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.

A quick note on open sets:

An interval on the number line is called open if it goes from A to B without containing A and B, ie it does not contain its boundary. The same can be said for subsets of the plane or hyperplanes – they are open if they do not contain any of their boundary. They are closed if they contain all of their boundary. The formal definition of a boundary is a point where no matter how small of a circle or ball you make around that point, it will contain points both inside and outside the set. Note that since an open set contains none of its boundary, it has the property that for any point in the open set, all points sufficiently close to that point are in that set, so open sets must always have some thickness to them. This is related to the idea that there is no smallest real number greater than another real number, and whether sets are open or closed turns out to be surprisingly important. Also, arbitrary unions of open sets are open since each point has a "ball" around it inside each of the open sets in the union and thus the whole thing, similarly arbitrary intersections of closed sets are closed. The reverse implications only hold for finite unions and intersections (since infinitely many balls intersection could be a point but finitely many is a ball).

We usually call open intervals (a,b) meaning everything from a to b not inclusive, and closed intervals [a,b] to mean everything from a to b inclusive, and [a,b) means everything from a to b including a but not b.

We need to recall the definition of the Lebesgue integral from the end of level 4. 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), and suppose f is integrable (ie its integral is finite).

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}. 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.

Note which will make sense later: We define the integral of the dirac delta parts separately, but the theorem above still applies to those – I will explain why in the same part where I define dirac deltas in the stats document.

Note also that the same argument can be used to prove the above theorem (and thus the theorems below that depend on it) for integrals of functions over 2D, 3D, or higher dimensional areas. This will be important later. Bolzano weierstrass (A theorem proved below) and its consequences can be used provided the set is not only bounded but also **closed**, since we want sequences to not only converge but have their limit be part of the set.

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 $f(x) = \lim_{n\to\infty} \inf f_n(x)$, then on any interval we have $\int f \leq \lim_{n\to\infty} \inf \int f_n$.

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 $\int f = \sup \int g_n = \lim_{n\to\infty} \inf \int g_n$. Because g_n is by definition not greater than f_n for all n at any input value, we have that $\int f = \sup \int g_n = \lim_{n\to\infty} \inf \int g_n \leq \lim_{n\to\infty} \inf \int f_n$, completing the proof of Fatou's lemma. From how we defined f, we have that $\int \lim_{n\to\infty} \inf f_n \leq \lim_{n\to\infty} \inf \int f_n$. 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: $\int \lim_{n\to\infty} \sup f_n \geq \lim_{n\to\infty} \sup \int f_n$.

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 level 4 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 $\lim_{n\to\infty} \sup|f - f_n| = 0$. 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 result says that if I have $\iint f(x,y)dxdy$ on a rectangle (that may go off to infinity) then I can swap the integrals provided $\iint |f(x,y)|dxdy$ is finite on that rectangle. An intuitition for we have this condition is because we know from level 4 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 $\iint |f(x,y)|dxdy$ converges. We will actually prove that in the multivariable case we can change the order of integrals.

Proof of the claim above:

We can define the integral as $\iint f(x,y)d(x,y)$ in the supremum sense analagous to above. Then if we show that this equals $\int(\int f(x,y)dx)\,dy$ then the result will follow by symmetry. If we have more than two variables, our proof will show that we can split the many variables into 2 and split it into multiple integrals. For example, $\iiiint f(x,y,z,w)d(x,y,z,w) = \iint[\iint f(x,y,z,w)d(x,y)]\,d(z,w)$

Part 1 – Proof for non-negative functions:

We will do this for the 2D case since it is precisely the same idea as the general case.

Set $g(x) \coloneqq \int f(x,y)dy$. Note that the theorem holds for the indicator of a rectangle by direct calculation, and thus it holds for all non-negative simple functions (since they are just finite sums of indicators). There exists a sequence of non-negative simple functions that is non-decreasing and converges to f and is such that $\iint f_n(x,y)d(x,y) \to \iint f(x,y)d(x,y)$ (In fact, by monotone convergence, pointwise convergence to f is enough to gurantee that $\iint f_n(x,y)d(x,y) \to \iint f(x,y)d(x,y)$ if f is non-negative and non-decreasing). Pick such a sequence f_n and define $g_n(x) \coloneqq \int f_n(x,y)dy$.

By the monotone convergence theorem applied to $\int f_n(x,y)dy = g_n(x)$, we have that $g_n(x) \to g(x)$ as $\int f_n(x,y)dy \to \int f(x,y)dy$, since f_n is a non-decreasing sequence of non-negative functions converging pointwise to f. After this, we can apply it again in x to $\int g_n(x)dx$ to show that $\int g_n(x)dx \to \int g(x)dx$, for exactly the same reason.

