We will have to build a lot of boring theory and definitions that may seem irrelevant to get to the results we need. While this may look intimidating, I will aim to make this more self contained than all of the other proofs I have seen of results like this, which annoyingly are all hard to follow because they are aimed at people who are better at maths than me even though I am at the level where I am sometimes expected to implicitly use these results, something I aim to fix in this document for future interested students like me. No more "Let (Symbol I've never seen) be a measure space (steps with symbols I haven't seen that I can't follow) intimidation" nonsense.

Definition: Pointwise supremum

Suppose I have a sequence of functions $f_1(x), f_2(x)$, ... then the pointwise supremum is the function g(x) such that g(a) is the least upper bound of the values $f_1(a), f_2(a), f_3(a)$, ...

Definition: Indicator function

An indicator function is a function that takes the value 1 if the input is in the set, and 0 otherwise. For example, the indicator function for the set (0,1) U (1.5,2) looks like this:

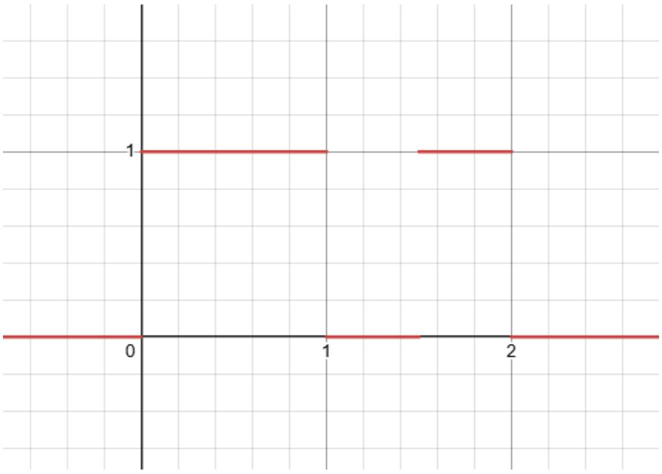

Image: Shows the graph of this indicator function: 1 in the interval, 0 elsewhere.

Here, I am estimating the integral $\int_0^3 3x - x^2 dx$ by a linear combination of indicator functions. The integral of an indicator function is defined as being equal to the total length of the sets in question, and in the image below we see that we have approximated the area by taking a bunch of indicator functions, multiplying them by something, and adding them together.

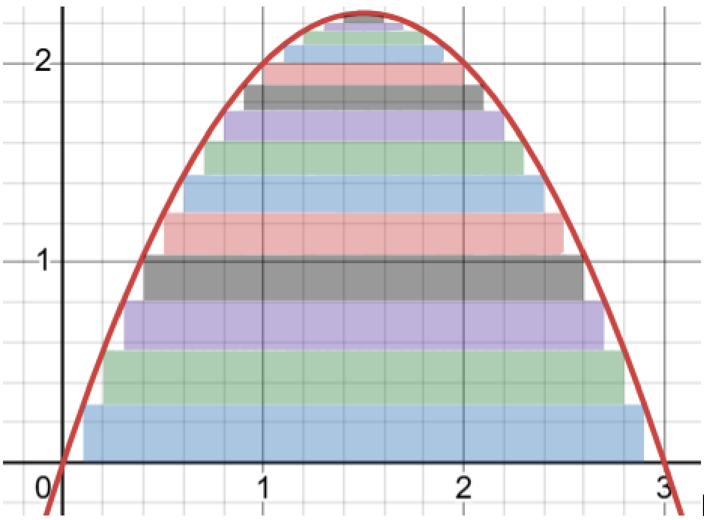

Image: Shows the integral of a graph approximated below by rectangles, instead of tall and thin they are short and thick.

Notice how this is an underestimate since our linear combination of indicator functions never exceeds our function. The integral is then defined as the least upper bound of the area we can get by underestimating it this way. That's the formal definition of the lebesgue integral.

Definition: Simple function

A simple function is a linear combination of indicator functions.

Now we prove a useful lemma known as the monotone convergence theorem. Suppose we have a sequence of non-negative functions $0 < f_1 < f_2 < f_3$ ... converging to f in the limit pointwise (ie, f is the pointwise supremum).

