Math Foundations Notes

Srikar Katta and Edoardo Airoldi

Contents

Math Foundations

Sets, Functions, and Notation

Sets are at the heart of mathematics, applied statistics, and data science. The mathematical realm that we most often work in – the real numbers – is a set, and understanding the ideas behind sets can help a data scientist understand

1. the terminology used in this book and by other empirical researchers
2. the mathematical foundations of probability/statistics.

While the foundations of statistics from a mathematical perspective is far beyond the scope of this book, it is nevertheless useful to understand the basic idea behind sets.

Definition 0.1: Set

A **set** is a collection of unique objects (elements), such as numbers, and is often denoted by braces (i.e., $\{\ldots\}$) or capital letters (e.g., A). The set with no elements is called the **empty set** and is denoted by $\emptyset$.

For example, {1, 2, 3, 4} and {statistics, computer science, data science} are considered to be sets because they are both collections of distinct elements. Even though the second example is not mathematical, it satisfies the basic definition of a set and is therefore a set. However, {1, 1, 2, 3, 4} is not a set because it contains 1 twice.

Set Operations

Just as numbers can be added and multiplied, sets have their own operations. We are concerned with set *union* and set *intersection* as these show up in the basic probability theory that is required for data science.

Definition 0.2: Set Union

The **union** of a collection of sets (denoted by $\cup$) is the set of all elements in the collection.

Definition 0.3: Set Intersection

The **intersection** of a collection of sets (denoted by $\cap$) is the set of all elements shared by every set in the collection.

For example, suppose we have sets $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$. The union of A and B is $A \cup B = \{1, 2, 3, 4, 5\}$ since the union contains all the elements in A and B. On the other hand, the intersection of A and B is $A \cap B = \{3\}$ since 3 is the only element in both sets.

While set operators are useful, very rarely do people say "take the intersection of sets A and B." We implicitly use these set operators in our everyday lives. Suppose someone is majoring in mathematics and statistics and needs to identify what courses they need to take. For their math curriculum, they take courses like calculus, probability theory, algebra, analysis, etc. And for their statistics curriculum, they take courses like calculus, probability theory, regression analysis, applied statistics, etc. The math and statistics curricula are each their own sets because they are collections of distinct courses.

Suppose someone asks, "What are the required courses in math **and** statistics?" The use of "and" in that statement suggests the intersection of the sets of courses because they want courses shared by both sets of courses. If someone wanted to find out what all courses they need to take, then that is the same as finding the union. In everyday English, that is equivalent to saying, "What are the courses in math **or** statistics?" The use of "or" in that statement suggests the union of the sets of courses because they want the set of all courses. Understanding the relationship between "and"/"or" and sets allows a data scientist to translate domain problems into math.

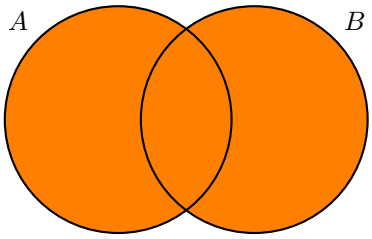


Figure 1: $A \cup B$

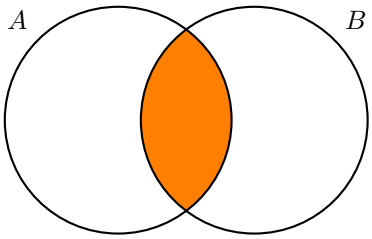


Figure 2: $A \cap B$

Functions

While sets on their own are very powerful tools, sets can be related to one another through **functions**. Secondary school algebra introduces functions to students, but it is not as rigorously defined as it is in this

section. In applied statistics, understanding how functions work can guide the practitioner to decide what probability distributions most likely fit with certain quantities. This will be expanded upon more in later sections, but functions are essential to mathematics and statistics.

