consider this integral. Instead of integrating over a box, now I'm integrating this expression.

OK. So why is this true? Well, one of the directions is easy to see, because the right-hand side is strictly an enlargement of the left-hand side. So by taking u to be the indicator function of S and v to be the indicator of function of T , you see that the right-hand side, in fact, includes the left-hand side in terms of what you are allowed to do.

But what about the other direction? So for the other direction, the main thing is to notice that the integral or the integrand, what's inside this integral, is bilinear in the values of u and v . So in particular, the extrema of this integral, as you allow to vary u and v , they are obtained. So they are obtained for u and v , taking values in the endpoints 0 , comma, 1 .

It may be helpful to think about the discrete setting, when, instead of this integral, you have a matrix and two vectors multiplied from left and right. And you had to decide, what are the coordinates of those vectors? It's a bilinear form. How do you maximize it or minimize it? You have to change every entry to one of its two endpoints. Otherwise, it can never be-- you never lose by doing that.

OK, so think about it. So this is not difficult once you see it the right way. But now, we have this cut norm expressed over not sets, but over bounded functions. And now I'm ready to prove the counting lemma.

And instead of writing down the whole proof for general H , let me write down the calculation that illustrates this proof for triangles. And the general proof is the same once you understand how this argument works. And the argument works by considering the difference between these two F densities. And what I want to do is-- so this is some integral, right? So this is this integral, which I'll write out.

So we would like to show that this quantity here is small if u and w are close in cut norm. So let's write this integral as a telescoping sum where the first term is obtained by-- so by this, I mean w of x , comma, y minus u of x , comma, y .

And then the second term of the telescoping sum-- so you see what happens. I change one

factor at a time. And finally, I change the third factor.

So this is the identity. If you expand out all of these differences, you see that everything intermediate cancels out. So it's a telescoping sum.

But now I want to show that each term is small. So how can I show that each term is small? Look at this expression here. I claim that for a fixed value of z -- so imagine fixing z . And let x and y vary in this integral.

It has the form up there, right? If you fix z , then you have this u and v coming from these two factors. And they are both bounded between 0 and 1. So for a fixed value of z , this is at most w minus u -- the cut norm difference between w and u in absolute value.

So if I left z vary, it is still bounded in absolute value by that quantity. So therefore each is bounded by w minus u cut norm in absolute value. Add all three of them together. We find that the whole thing is bounded in absolute value by 3 times the cut normal difference.

OK, and that finishes the proof of the counting lemma. For triangles, of course, if you have general H , then you just have more terms. You have a longer telescoping sum, and you have this bound.

OK. So this is a counting lemma. And I claim that it's exactly the same proof as the second proof of the counting lemma that we did when we discussed Szemerédi's regularity lemma and this counting lemma. Any questions? Yeah.

AUDIENCE: Why did it suffice to prove over the [INAUDIBLE]?

PROFESSOR: OK. So let me answer that in a second. So first, this should be H , not F . OK, so your question was, up there, why was it sufficient to prove this version instead of that version? Is that the question?

AUDIENCE: Yeah.

PROFESSOR: OK. Suppose I prove it for this version. So I know this is true. Now I take infimum of both sides. So now I consider infimum of both sides.

So then this is true, right? Because it's true for every ϕ . But the left-hand side doesn't change, because the F density in a relabeling of the vertices, it's still the same quantity, whereas this one here is now that.

All right. So what we see as a corollary of this counting lemma is that if you are a Cauchy sequence with respect to the cut distance, then the sequence is automatically convergent. So recall the definition of convergence. Convergence has to do with F densities converging. And if you have a Cauchy sequence, then the F densities converge.

And also, a related but different statement is that if you have a sequence w_n that converges to w in cut distance, then it implies that w_n converges to w in the sense as defined for F densities. So qualitatively, what the counting lemma says is that the cut norm is stronger than the notion of convergence coming from subgraph densities.