Now the theorem which we will show says that the limit of the integals of f_n is the integral of f on any interval. To do this, we will prove that eventually, $\int f_n$ eventually gets larger than$(\int f) - \varepsilon$ regardless of how small we make ε.

We know we can find a simple non-negative function g(x) such that $\int g \geq (\int f) - \frac{\varepsilon}{3}$ because $\int f$ is defined as the least upper bound of integrals of simple functions, meaning if we could not find simple functions below f whose integrals are arbitrarily close to the integral of f, f would not be the least upper bound. The reason for our choice of $\frac{\varepsilon}{3}$ will become clear eventually. Now we want to shift our g downwards by a constant δ in such a way that $\int g^- \geq (\int f) - \frac{2\varepsilon}{3}$ where g^- is defined as $g - \delta$ if $g > \delta$ and 0 otherwise. This ensures g^- is always non-negative and that it is at most $g - \delta$. So we have two cases: If the length of our interval that we are integrating along is a finite length l, then we pick δ to be $\frac{\varepsilon}{3l}$, since then the total amount that the rectangles get moved down by which is $\frac{\varepsilon}{3l}$ times the total length of the rectangles which is l is at most $\frac{\varepsilon}{3}$. If we are integrating from $-\infty$ to ∞ then we can still apply the same argument since the length on which our g is non-zero is finite. This is true because the integral is defined as the least upper bound of the integrals of all such g. Now define the set $S_n$as the set of values x such that $f_n(x) \geq g^-$. Then since the f's are increasing, each set S_n must contain the previous set S_{n-1}, ie the sets S are strictly growing. Now I claim there is an N such that the integral over the parts of our interval not containing S_N of f is less than or equal to $\frac{\varepsilon}{3}$. To do this, suppose the contrary, that there is a set T such that the integral of f over T is more than $\frac{\varepsilon}{3}$ but no point in T ever goes into any of the S_n's. We know that f is not 0 anywhere in T because if f(a) is 0 then a is in all S_n's since all f_n's will be 0 at those points by definition which means they satisfy $f_n(x) \geq g^-$since g^- would also be 0 at those points as g^- is non negative and never bigger than f which is 0. But then, this means that everywhere in T, g^- is strictly less than f, since if $0 < f < \delta$ then $g^- = 0$ and otherwise $g^- = f - \delta$. Therefore, $f_n(x) \leq g^-$ for all n and x in T by definition of the S_n's and of T, so the least upper bound of $f_n(a)$ for any point a in T is at most $g^-(a)$ which is strictly less than f(a), which contradicts the definition of f. This means if we let T_N be the interval not including S_n, then $\int_{T_N} f \leq \frac{\varepsilon}{3}$. Therefore, putting everything together, $\int f_n \geq \int_{S_N} f_n \geq \int_{S_N} g^- = \int g^- - \int_{T_N} g^- \geq \int g^- - \int_{T_N} f \geq \int g^- - \frac{\varepsilon}{3} \geq (\int f) - \varepsilon$. This completes the proof of the lemma.

Now we will prove another lemma known as **Fatou's lemma**.

Now, Fatou's lemma states that if we have a sequence of non-negative functions $f_n(x)$, then for every x define $f(x) = \liminf_{n\to\infty} f_n(x)$, then on any interval we have $\int f \leq \liminf_{n\to\infty} \int f_n$.

Now, for all fixed x, we define $g_n(x)$ as $\inf_{k\geq n} f_k(x)$. Then we know that for any fixed x, this is not decreasing as chopping more terms of the start cannot make the infimum lower. We also have the following by stuff we have discussed so far:

$$f(x) = \liminf_{n\to\infty} f_n(x) = \sup_n \inf_{k\geq n} f_k(x) = \sup_n g_n(x)$$