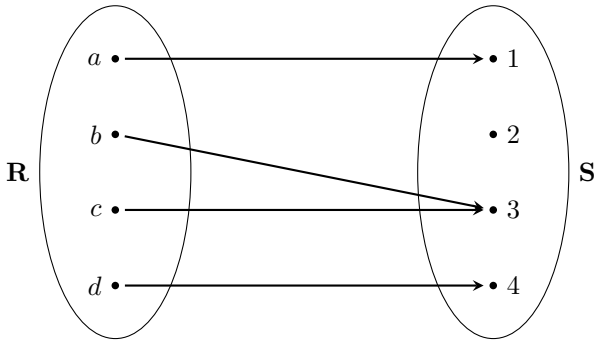
Definition 0.4: Function

Suppose we have two sets A and B. A **function** is a process that associates each element A – referred to as the **domain** – to a single value of the set B – called the **co-domain**. In other words, if f is a function, then for all elements a in A, there exists one and only one element b contained in the set B such that $f(a) = b$.

Suppose we have a relationship f that maps values from the set \$R = \$ {a, b, c, d} to the set $S = \{1, 2, 3, 4, 5\}$ defined by

$$f(r) = \begin{cases} 1 \text{ if } & r = a \\ 3 \text{ if } & r = b \text{ or } r = c \\ 4 \text{ if } & r = d. \end{cases}$$

Graphically, the function would look as follows:

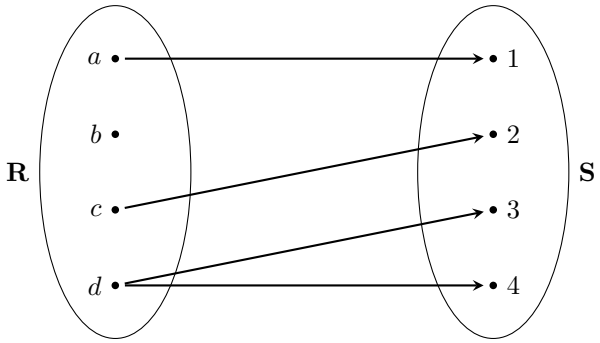


Notice that *every* element of R maps to *one and only one* value of S, so f is a function. Also, in this example, there is no element r in R such that $f(r) = 2$. People often make the mistake of saying that since nothing in R maps onto 2 that f is not a function. However, this is not a requirement of functions and is therefore not a problem. Additionally, notice that $f(b) = 3$ and $f(c) = 3$. Even though both values map into 2, this is not a problem.

Now consider the relation g that maps values R to S defined by

$$g(r) = \begin{cases} 1 \text{ if } & r = a \\ 2 \text{ if } & r = a \\ 3 \text{ if } & r = b \\ 4 \text{ if } & r = d. \end{cases}$$

Graphically, this relationship would look as follows:



This relation g is *not* a function for two reasons:

1. Even though b is an element of R, it is not mapped to any value in S
2. An element of the domain, d, maps to *two* values and it can map to one and only one for g to be considered a function.

Another key concept of functions is the **image** of the function. Suppose f maps values from arbitrary sets A to B; this does not mean that all the values in B must have some corresponding value in A (i.e., for each element b in B, there does not need to exist a value in A such that $f(a) = b$). Because of this, there may be some values in B that are not mapped to. In order to "get rid of these," the **image** is the subset of B that contains all values that f maps values of A onto.

Definition 0.5: Image of a Function

Let f be a function from A to B. The **image** of f is all the output values f may produce.

Example 1. Let f be defined from $(-\infty, \infty)$ to $(-\infty, \infty)$ defined by $f(x) = 0$. The image of f is the set of output values of f; since 0 is the only output value of f, the image of f is the set $\{0\}$.
Example 2. Let f be defined from $(-\infty, \infty)$ to $[0, \infty)$ defined by $f(x) = x^2$. The image of f is the set of output values of f. Since x^2 has output values $[0, \infty)$, the image of f is its co-domain.
Example 3. Let f be defined from $[10, \infty)$ to $[0, \infty)$ defined by $f(x) = x^2$. The image of f is the set of output values of f. While this looks very similar to the previous example, there is one key difference: the domain is now restricted to be greater than or equal to 10. Because of this, f is always greater than or equal to $10^2 = 100$. So, the image of f is $[100, \infty)$.