So this is one part of this regularity method, so the counting lemma. Of course, the other part is the regularity lemma itself. So that's the next thing we'll do. And it turns out that we actually don't need the full strength of the regularity lemma. We only need something called a weak regularity lemma.

What the weak regularity lemma says is-- I mean, you still have a partition of the vertices. So let me now state it for graphons. So for a partition p -- so I have a partition of the vertex set-- and a symmetric, measurable function w -- I'm just going to omit the word "measurable" from now on. Everything will be measurable.

What I can do is, OK, all of these assets are also measurable. I can define what's known as a stepping operator that sends w to this object, $w_{\text{sub } p}$, obtained by averaging over the steps s_i cross s_j and replacing that graphon by its average over each step.

Precisely, so I obtain a new graphon, a new symmetric, measurable function, $w_{\text{sub } p}$, where the value on x, y is defined to be the following quantity-- if x, y lies in s_i cross s_j . So pictorially, what happens is that you look at your graphon. There's a partition of the vertex set, so to speak, the interval. Doesn't have to be a partition into intervals, but for illustration, suppose it looks like that.

And what I do is I take this w , and I replace it by a new graphon, a new symmetric, measurable function, $w_{\text{sub } p}$, obtained by averaging. Take each box. Replace it by its average. Put that average into the box. So this is what $w_{\text{sub } p}$ is supposed to be.

Just a few minor technicalities. If this denominator is equal to 0, let's ignore the set. I mean, then you have a zero measure set, anyway, so we ignore that set. So everything will be treated up to measure zero, changing the function on measure zero sets. So it doesn't really

matter if you're not strictly allowed to do this division.

OK. So this operator plays an important role in the regularity lemma, because it's how we think about partitioning, what happens to a graph under partitioning. It has several other names if you look at it from slightly different perspectives.

So you can view it as a projection in the sense of Hilbert space. So in the Hilbert space of functions on the unit square, the stepping operator is a projection unto the subspace of constants, subspace of functions that are constant on each step. So that's one interpretation.

Another interpretation is that this operation is also a conditional expectation. If you know what a conditional expectation actually is in the sense of probability theory, so then that's what happens here. If you view $[0, 1]$ squared as a probability space, then what we're doing is we're doing conditional expectation relative to the sigma algebra generated by these steps.

So these are just a couple of ways of thinking about what's going on. They might be somewhat helpful later on if you're familiar with these notions. But if you're not, don't worry about it.

Concretely, it's what happens up there.

OK. So now let me state the weak regularity lemma. So the weak regularity lemma is attributed to Frieze and Kannan, although their work predates the language of graphons. So it's stated in the language of graphs, but it's the same proof. So let me state it for you both in terms of graphons and in graphs.

What it says is that for every ϵ and every graphon w , there exists a partition denoted p of the $[0, 1]$ interval. And now I tell you how many sets there are. So it's a partition into-- so not a tower-type number of parts, but only roughly an exponential number of parts-- 4 to the $1/\epsilon^2$ measurable sets such that if we apply the stepping operator to this graphon, we obtain an approximation of the graphon in the cut norm.

So that's the statement of the weak regularity lemma. There exists a partition such that if you do this stepping, then you obtain an approximation. So I want you to think about what this has to do with the usual version of Szemerédi's regularity lemma that you've seen earlier. So hopefully, you should realize, morally, they're about the same types of statements. But more importantly, how are they different from each other?

And now let me state a version for graphs, which is similar but not exactly the same as what

we just saw for graphons. So let me state it. So for graphs, the weak regularity lemma says that, OK, so for graphs, let me define a partition p of the vertex set is called weakly ϵ -regular if the following is true.

If it is the case that whenever I look at two vertex subsets, A and B , of the vertex set of G , then the number of edges between A and B is what you should expect based on the density information that comes out of this partition. Namely, if I sum over all the parts of the partition, look at how many vertices from A lie in the corresponding parts. And then multiply by the edge density between these parts. So that's your predicted value based on the data that comes out of the partition.