We have $\int f = \int \sup g_n = \sup \int g_n$ where the second equality is because g is increasing so we can apply the monotone convergence theorem. Since $\int g_n$ is an increasing sequence, the lim inf of this sequence is going to equal the limit of the sequence, as the infima of the tails are the same as the terms themselves. But the limit of an increasing sequence is the same as its supremum, so we now have that $\int f = \sup \int g_n = \liminf_{n\to\infty} \int g_n$. Because g_n is by definition not greater than f_n for all n at any input value, we have that $\int f = \sup \int g_n = \liminf_{n\to\infty} \int g_n \leq \liminf_{n\to\infty} \int f_n$, completing the proof of Fatou's lemma. From how we defined f, we have that $\int \liminf_{n\to\infty} f_n \leq \liminf_{n\to\infty} \int f_n$. We can also get that if f_n is bounded above by a function g with a finite integral on our interval for all n, then applying Fatou's lemma to $g - f_n$ gives us the reverse fatou lemma: $\int \limsup_{n\to\infty} f_n \geq \limsup_{n\to\infty} \int f_n$.

The **dominated convergence theorem** states that if we have a sequence of functions $f_n(x)$ which for all x converge to a function f(x) as n goes to infinity, and there is a function g(x) which has a finite integral on the interval we are working in, and that $|f_n(x)| \leq g(x)$ for all n and all x, then we have that $\int \lim_{n\to\infty} f_n = \lim_{n\to\infty} \int f_n$, essentially giving us a condition for when we can swap limits and integrals. Recall how at some places in the A level maths document we talked about the triangle inequality which says that |a+b|≤|a|+|b|, we will use this now to say that $|f - f_n| \leq |f| + |-f_n| = |f| + |f_n| \leq 2g$ (note that f is at most g at all places since it is the pointwise limit of functions that are at most g so it would not make sense for it to exceed g.) We also have by definition that $\limsup_{n\to\infty} |f - f_n| = 0$. Also, in this next step, we will use the fact analagous to the triangle inequality which says that $|\int h(x)| \leq \int |h(x)|$ for any function h(x). For real numbers this is the case because h is always between -|h| and |h|, so the integral of h will always be between the integral of -|h| and the integral of |h|, so the absolute value of that will be at most the integral of |h|. For complex numbers (which I am doing to show that we can generalize this beyond real integrals), we have $\int h = re^{i\theta}$ so $|\int h| = r$ and $\int he^{-i\theta} = r$. Since r is real, we have that $Re(\int he^{-i\theta}) = r$. Since the real part of the integral is the integral of the real part, we have $\int Re(he^{-i\theta}) = r \leq \int |he^{-i\theta}| = \int |h|$. Since $|\int h| = r \leq \int |h|$ this completes the proof. Now back to dominated convergence:

$$\left|\int f - \int f_n\right| = \left|\int (f - f_n)\right| \leq \int |f - f_n|$$

Now, we use the reverse fatou lemma:

$$\limsup_{n\to\infty} \int |f - f_n| \leq \int \limsup_{n\to\infty} |f - f_n| = \int 0 = 0$$

Therefore, since the lim sup is at most 0 and the terms in $\int |f - f_n|$ are non-negative, both the lim sup and the lim of this sequence must be exactly 0. So, $|\int f - \int f_n| \leq \int |f - f_n| \to 0$ so $|\int f - \int f_n| \to 0$ so $\int f_n$ approaches $\int f$ so the limit and integral interchange is justified.

Now, if $f_n(x)$ is defined as n for 0<x<1/n then the integral of this will be 1, but at all points between 0 and 1 f_n will eventually go to 0, so the integral of the limit is not the limit of the integral. It turns out that in this case it turns out we cannot find a function g such that for all n we have $\left|\int_0^1 f_n\right| \leq \left|\int_0^1 g\right| < \infty$. This is a standard textbook counterexample.

Another much easier result says that if I have $\iint f(x,y)dxdy$ then I can swap the integrals provided $\iint |f(x,y)|dxdy$ is finite. The reason we have this condition is because we know from the A level maths document that we can safely rearrange terms in a sum if the absolute value of the terms has a finite sum, and so the same applies for all the sums of $f(x,y)dxdy$ as dx and dy get smaller, as by moving the integrals around we are just changing the order in which we add the terms, which we already know is allowed whenever $|f(x,y)|dxdy$ converges.