The distinction between the image and co-domain may seem quite tedious, but it is crucial in defining other ideas. Suppose someone is traveling from the United States to Japan to India. From the United States to Japan, the traveler exchanges the US dollar to Japanese yen; for each amount in US dollars, there is one and only one equivalent in Japanese yen and all US dollar conversions are considered. So this conversion rate is a function. From Japanese yen to Indian rupees, there is another conversion rate, which is also a function. So, the output of the US-Japan function is the input of the Japan-India function. Such an idea is referred to as the **composition of functions**.

Definition 0.6: Function Compositions

A **function composition** is an operation that takes two functions f and g and creates a new *function* h such that $h(x) = g(f(x))$. In other words, the output of one function is the input of another. From this, it is easy to recognize that the domain of h is equivalent to the domain of f and the co-domain of h is the same as the codomain of g. This is often denoted as $(g \circ f)(x) = g(f(x))$.

Because the output of one function is the input of another, key restrictions must be placed. Namely, in order for the composition of functions to be a function itself, the image of the input function must itself be a subset of the domain of the output function. So for example, suppose a function f is to be the input function of the function g. Then, the composition of these functions would be $g(f(x))$. In order for $g(f(x))$ to be a function, all of the values in its domain must map to one and only one value in its co-domain. Suppose the image of f is not a subset of the domain of g. Then, there exists some element in f's domain, call it x_0, whose output that has no mapping in g. So, that means that $g(f(x_0))$ does not exist, and thus, $(g \circ f)(x)$ is not a function. Thus, in order for the composition itself to be a function, the image of f must be a subset of the domain of g. The following examples can further illustrate these points.
Example 4. Let f map values from $R = \{\text{a, b, c, d}\}$ to $S = \{1, 2, 3, 4\}$ defined by

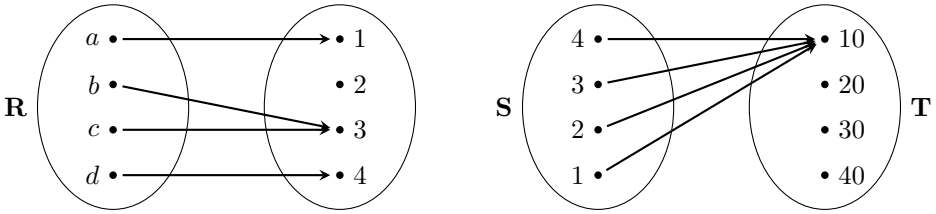
$$f(r) = \begin{cases} 1 \text{ if} & r = a \\ 3 \text{ if} & r = b \text{ or } r = c \\ 4 \text{ if} & r = d. \end{cases}$$

Let g map values from S to $T = \{10, 20, 30, 40\}$ defined by

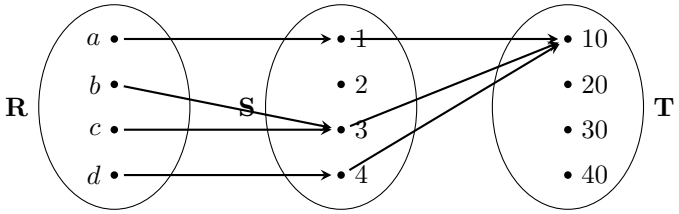
$$g(s) = 10.$$

Is it possible to create $(g \circ f)(x)$? Is it possible to create $(f \circ g)(x)$?