So I claim that this is the actual number of edges. This is the predicted number of edges. And those two numbers should be similar to each other but at most ϵn^2 , where n is the number of vertices. So this is the definition of what it means for a partition to be weakly ϵ -regular.

So it's important to think about why this is weaker. It's called weak, right? So why is it weaker than a notion of ϵ -regularity? So why is it weaker?

So previously, we had ϵ -regular partition in the definition of Szemerédi's regularity lemma, this ϵ -regular partition. And here, notion of weakly ϵ -regular. So why is this a lot weaker?

It is not saying that individual pairs of parts are ϵ -regular. And eventually, we're going to have this number of parts. So I'll state a theorem in a second. So the sizes of the parts are much smaller than ϵn .

But what this weak notion of regularity says, if you look at it globally-- so not looking at specific parts, but looking at it globally-- then this partition is a good approximation of what's going on in the actual graph, whereas-- OK, so it's worth thinking about. It's really worth thinking about what's the difference between this weak notion and the usual notion.

But first, let me state this regularity lemma. So the weak regularity lemma for graphs says that for every ϵ and every graph G , there exists a weakly ϵ -regular partition of the vertex set of G into at most $\frac{1}{\epsilon^2}$ parts.

Now, you might wonder why did Frieze and Kannan come up with this notion of regularity. It's a weaker result if you don't care about the bounds, because an ϵ -regular partition will be automatically weakly ϵ -regular. So maybe with small changes of ϵ if you wish, but

basically, this is a weaker notion compared to what we had before.

But of course, the advantage is that you have a much more reasonable number of parts. It's not a tower. It's just a single exponential.

And this is important. And their motivation was a computer science and algorithm application. So I want to take a brief detour and mention why you might care about weakly epsilon-regular partitions.

In particular, the problem that is of interest is in approximating something called a max cut. So the max cut problem asks you to determine-- given a graph G , find the maximum over all subsets of vertices, the maximum number of vertices between a set and its complement. That's called a cut. I give you a graph, and I want to know-- find this s so that it can have as many edges across this set as possible.

This is an important problem in computer science, extremely important problem. And the status of this problem is that it is known to be difficult to get it even within 1%. So the best algorithm is due to Goemans and Williamson.

It's an important algorithm that was one of the foundational algorithms in semidefinite programming, so related-- the words "semidefinite programming" came up earlier in this course when we discussed growth index inequality. So they came up with an approximation algorithm.

So here, I'm only talking about polynomial time, so efficient algorithms. Approximation algorithm with approximation ratio around 0.878. So one can obtain a cut that is within basically 13% of the maximum. So it's an approximation algorithm. However, it is known that it is hard in the sense of complexity theory. It'd be hard to approximate beyond the ratio $\frac{16}{17}$, which is around 0.491.

And there is an important conjecture in computer science called a unique games conjecture that, if that conjecture were true, then it would be difficult. It would be hard to approximate beyond the Goemans-Williamson ratio. So this indicates the status of this problem. It is difficult to do an epsilon approximation.

But if the graph I give you is dense-- "dense" meaning a quadratic number of edges, where n is a number of vertices-- then it turns out that the regularity-type algorithms-- so that theorem

combined with the algorithmic versions allows you to get polynomial time approximation algorithms. So this is polynomial time approximation schemes.

So one can approximate up to $1 - \epsilon$ ratio. So one can approximate up to ϵn^2 additive error in polynomial time. So in particular, if I'm willing to lose $0.01 n^2$, then there is an algorithm to approximate the size of the max cut.

And that algorithm basically comes from-- without giving you any details whatsoever, the algorithm essentially comes from first finding a regularity partition. So the partition breaks the set of vertices into some number of pieces. And now I search over all possible ratios to divide each piece.

So there is a bounded number of parts. Each one of those, I decide, do I cut this up half-half? Do I cut it up $1/3$, $2/3$, and so on? And those numbers alone, because of this definition of weakly ϵ -regular, once you know what the intersection of A , B is, let's say, a complement is with individual sets, then I basically know the number of edges.