Also, in my exponentials and logarithms video, I briefly mentioned that if a function has a single-valued antiderivative then the integral along a path of that function does not depend on the path. By "integrate along a path", I mean take the sum of the function times the distance you move by on the path (eg if you move from 1.07+3.12i to 1.08+3.14i you add (0.01+0.02i)f(1.07+3.12i)), then take the limit of these sums as these distances go to 0. Intuitively, similar to actual integration, taking the sum like this moves you along the antiderivative, which if it is single valued (unlike log, for example) will not allow you to possibly have path dependence.

Lemma (**Bolanzo weierstrass**): Every sequence that is bounded in absolute value has a(n infinite) convergent subsequence.

Idea: Plot the sequence on a graph then take the terms that are maxima of the first n terms, like this

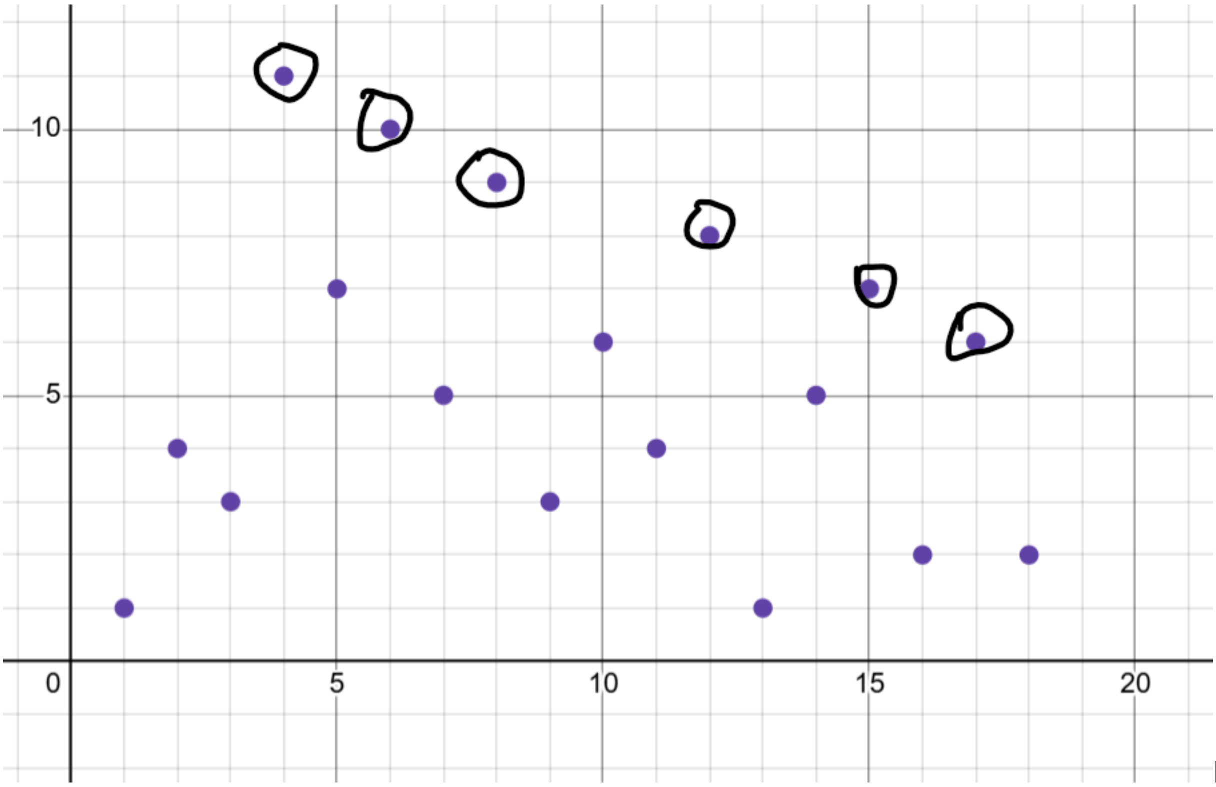

Image: Shows a sequence graph with peaks circled.