It would be useful to consider these functions graphically. The function f is on the left while g is on the right. Here, it is easy to recognize that the image of f is $\{1, 3, 4\}$, which is a subset of the domain of g, $S = \{1, 2, 3, 4\}$. However, the image of g, which is $\{10\}$, is not an input of f.



So, for the composition $(g \circ f)(x)$, the values in R all have somewhere to map (visually represented below). Even though the image of f and the domain of g are not equivalent, that is okay; the fact that the image of f is a subset of the domain of g is enough. Since each of the values in R is mapped to one and only one value of T, $(g \circ f)(x)$ is a valid function. On the other hand, because the image of g is not a subset of the domain of f (since 10 is not an element of R), it is not possible to create the composition $(f \circ g)(x)$.



Derivatives

In high school algebra, students often discuss *rates of change*, a concept that defines how a change in one quantity relates to a change in another quantity. This idea is linked very closely to sets and functions as the most common use of rates of change is with functions: how does a change in inputs impact the output of a function? In earlier math classes, the phrase "rise over run" is an intuitive way of calculating rates of change. The "run"

Definition 0.7: Rate of Change

Suppose f is a function. Let x_0 and x_1 be elements of the domain of f. The rate of change of $f(x)$ from x_0 to x_1 is then defined as

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

since that is change in f (i.e., $f(x_1) - f(x_0)$) with respect to the change in x (i.e., $x_1 - x_0$).

This is perhaps best explained through an example.

Example 5. Let f be a function from $[0, 1]$ (i.e., all the values between 0 and 1, inclusive) to $[0, 2]$ defined by

$$f(x) = 2x.$$

So for example, $f(0.5) = 2(0.5) = 1$. In other words, the output is just equal to twice the input.

This is a linear function, so the rate of change is equivalent to the slope. Intuitively, since 2 is the coefficient on x, the rate of change is 2. However, we can check this using definition 0.7. Let $x_0 = 0$ and $x_1 = 1$. The

choice of values is arbitrary since the function is linear, so all that matters is that x_0 and x_1 both belong to the domain of f. Then, the rate of change of f is

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{2-0}{1-0} = \frac{2}{1} = 2,$$

which is also the slope of $f(x)$.

The Derivative

Notice that the formula outlined in Definition 0.7 works for arbitrary x_0 and x_1 values, *unless* $x_0 = x_1$ because it falls victim to the "divide by zero" problem (i.e., 0 can never be in the denominator). So, instead of calculating

$$\frac{f(x_0) - f(x_1)}{x_0 - x_1}, \text{ where } x_0 = x_1,$$

calculate

$$\lim_{x_1 \to x_0} \frac{f(x_0) - f(x_1)}{x_0 - x_1}.$$

Then, it is possible to find the instantaneous rate of change.

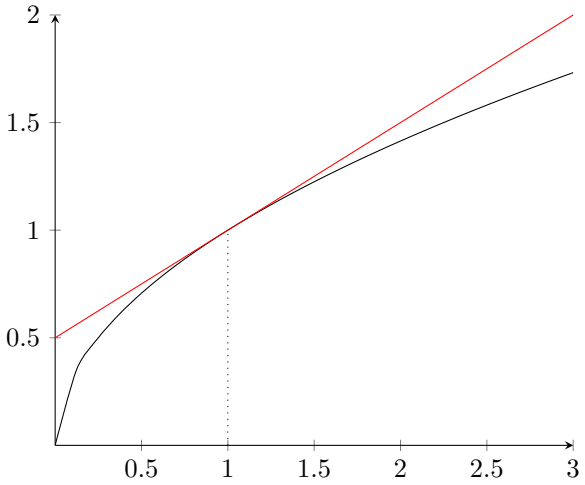
Definition 0.8: Instantaneous Rate of Change

The **instantaneous rate of change** is the rate of change at a single point x_0. For a function f, it is calculated as

$$\lim_{x_1 \to x_0} \frac{f(x_0) - f(x_1)}{x_0 - x_1},$$

often denoted as $f'(x)$ or $\frac{df}{dx}$.