Finally, we already know by how we picked f_n that $\iint f_n(x,y)d(x,y) \to \iint f(x,y)d(x,y)$. For each n, $\int g_n(x)dx = \iint f_n(x,y)d(x,y)$ as the theorem is proved for simple functions.

But then the limits as n goes to infinity of $\int g_n(x)dx$ and $\iint f_n(x,y)d(x,y)$ must be the same as the integrals are equal for all n. This means $\int\int f(x,y)dy\,dx = \int g(x)dx = \iint f(x,y)d(x,y)$ as required.

The symmetric argument allows us to conclude that $\iint f(x,y)dxdy = \iint f(x,y)d(x,y) = \iint f(x,y)dydx$ when f is non-negative.

Part 2 – Extension to general case

Since the integral of the absolute value of f is finite, it means if I restrict f to where it is positive, and define a function g as -f where f is negative, I can apply part 1 to both of those, and then subtract the results. This is only problematic when we do not have absolute convergence and then we get $\infty - \infty$.

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

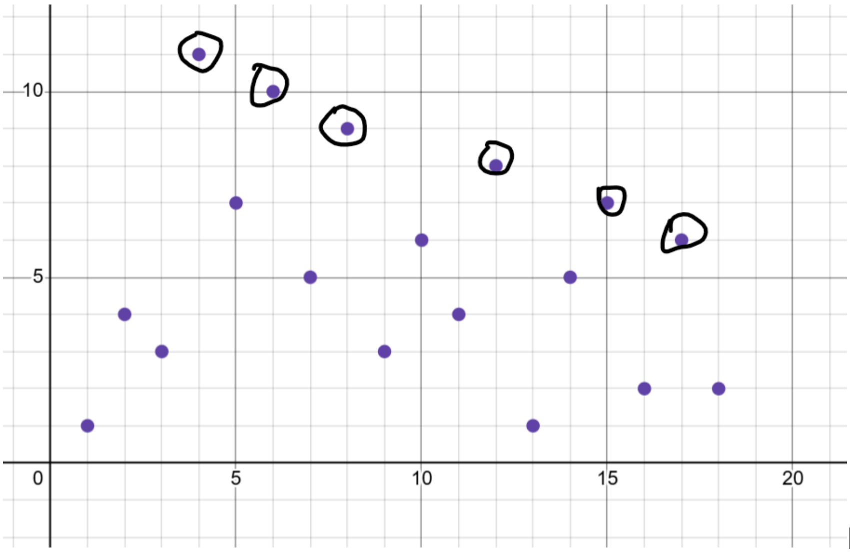


Image: Shows a sequence graph with peaks circled.

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 or 0's (to ensure the number has a unique decimal representation. For the numbers in our list we pick the one with trailing 0's and not trailing 9's whenever we have to make this choice). 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:

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

	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										

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.

Definition (big O notation):

A function g(x) is O(f(x)) if when x is sufficiently close to x_0, |g(x)| is bounded by M|f(x)| where M is some fixed positive constant. If f(x) is not 0 in the vicinity of x_0 we can write that $|\frac{g(x)}{f(x)}|$ is bounded by M when x is sufficiently close to x_0. We can also say, and this is often done in computer science when analyzing how long an algorithm will take, that a function is a function g(x) is O(f(x)) as x goes to infinite if when x is large enough, |g(x)| is bounded by M|f(x)|. In previous documents, the value of x_0 has been implied. Sometimes people write g(x)=O(f(x))

Definition (little o notation):

Like big O notation except if g(x) is o(f(x)) at x_0 it means that g(x) gets much smaller than f(x) when x is sufficiently close to x_0. Precisely, it means that M can be made arbitrarily small by making x sufficiently close to x_0, or by making x sufficiently large in the case x_0 is infinity.

Note that if f(x)=o(g(x)), then we necessarily also have f(x)=O(g(x)).

Examples:

$x \neq O(x^2)$ as $x \to 0$ since x^2 is much smaller than x. In fact, $x^2 = o(x)$ as $x \to 0$.

$x = O(x^2) = o(x^2)$ as $x \to \infty$ but $x^2 \neq O(x)$ as $x \to \infty$.

$x = O(\sqrt{x})$ as $x \to 0$.

$\sin(2x) = O(x)$ as $x \to 0$, but $\sin(2x) \neq o(x)$ as $x \to 0$.