So I can approximate the size of the max cut using a weakly ϵ -regular partition. So that was the motivation for these weakly ϵ partitions, at least the algorithmic application. OK. Any questions?

OK. So let's take a quick break. And then afterwards, I want to show you the proof of the weak regularity lemma.

All right. So let me start the proof of the weak regularity lemma. And the proof is by this energy increment argument. So let's see what this energy increment argument looks like in the language of graphons.

So energy now means L^2 , so L^2 energy increment. So the statement of this lemma is that if you have w , a graphon, and p , a partition, of $[0, 1]$ interval such that-- always measurable pieces. I'm not going to even write it. It's always measurable pieces-- such that the difference between w and w averaged over steps p is bigger than ϵ .

So this is the notion of being not ϵ -regular in the weak sense, not weakly ϵ -regular. Then there exists a refinement, p' prime of p , dividing each part of p into at most four parts such that the true norm increases by more than ϵ^2 under this refinement. So it should be similar. It should be familiar to you, because we have similar arguments from Szemerédi's regularity lemma. So let's see the proof.

Because you have violation of weak epsilon regularity, there exists sets S and T , measurable subsets of $[0, 1]$ interval, such that this integral evaluated over $S \times T$ is more than epsilon in absolute value. So now let me take p' to be the common refinement of p by introducing S and T into this partition. So throw S and T in and break everything according to S and T .

And so each part becomes at most four subparts. So that's the at most four subparts. I now need to show that I have an energy increment. And to do this, let me first perform the following calculation.

So remember, this symbol here is the inner product obtained by multiplying and integrating over the entire box. I claim that that inner product equals to the inner product between w_p and $w_{p'}$, because what happens here is we are looking at a situation where $w_{p'}$ is constant on each part.

So when I do this inner product, I can replace w by its average. And likewise, over here, I can also replace it by its average. And you end up having the same average. And these two averages are both just what happens if you do stepping by p .

You also have that w has inner product with $1_{S \times T}$ the same as that of $w_{p'}$ by the same reason, because over $S \times T$. So $S \times T$ is a union of the parts of p' . So S is union of parts of p' .

OK. So let's see. With those observations, you find that-- so this is true. This is from the first equality.

So now let me draw you a right triangle. So you have a right angle, because you have an inner product that is 0. So by Pythagorean theorem, so what is this hypotenuse? So you add these two vectors. And you find out this $w_{p'}$. So by Pythagorean theorem, you find that the L^2 norm of $w_{p'}$ equals to the L^2 norm of the sum of the L^2 norm squares of the two legs of this right triangle.

On the other hand, this quantity here. So let's think about that quantity over there. It's an L^2 norm. So in particular, it is at least this quantity here, which you can derive in one of many ways-- for example, by Cauchy-Schwarz inequality or go from L^2 to L^1 and then pass down to L^1 . So this is true. So let's say by Cauchy-Schwarz.

But this quantity here, we said was bigger than epsilon. So as a result, this final quantity, this

L2 norm of the new refinement, increases from the previous one by more than ϵ squared. OK.

So this is the L2 energy increment argument. I claim it's the same argument, basically, as the one that we did for Szemerédi's regularity lemma. And I encourage you to go back and compare them to see why they're the same.

All right, moving on. So the other part of regularity lemma is to iterate this approach. So if you have something which is not ϵ regular, refine it. And then iterate. And you cannot perceive more than a bounded number of times, because energy is always bounded between 0 and 1.

So for every ϵ bigger than 0 and graphon w , suppose you have P_0 , a partition of $[0, 1]$ interval into measurable sets. Then there exists a partition p that cuts up each part of P_0 into at most $4/\epsilon$ parts such that $w - w|_p$ is at most ϵ . So I'm basically restating the weak regularity lemma over there but with a small difference, which will become useful later on when we prove compactness.

Namely, I'm allowed to start with any partition. Instead of starting with a trivial partition, I can start with any partition. This was also true when we were talking about Szemerédi's regularity lemma, although I didn't stress that point. That's certainly the case here.