Notice that there is nothing limiting this definition to only linear functions. The instantaneous rate of change can be calculated for polynomials (e.g., $f(x) = 5x^5 + 2x^3 + 9x + 6$), exponential functions (e.g., $f(x) = e^x$), logarithms (e.g., $f(x) = log_2((x))$, and several others. The instantaneous rate of change describes the slope of the line that "just touches" the function at a given point. For example, the red line in the following graph "just touches" the black line at the dotted line; this is the instantaneous rate of change/derivative.



Definition 0.9: Tangent line

Geometrically, the derivative at a point x_0 is interpreted as the rate of change of the line that "just touches" x_0, also known as the **tangent line**. If the derivative exists, then there is only one possible

tangent line.

Unlike the rate of change calculation in Definition 0.7, the derivative has two restrictions:

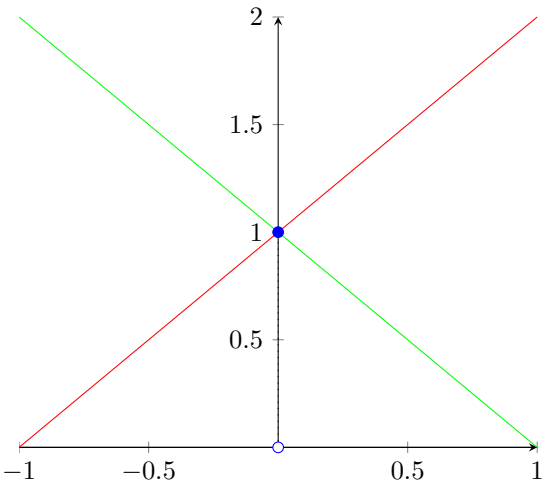
1. The function must not have an abrupt change in values at the point of interest (i.e., the function must be continuous at the point for which one wants to find the instantaneous rate of change)
2. The function at the point of interest cannot be a cusp or kink

To illustrate the need for these, consider the following examples.

Example 6. Consider the function g from $(-\infty, \infty)$ to $(-\infty, \infty)$ defined by

$$g(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, this function is not continuous at $x = 0$ because there is an abrupt change in values. Intuitively, consider what the tangent line may look like at $g(0)$. Because $g(0)$ is a single point, both the red and green lines "just touch" the blue dot, so $g(0)$ has more than one tangent line, which means its instantaneous rate of change is not unique and therefore does not have a derivative.



Example 7. Consider the function g from $(-\infty, \infty)$ to $(-\infty, \infty)$ defined by

$$g(x) = |x| = \begin{cases} x, & \text{if } x \geq 0 \\ -x, & \text{otherwise.} \end{cases}$$

What is the instantaneous rate of change at 0? Apply the formula from Definition 0.8:

$$\lim_{x_1 \to 0} \frac{g(0) - g(x_1)}{0 - x_1} = \lim_{x_1 \to 0} \frac{-g(x_1)}{-x_1}.$$

Consider the values $-1, \frac{-1}{2}, \frac{-1}{3}, \frac{-1}{4}, \ldots$. Notice that this sequence of values goes to 0 since the numerator is fixed at -1 but the denominator increases to ∞. Let x_1 be an arbitrary value selected from this set. Since all the values in this set are less than 0, $g(x_1) = -x_1$. So,