$x \neq O(x\sin(x))$ as $x \to \infty$, because although x does not seem to be much larger than xsin(x), the ratio $\frac{x}{x\sin(x)}$ is unbounded when x gets close to integer multiples of π.

Also, a principle here is the idea that a polynomial is dominated by its leading term. For example

$2x^3 + 4x + 12 = O(x^3)$ as $x \to \infty$ because when x>1, $|2x^3 + 4x + 12| \leq |2x^3| + |4x| + |12| < |2x^3| + |4x^3| + |12x^3| = 18|x^3|$ so we can put M=18, then done. Also, big o notation can capture the idea that exponentials beat polynomials (by the way you can prove this as $\frac{a^x}{x^b}$ goes to infinity as x goes to infinity which you can see if you apply l'hopital's rule at least b times), as we can write that P(x)=o(exp(x)) for any polynomial P. Also, $\log(x) = o(x^\varepsilon)$ for all positive ε and this can be proven the same way, or by thinking of it intuitively as a sort of inverse of the exponentials beat polynomials idea. Note exp(x) means e^x.

Another idea is that non-zero constants don't matter here since we're looking at the bigger picture with the growth rate of these functions.

Example:

Since $\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = f'(x)$, $\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} - f'(x) = 0$ so $\frac{f(x+h)-f(x)}{h} - f'(x) = o(1)$ as x goes to h by definition. Therefore, $f(x+h) - f(x) - hf'(x) = o(h)$ by multiplying by h on both sides, so we can write that $f(x+h) - f(x) = hf'(x) + o(h)$, which looks a lot like what I did when I justified differentiating infinite power series.

Definition: Taylor polynomial

A taylor polynomial of degree n is a taylor series for a function that is n times differentiable but only up to the term in x^n. We write this as $T_n(f)(x)$. This is a polynomial which gives an approximation of f near some point x_0, with error equal to $o((x - x_0)^n)$.

Claim 1: If f is n times differentiable at x_0 then $f(x) - T_n(f)(x) = o((x - x_0)^n)$ as $x \to x_0$ if x_0 is the point we are doing a taylor series around.

Claim 2: If f is n+1 times continuously differentiable on the interval $[x_0, x]$ then $f(x) - T_n(f)(x) = O((x - x_0)^{n+1})$ as $x \to x_0$ if x_0 is the point we are doing a taylor series around.

Claim 3: If the conditions for claim 2 hold, then there exists some $t \in (x_0, x)$ such that $f(x) - T_n(f)(x) = \frac{(x-x_0)^{n+1}}{(n+1)!} f^{(n+1)}(t)$.

We will prove this.

Lemma 1 (Extreme value theorem):

Any function continuous on a finite closed interval is bounded.

Note: It is very important that the interal is closed. On the open interval (0, 1) we can define 1/x which is continuous on that interval, but not on the closed interval [0,1] because the closed interval contains x=0 but the open interval does not. But there's not much you can do to make a continuous unbounded function defined on a finite closed interval – Try it!

We will use the fact that every bounded sequence has a convergent subsequence since that was proven in the level 6 technical results document.

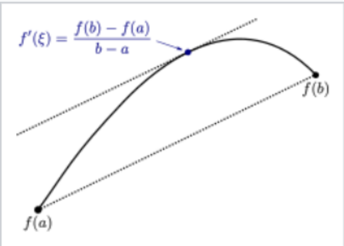
Suppose for a contradiction f is unbounded. Then for all n, we can choose x_n in [a,b] with $|f(x_n)| > n$. Now pick a convergent subsequence x_{n_k} of the sequence formed by each each x_n, and call its limit L, which is in [a,b] since [a,b] is closed so limits of sequences in [a,b] are in [a,b]. By continuity, $f(x_{n_k}) \to f(L)$, but since $f(x_{n_k})$ is unbounded, it follows that this is a contradiction, since $f(L)$ cannot be defined.

Important note: Bolzano weierstrass holds for sequences of points in k dimensions, provided the bounded set the sequence is in is closed as mentioned above – there is a subsequence whose x coordinate converges and a subsequence of that whose y coordinate converges etc, and thus the above theorem holds for functions of more than one variable, as well as the theorem about uniform continuity and about riemann integrals for continuous functions, as the definition of continuity is analagous. This is important since we use these theorems in the context of functions of multiple variables.

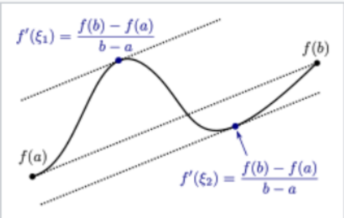
Lemma 2 (Mean value theorem):

This lemma assumes the values are real, so everything proven here applies to real numbers, but not general complex numbers.