Here I have circled peaks, ie points that are larger/higher than (or equally as large/high as) all points after them. If there are infinitely many of these, then these form an infinite non-increasing sequence which is bounded below, which has a highest lower bound that it must converge to. If there are finitely many peaks, this means there has to be an infinite sequence of non-decreasing points, since after the last peak, the next term is not a peak meaning there is a term after that term that is larger than it, and that term is also not a peak so there is a larger term after that and so on. This forms an infinite increasing sequence which is bounded above, which converges to its least upper bound. So done.

Formal definition of a continuous function:

We want to say that as the inputs get close together the outputs of the function get arbitrarily close together since that is what it intuitively means to get continuous. We will write this precisely as follows:

A function f(x) is continuous at x if for any ε, no matter how small, you can pick δ small enough that for any a with |x-a|< δ, |f(a)-f(x)|< ε.

We will need this definition because we will prove that if a function is continuous everywhere on a finite interval, including the endpoints of the finite interval, then it is uniformly continuous (I will explain what this means shortly). This is because, and I can't believe I'm saying this, this is a technical dependency for a technical dependency (convergence in distribution implies convergence in characteristic function) of a technical dependency (cramer wold device) of a technical dependency (multivariate clt) for the chi squared result. In fact bolzano weierstrass was for FIVE levels of dependency. Uniformly continuous means that not only are we able to pick a δ for each x, but a δ that works for all x. An example of a function that is not uniformly continuous is 1/x on the open interval (0,1), since if I pick a certain ε, then no matter how small I pick δ, I can go close enough to 0 that the |x-a|< δ implies |f(a)-f(x)|< ε condition is not satisfied, so I cannot pick a δ that works for all x.

Proof of result: Since we are assuming our function is continuous on a closed bounded interval, bolzano weierstrass applies by boundedness, and the subsequence in question converges to a limit in the interval by closure of the interval.

Assume for a contradiction that our function is continuous but not uniformly continuous on our closed bounded interval. Then there is an ε>0 such that if I pick $\delta = \frac{1}{n}$ there is some x_n, y_n with $|x_n - y_n| < \frac{1}{n}$ but $|f(x_n) - f(y_n)| > \varepsilon$. We know that x_n and y_n both have convergent subsequences converging to places on our closed interval, and in fact they converge to the same place since the difference betweenx_n and y_n gets arbitrarily small. If the subsequences in question are x_{n_i} and y_{n_i} and their limits are x, then $f(x_{n_i}) \to f(x)$ and $f(y_{n_i}) \to f(x)$ because continuous have functions that whenever the inputs are close the outputs are close, so intuitively you can pass limits through continuous functions like this. But $f(x_{n_i})$ and $f(y_{n_i})$ somehow converge to the same limit but are always > ε, which is a contradiction.

Now, I will talk about cardinality. This is really fun stuff, but I will not go through all the fun results here, only the results we need. The motivation of this is to show that there are, in a precise sense, "more" real numbers than natural numbers, so I can show that you cannot take a sum over all the real numbers of positive numbers and still get a finite number.

We consider two sets to be the same size if there is a one to one correspondence between their elements. This sounds obvious, but it shows that there are equally as many even positive integers as positive integers, because we can map 1—2 and 2—4 and 3—6 and so on, giving us a one to one correspondence.

Now we prove we cannot find a one to one correspondence between real numbers on any interval and positive integers. We will do this for the interval (0,1) then note that we can get to any interval by scaling, or the entre reals by doing cot(πx). The property of there being a correspondence to positive integers is called being countable. But, I think listable is a better term.

So, suppose we do, in fact, have a list of all the real numbers between 0 and 1.

1. 0.3141592653589793…
2. 0.2718281828459045…
3. 0.1414213562373095…
4. 0.3498579345858968…
5. 0.9988237478199283…

Now construct a new number where the n'th decimal digit after the point is not the red number, and such that we do not have infinite trailing 9's (to ensure the decimal expansions determine a unique number. We have also assumed this about the numbers on our list). For example, adding 1 to each red we could have 0.48293… But this is clearly not on the list – It differs from everything on the list by at least one digit. Contradiction. Essentially, this makes it so if you think you found a way to list them, I can prove you are wrong.