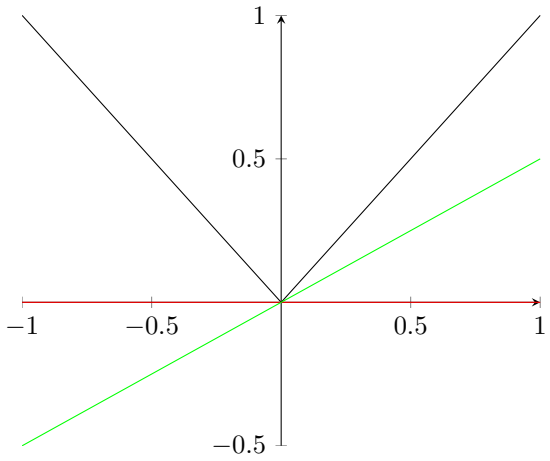
$$\begin{aligned} \lim_{x_1 \to 0} \frac{-g(x_1)}{-x_1} &= \lim_{x_1 \to 0} \frac{-(-x_1)}{-x_1} \\ &= \lim_{x_1 \to 0} \frac{x_1}{-x_1} \\ &= \lim_{x_1 \to 0} -1 \\ &= -1, \end{aligned}$$

so the instantaneous rate of change must be -1. However, now consider the values $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$. This sequence of values als goes to 0 since the numerator is fixed at 1 but the denominator increases to ∞. Let x_1 be an arbitrary value selected from this set. Since all the values in this set are less than 0, $g(x_1) = x_1$. So,

$$\begin{aligned}\lim_{x_1 \to 0} \frac{-g(x_1)}{-x_1} &= \lim_{x_1 \to 0} \frac{-x_1}{-x_1} \\ &= \lim_{x_1 \to 0} 1 \\ &= 1,\end{aligned}$$

so the instantaneous rate of change must be 1. The instantaneous rate of change cannot be two values at the same time, so it does not exist. This is because g is a kink at 0.

Geometrically, there can again be multiple tangent lines with different rates of change at $g(0)$, so no derivative exists for $g(0)$.



If a function f is continuous and has no kinks across its domain, often times, data scientists define a function f' that maps all the values in the domain to the instantaneous rate of change at each point.

Definition 0.10: Derivative of a Function

Let f be a function from some set A to $(-\infty, \infty)$ that is both continuous and has no kinks. Then, if x is an element of the set A, there exists an instantaneous rate of change for x. So, the **derivative of a function** is a function, denoted as f', itself from A to $(-\infty, \infty)$ that maps each value x in A to the instantaneous rate of change of $f(x)$, often denoted as f'.

Derivative "Shortcuts"

While the derivative definition is often very useful for solidifying theoretical interpretations of the derivative, using the limit for actual calculations is very tedious. Mathematicians discovered derivative rules that are often used in calculus. Proving these rules is left to the reader.

Theorem 0.1: Derivative Rules

Let f and g be differentiable functions from some set A to $(-\infty, \infty)$. Then,

1. If f is a polynomial of order n (i.e., $f(x) = x^n$, then $f'(x) = nx^{n-1}$
2. If f is of the form $f(x) = e^x$, then $f'(x) = e^x$
3. If f is of the form $f(x) = ln(x)$, then $f'(x) = \frac{1}{x}$

4. $(f+g)'(x) = f'(x) + g'(x)$: the derivative of a sum of functions, is the sum of the functions' derivatives
5. $(cf)'(x) = cf'(x)$: the derivative of a constant times a function is the constant times the function's derivative
6. $(f \cdot g)'(x) = f'(x)g(x) + f(x)g'(x)$: the derivative of a product of functions is equal to the sum of the derivative of the first function times the second function and the derivative of the second function times the first function
7. $\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$: the derivative of the quotient of two functions is equal to the bottom times the derivative of the top minus the top times the derivative of the bottom divided by the bottom squared

Example 8. <!–
Example 9. <!–

The Chain Rule

Along with the arithmetic derivative rules comes the property of the **chain rule**, a formula to compute the derivative of a composition of functions. While neural networks – a famous and commonly used computing system in artificial intelligence – are not discussed in this textbook, the chain rule has guided the development of research in that area of research. It is useful even in the techniques laid out here, especially in discussions on the transformation theorem.