First of all, a picture (from wikipedia):



The function f attains the slope of the secant between a and b as the derivative at the point $\xi \in (a, b)$.



It is also possible that there are multiple tangents parallel to the secant.

You can see that it's kind of obvious that the function's derivative attains its mean value, since the function is differentiable everywhere by assumption so there are no spikes. This is essentially what the theorem says. However, to prove it, we want to subtract the line from (a, f(a)) to (b, f(b)) from y. This line has the slope we want so we want to prove that our new function, which is 0 at a and b, attains a derivative of 0 between a and b. This new function is either constant, in which case the theorem is obvious, or it goes above 0 at some point (If not prove the theorem for minus the function). Since it is continuous on a closed, bounded interval, it is bounded, and has either a maximum or a minimum. At such a point, the derivative exists by our hypothesis, and it can't possibly be something other than 0. Why? Suppose the derivative were positive at a maximum point c, then by definition of limits, $\frac{f(x)-f(c)}{x-c}$ would be positive if we make x close enough to c with x>c, implying f(x)>f(c) which is a contradiction.

In particular, we have Rolle's theorem, which says that if f(a)=f(b)=0 and f is differentiable on (a,b) then there is a point a<x<b such that f'(x)=0

Lemma 3 (Generalized Rolle's theorem):

Here is the statement and proof taken from Cambridge notes:

Theorem (Higher-order Rolle's theorem). Let f be continuous on $[a, b]$ $(a < b)$ and n-times differentiable on an open interval containing $[a, b]$. Suppose that

$$f(a) = f'(a) = f^{(2)}(a) = \cdots = f^{(n-1)}(a) = f(b) = 0.$$

Then $\exists x \in (a, b)$ such that $f^{(n)}(x) = 0$.

Example: The function $x^3 - x^2$ has its first derivative and value equal to 0 when x=0, and its value equal to 0 at x=1, so between 0 and 1 it must have a point of inflection. Indeed it does when x=1/3.

Theorem 1:

If $f(x)$ is n times differentiable at $x = x_0$ then

$$f(x) - f(x_0) - (x - x_0)f'(x_0) - \frac{(x-x_0)^2}{2!}f''(x_0) - \cdots - \frac{(x-x_0)^n}{n!}f^{(n)}(x_0)$$

is $o((x - x_0)^n)$ as $x \to x_0$.

Proof:

The first n derivatives of the above expression are all 0 when evaluated at x_0 because for k between 0 and n inclusive the $\frac{(x-x_0)^k}{k!}f^{(k)}(x_0)$ term has k'th derivative equal to $f^{(k)}(x_0)$ and all its other derivatives are 0 at x_0 by repeated application of the power rule so therefore it cancels the k'th derivative of the $f(x)$ term to make it equal to 0.

So, by the definition of o, we need to show that $\frac{f(x)}{(x-x_0)^n} \to 0$ as $x \to x_0$. This is fine as $(x - x_0)^n$ is not 0 when you make it arbitrarily small so it can go in the denominator.

Now, if we can show that any function whose first n derivatives are all 0 at x_0 is $o((x - x_0)^n)$, we will be done. It is clear that a function whose value and first derivative is 0 at x_0 is $o(x - x_0)$ since $\frac{g(x)}{x-x_0} = \frac{g(x)-g(x_0)}{x-x_0} \to g'(x_0) = 0$, so this theorem is true if n=1. But now suppose that we know that this theorem is true for n=k, then we will prove it is also true for n=k+1, therefore proving this by induction. Let g(x) be such that at x_0 its value and first k+1 derivatives are all 0. By the induction hypothesis, g'(x) has its first k derivatives and value equal to 0 at x_0 so $g'(x_0 + h) = h^k\varepsilon(h)$ with $\varepsilon(h) \to 0$ as $h \to 0$, by the definition of little o notation and the induction hypothesis. By the mean value theorem, there exists some θ between 0 and 1 such that

$g'(x_0 + \vartheta h) = \frac{g(x_0+h)-g(x_0)}{h}$, or equivalently $hg'(x_0 + \vartheta h) = g(x_0 + h)$, meaning that by the induction hypothesis $g(x_0 + h) = h\left(h^k\varepsilon(\vartheta h)\right) = h^{k+1}\varepsilon(\vartheta h) = o(h^{k+1})$. So induction done and proof complete.