I mean, the proof is exactly the same with or without this extra. This extra P_0 really plays an insignificant role. What happens, as in the proof of Szemerédi's regularity lemma, is that we repeatedly apply the previous lemma to obtain the sequence of partitions of the $[0, 1]$ interval where, each step, either we find that we obtain some partition p_i such that it's a good approximation of w , in which case we stop, or the L2 energy increases by more than ϵ squared.

And since the final energy is always at most 1-- so it's always bounded between 0 and 1-- we must stop after at most $1/\epsilon$ steps. And if you calculate the number of parts, each part is subdivided into at most four parts at each step, which gives you the conclusion on the final number of parts. OK, so very similar to what we did before.

All right. So that concludes the discussion of the weak regularity lemma. So basically the same proof. Weaker conclusion and better quantitative balance.

The next thing and the final thing I want to discuss today is a new ingredient which we haven't seen before but that will play an important role in the proof of the compactness-- in particular, the proof of the existence of the limit. And this is something where I need to discuss martingales.

So martingale gill is an important object in probability theory. And it's a random sequence. So we'll look at discrete sequences, so indexed by non-negative integers. And is martingale is such a sequence where if I'm interested in the expectation of the next term and even if you know all the previous terms-- so you have full knowledge of the sequence before time n , and you want to predict on the expectation what the n th term is-- then you cannot do better than simply predicting the last term that you saw.

So this is the definition of a martingale. Now, to do this formally, I need to talk about filtrations and what not in measured theory. But let me not do that.

OK, so this is how you should think about martingales and a couple of important examples of martingales. So the first one comes from-- the reason why these things are called martingales is that there is a gambling strategy which is related to such a sequence where let's say you consider a sequence of fair coin tosses. So here's what we're going to do.

So suppose we consider a betting strategy. And x sub n is equal to your balance time n . And suppose that we're looking at a fair casino where the expectation of every game is exactly 0.

Then this is a martingale. So imagine you have a sequence of coin flips, and you win \$1 for each head and lose \$1 for each tail. When you're at time five, you should have \$2 in your pocket. Then time five plus 1, you expect to also have that many dollars. It might go up. It might go down. But in expectation, it doesn't change. Is there a question?

OK. So they're asking about, is there some independence condition required? And the answer is no. So there's no independence condition that is required. So the definition of a martingale is just if, even with complete knowledge of the sequence up to a certain point, the difference going forward is 0 in expectation.

OK, so here's another example of a martingale, which actually turns out to be more relevant to our use-- namely, that if I have some hidden-- think of x as some hidden random variable, so something that you have no idea. But you can observe it at time n based on information up to time n .

So for example, suppose you have no idea who is going to win the presidential election. And really, nobody has any idea. But as time proceeds, you make an educated guess based on the information that you have, all the information you have up to that point.

And that information becomes a larger and larger set as time moves forward. Your prediction is going to be a random variable that goes up and down. And that will be a martingale, because-- so how I predict today based on what are all the possibilities happening going forward, well, one of many things could happen. But if I knew that my prediction is going to, in expectation, shift upwards, then I shouldn't have predicted what I predict today. I should have predicted upwards anyway.

OK. So this is another construction of martingales. So this also comes up. You could have other more pure mathematics-type explanations, where suppose I want to know what is the chromatic number of a random graph. And I show you that graph one edge at a time.

You can predict the expectation. You can find the expectation of this graph's statistic based on what you've seen up to time n . And that sequence will be a martingale.

An important property of a martingale, which is known as the martingale convergence theorem-- and so that's what we'll need for the proof of the existence of the limit next time-- says that every bounded martingale-- so for example, suppose your martingale only takes values between 0 and 1. So every bounded martingale converges almost surely.

You cannot have a martingale which you expect to constantly go up and down. So I want to show you a proof of this fact. Let me just mention that the bounded condition is a little bit stronger than what we actually need. From the proof, you'll see that you really only need them to be L^1 bounded. It's enough.