Definition 0.11: Chain Rule

Let f and g be differentiable functions such that $f(g(x))$ exists. Because f and g are both differentiable, $(f \circ g)$ must be differentiable. Then, $(f \circ g)'(x) = f'(g(x))g'(x)$.

Example 10. Let f be a function from $(-\infty, \infty)$ to $[0, \infty)$ defined by $f(x) = x^2$. Let g be a function from $[0, \infty)$ to $[0, \infty)$ defined by $g(x) = 2x + 2$. Find $(f \circ g)'(x)$.

First, recognize that $(f \circ g)(x) = f(g(x))$ and this can be seen as a chain rule problem. So both $f'(g(x)) = \frac{d}{dx}(g(x))^2$ and $g'(x) = \frac{d}{dx}(2x + 2)$ are needed. To find $f'(g(x))$, recognize that this is a polynomial, so using the polynomial derivative rule, $f'(g(x)) = 2g(x) = 2(2x + 2)$. And to find $g'(x)$, recognize that g is the sum of two individual functions: $2x$ and 2, so $g'(x)$ can be found by summing the derivatives of $2x$ and 2, which are 2 and 0 respectively because of the polynomial rules. So, $g'(x) = 2 + 0 = 2$. Thus, $(f \circ g)'(x) = f'(g(x))g'(x) = (2(2x + 2))(2) = 4(2x + 2) = 8x + 8$.

This can also be computed without the chain rule, by taking advantage of the fact that $(f \circ g)(x) = f(g(x)) = (2x + 2)^2 = 4x^2 + 8x + 4$. Since this is function is a sum of polynomials, $\frac{d}{dx}(4x^2 + 8x + 4)$ simplifies to $8x + 8 + 0 = 8x + 8$. Thus, $(f \circ g)'(x) = 8x + 8$.

Finding Minima/Maxima

One of the key applications of the derivative is its ability to help practitioners find the points at which a function achieves minimum and maximum points. For example, consider the function $f(x) = x^2$. It is quite easy to recognize that the minimum of f occurs at $x = 0$, but for functions like $f(x) = x^6 + 4x^4 - 12x^3 + x^2 - 1$, finding minima/maxima is no longer a simple task. The derivative can help simplify this process. But first, consider the following definitions of minima and maxima.

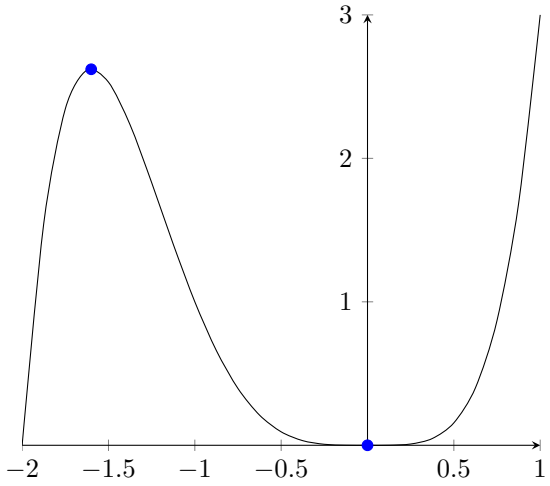
Definition 0.12: Global Minimum

Let f be a function from some set A to $(-\infty, \infty)$. If f has a **global minimum**, then there exists some element of A denoted as x_0 such that $f(x_0) \leq f(x)$ for all elements x in A.

Definition 0.13: Global Maximum

Let f be a function from some set A to $(-\infty, \infty)$. If f has a **global maximum**, then there exists some element of A denoted as x_0 such that $f(x_0) \geq f(x)$ for all elements x in A.

Often, people discuss global minima/maxima, but there are other types of minima and maxima that may be useful. Consider the function f from $(-\infty, \infty)$ to $(-\infty, \infty)$ defined by $f(x) = x^5 + 2x^4$. As x approaches $-\infty$, f goes to $-\infty$' and as x approaches ∞, f goes to ∞. So the global minimum/maximum do not exist since there is no single point x_0 such that $f(x_0)$ is less than/greater than all other points in the domain of f. However, in the following graph of f, it is quite easy to recognize that there are values that can be considered minima/maxima from $(-2, 1)$, a subset of the domain, highlighted in blue:



Such minima/maxima are referred to as **local minima/maxima**.