Remark:

Just because each remainder is small does not mean that the taylor series converges, since convergence at some x requires that there, the remainder is uniformly small, but it could be that the interval where the remainder is small shrinks and is zero in the limit. There are functions such that the taylor series does not converge in any interval despite the functions being infinitely differentiable, like $\exp(-x^{-2})$ around x=0. However, it turns out that if the function has a complex derivative the taylor series always converges in some interval, see IB complex analysis. I'm not proving it because we don't need to use it for this purpose.

Theorem 2:

If $f(x)$ is n+1 times continuously differentiable in an open interval containing x_0 then $f(x) - f(x_0) - (x - x_0)f'(x_0) - \frac{(x-x_0)^2}{2!}f''(x_0) - \cdots - \frac{(x-x_0)^n}{n!}f^{(n)}(x_0) - \cdots$ is $O((x - x_0)^{n+1})$ as $x \to x_0$.

Proof:

Lemma:

Let f be n+1 times differentiable on [a,b] with it's (n+1)'th derivative continuous. Then we have that $f(b) = f(a) + (b - a)f'(a) + \frac{(b-a)^2}{2!}f^{(2)}(a) + \cdots + \frac{(b-a)^n}{n!}f^{(n)}(a) + \int_a^b \frac{(b-t)^n}{n!}f^{(n+1)}(t)dt.$

Proof of lemma: We will do this by induction on n. When n=0 the theorem says

$$f(b) = f(a) + \int_a^b f'(t)dt$$

Which is known to be true. Now we will do integration by parts to get

$$\int_a^b \frac{(b-t)^n}{n!}f^{(n+1)}(t)dt = \left[\frac{-(b-t)^{n+1}}{(n+1)!}f^{(n+1)}(t)\right]_a^b + \int_a^b \frac{(b-t)^{n+1}}{(n+1)!}f^{(n+2)}(t)dt$$

$$= \frac{(b-a)^{n+1}}{(n+1)!}f^{(n+1)}(a) + \int_a^b \frac{(b-t)^{n+1}}{(n+1)!}f^{(n+2)}(t)dt$$

So the lemma follows by induction.

Now this implies the result since $f^{(n+1)}$ is continuous on the closed interval [a,b] and thus bounded there, say it is bounded in absolute value by M. Then the integral is at most $\int_a^b \frac{(b-t)^n}{n!}Mdt = M\frac{(b-a)^{n+1}}{n+1!}$ which is $O((x - a)^{n+1})$ as required.

Theorem 3:

$f(x) - f(x_0) - (x - x_0)f'(x_0) - \frac{(x-x_0)^2}{2!}f''(x_0) - \cdots - \frac{(x-x_0)^n}{n!}f^{(n)}(x_0) = \frac{(x-x_0)^{n+1}}{(n+1)!}f^{(n+1)}(t)$ for some t(x) between x_0 and x, provided the first n+1 derivatives of f exist and are continuous from x_0 to x.

Proof:

The n=0 case is just the mean value theorem. Consider $g(x) \coloneqq f(x) - f(x_0) - (x - x_0)f'(x_0) - \frac{(x-x_0)^2}{2!}f''(x_0) - \cdots - \frac{(x-x_0)^n}{n!}f^{(n)}(x_0)$. This has its value and first n derivatives at x_0 equal to 0 by the same reasoning as we gave in one of the earlier proofs

Set $\frac{g(x)}{(x-x_0)^{n+1}} = C$. Then $g(x) - C(x - x_0)^{n+1}$ is 0 at x from how we defined C and has its first n derivatives and value equal to 0 at x_0 meaning it satisfies the conditions for generalized rolle's theorem. Therefore, there exists a point y between x_0 and x such that the n+1'th derivative of

$g(y) - C(y - x_0)^{n+1}$ is 0. But the derivative of the second term is just C(n+1)! by the power rule, thus the n+1'th derivative of g at y is also C(n+1)!, thus $C = \frac{g^{(n+1)}(y)}{(n+1)!}$. Therefore, since we have that $g(x) - C(x - x_0)^{n+1}$ is 0 at x from earlier, we thus have that

$$g(x) = C(x - x_0)^{n+1} = \frac{g^{(n+1)}(y)}{(n+1)!}(x - x_0)^{n+1}$$

Which is exactly what we wanted to prove.

Example: There exists some point between 0 and 1 such that $\frac{e^x}{2} = e - 2$, because if $x = 1$ and $x_0 = 0$, then $e^x - x - 1 = \frac{x^2}{2}e^y$ for some y between 0 and 1, but x=1 so this simplifies to $\frac{e^y}{2} = e - 2$.