And more generally, there is a condition called uniform integrability, which I won't explain. All right. OK. So let me show you a proof of the martingale convergence theorem. And I'm going to be somewhat informal and somewhat cavalier, because I don't want to get into some of the fine details of probability theory. But if you have taken something like 18.675 probability theory, then you can fill in all those details.

So I like this proof, because it's kind of a proof by gambling. So I want to tell you a story which should convince you that a martingale cannot keep going up and down. It must converge almost surely.

So suppose x_n doesn't converge. OK, so this is why I say I'm going to be somewhat cavalier with probability theory. So when I say this doesn't converge, I mean a specific instance of the sequence doesn't converge or some specific realization. If it doesn't converge, then there exists a and b , both rational numbers between 0 and 1, such that the sequence crosses the interval a, b infinitely many times.

So by crossing this interval, what I mean is the following. OK. So there's an important picture which will help a lot in understanding this theorem. So imagine I have this time n , and I have a and b . So I have this martingale. Its realization curve will be like that. So that's an instance of this martingale.

And by crossing, I mean a sequence that-- OK, so here's what I mean by crossing. I start below a and-- let me use a different color. So I start below a , and I go above b and then wait until I come back below a . And I go above b . Wait until I come back. So do like that. Like that.

So I start below a until the first time I go above b . And then I stop that sequence. So those are the upcrossings of this martingale. So upcrossing is when you start below a , and then you end up above b .

So if you don't converge, then there exists such a and b such that there are infinitely many such crossings. So this is just a fact. It's not hard to see.

And what we'll show is that this doesn't happen except with probability 0. So we'll show that this occurs with probability 0. And because there are only countably many rational numbers, we find that x_n converges with probability 1.

So these are upcrossings. So I didn't define it, but hopefully you understood from my picture and my description. And let me define by u_n to be the number of upcrossings up to time n , so the number of such upcrossings.

Now let me consider a betting strategy. Basically, I want to make money. And I want to make money by following these upcrossings.

OK. So every time you give me a number and-- so think of this as the stock market. So it's a fair stock market where you tell me the price, and I get to decide, do I want to buy? Or do I want to sell?

So consider the betting strategy where at any time, we're going to hold either 0 or 1 share of the stock, which has these moving prices. And what we're going to do is if x_n is less than a , is less than the lower bound, then we're going to buy and hold, meaning 1, until the first time that the price reaches above b and then sell as soon as the first time we see the price goes above b .

So this is the betting strategy. And it's something which you can implement. If you see a sequence of prices, you can implement this strategy. And you already hopefully see, if you have many upcrossings, then each upcrossing, you make money. Each upcrossing, you make money.

And this is almost too good to be true. And in fact, we see that the total gain from this strategy-- so if you start with some balance, what you get at the end-- is at least this difference from a to b times the number of upcrossings.

You might start somewhere. You buy, and then you just lose everything. So there might be an initial cost. And that cost is bounded, because we start with a bounded martingale. So suppose the martingale is always between 0 and 1. We start with a bounded martingale.

But on the other hand, there is a theorem about martingales, which is not hard to deduce from the definition, that no matter what the betting strategy is, the gain at any particular time must be 0 in expectation. So this is just the property of the martingale. So 0 equals the expected gain, which is at least b minus a times the expected number of upcrossings minus 1. And thus the expected number of upcrossings up to time n is at most 1 over b minus a .

Now, we let n go to infinity. And let u_{∞} be the total number of upcrossings. By the monotone convergence theorem in this limit, the limit of these u_n 's, it can never go down. It's always weakly increasing. It converges to the expectation of the total number of upcrossings.

So now, in particular, you know that the total number of upcrossings is at most some finite number. So in particular, the probability that you have infinitely many crossings is 0. So with probability 0, you cross infinitely many times, which proves the claim over there and which concludes the proof of the claim that the martingale converges almost surely.

OK, so that proves the martingale converge theorem. So next time, we'll combine everything that we did today to prove the three main theorems that we stated last time on graph limits.