Now if we had a sum over the real numbers of positive numbers x, split them into subsets:

$x \geq 1, \frac{1}{2} \leq x < 1, \frac{1}{3} \leq x < \frac{1}{2}, \frac{1}{4} \leq x < \frac{1}{3}.$

Then one of these subsets has to have infinitely many elements. Why? If they all had finitely many elements then they would form a subset of this infinite table:

	1	2	3	4	5	6	7	8	9	10
Set 1										
Set 2										
Set 3										
Set 4										
Set 5										
Set 6										
Set 7										
Set 8										
Set 9										

Table: empty table with rows as the sets defined above, meant to be filled with the terms in our uncountable sum.

So I could then go along the table in a zigzag pattern like this in the order shown below, adding to my list whenever the elements are in the table, and not adding them otherwise.

	1	2	3	4	5	6	7	8	9	10
Set 1	1	2	6	7	15	…				
Set 2	3	5	8	14						
Set 3	4	9	13							
Set 4	10	12								
Set 5	11									
Set 6										
Set 7										
Set 8										
Set 9										

Table: Shows 1, 2, 3, in the relevant cells to visually show a one to one mapping between the numbers and the table cells, if the table were to go on forever.

All cells get reached eventually, so this would imply we can list the cells. But we are assuming we are adding one term for each real number, and we cannot list the real numbers. Therefore, there is a set with infinitely many elements, so the sum is infinite.

Another theorem, this time about matricies:

We are in R^n (n dimensional vector space of reals) and for m<n there is an m-dimensional plane centered at the origin, then we can rotate the space around so that the first m basis vectors are in our plane, then consider where the rest of the basis vectors were before we did the rotation: Intuitively, we are considering a basis for vectors perpendicular to this plane. If S is a n*(n-m) matrix where each column of s is one of these vectors, we want to show that $I - SS^T$ projects any vector onto the m-dimensional plane. To do this, we just need to show three things:

1. A vector when this linear transformation is applied to it moves orthogonally to the direction of the plane
2. Any vector ends up on the plane after the transformation

Consider a vector v, then $(I - SS^T)v = Iv - SS^T v = v - SS^T v$

Condition 1: Orthogonality

The vector moves by $SS^T v$. Therefore we want to show that $(SS^T v).u = 0$ if u is on the plane. This is the same as $(SS^T v)^T u = v^T SS^T u$. But we know $S^T u$ is 0 since we are assuming u is on the plane (so all columns of S as a dot product with v return 0 so the result follows), so the whole thing becomes 0. So done.

Condition 2: A vector ends up on the plane after the transformation

We want to show $(I - SS^T)v$ ends up on the plane. Since the plane is defined by $S^T u$ is 0 if u is on the plane, we want to show that $S^T(I - SS^T)v = 0$. This is equal to $S^T v - S^T SS^T v = (I - S^T S)S^T v$.

Now what does $S^T S$ equal? It will be an (n-m)x(n-m) matrix where the i,j entry is equal to $s_i.s_j$. Since the s's are orthogonal unit vectors, this will be 1 when i=j and 0 otherwise, so we get the identity matrix. Therefore $I - S^T S = 0$. So done.

Now suppose B is a matrix where the columns are vectors that form an orthonormal basis of the plane (Essentially this means what you would expect: Where the basis vectors on the plane were before the rotation). Then $I - BB^T$ is the projection matrix onto the space perpendicular to the plane, since it is essentially the same idea with S and B being renamed to eachother, as they are both matricies with perpendicular columns within them and between them. Now a vector v decomposes into vectors v_s and v_b, where v_b is the component of v in the direction of the plane, and v_s is the component of v perpendicular to the plane. So $(I - BB^T)v = v_s$ and $(I - SS^T)v = v_b$, so $(I - BB^T)v + (I - SS^T)v = v_s + v_b = v$. So$(2I - BB^T - SS^T)v = v$. Since this is true for all v, we must have that $I = 2I - BB^T - SS^T$ so $I = BB^T + SS^T$. What I have proven is that if B and S are matricies whose columns together form a basis for a vector space with all vectors perpendicular to eachother then $I = BB^T + SS^T$. Alternatively, I have shown that if B forms a basis for the plane with all vectors in B perpendicular to eachother, then BB^T gives a projection onto the plane.