Definition 0.14: Local Minimum

Let f be a function from some set A to $(-\infty, \infty)$. If f has a **local minimum**, then there exists some element of A denoted as x_0 such that $f(x_0) < f(x)$ for all elements x in some connected subset of A where x and x_0 are distinct.

Definition 0.15: Local Maximum

Let f be a function from some set A to $(-\infty, \infty)$. If f has a **local maximum**, then there exists some element of A denoted as x_0 such that $f(x_0) > f(x)$ for all elements x in some connected subset of A where x and x_0 are distinct.

Finding Minima/Maxima

Consider some arbitrary function that is both continuous and un-kinked. Suppose the goal is to find a local minimum, which is known to exist. The local minimum (call it *min*) is, by definition, smaller than all the values in its neighborhood, so the function would look like a "U" around *min*. Notice that the points smaller

than *min* all must have a decreasing rate of change since they must all be decreasing to approach *min*; and the points greater than *min* all must have an increasing rate of change since they must all be increasing "away from" *min*. In other words, the points smaller than *min* must have a *negative derivative*, while the points greater than *min* must have a *positive derivative*. That means that the derivative at *min* must be 0.

However, just finding where the function's derivative is 0 is not enough to find the local minimum. Let *max* be some local maximum. By definition, *max* is greater than all the values in its neighborhood, so the function would look like a "∩" around *max*. Using a similar argument as before, this would then yield the following result: the derivative at *max* must be equal to 0.

Furthermore, *max* and *min* are not the only values at which the derivative is equal to 0. Consider the function f from $(-\infty, \infty)$ defined by $f(x) = x^3$. Graphically, it is easy to recognize at 0, $f'(0) = 0$ because the rate of change of the tangent line is 0. However, 0 is certainly not a local minimum or maximum because the points directly smaller than 0 are all less than $f(0)$ and the points directly greater than 0 are all greater than $f(0)$. Thus, $f'(0) = 0$ but 0 is not a local maximum/minimum for f. Such a point is known as a **saddle point.**

Definition 0.16: Saddle Point

Let f be a differentiable function (i.e., continuous and un-kinked). Let x_0 be an element in the domain of f. If $f'(x_0) = 0$ and x_0 is not a local minimum/maximum, then x_0 is known as a **saddle point**.

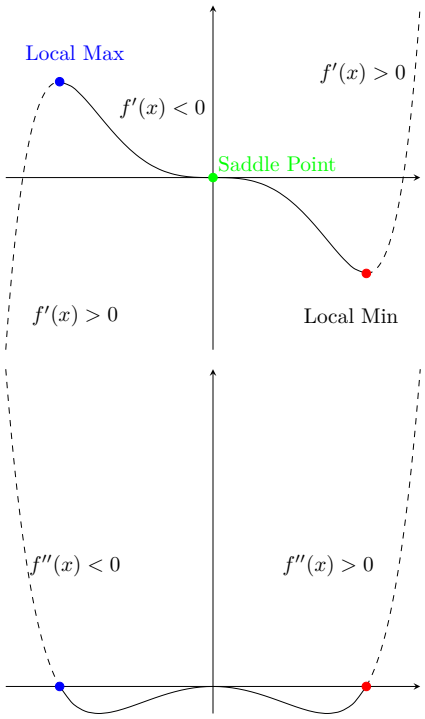
While finding the points at which the derivative is equal to 0 seems like a simple solution to finding local extrema, a little more analytical work is required. Notice that for a local maximum, the points directly to less than *max* are increasing towards *max*; in other words the derivative is positive there. Additionally, the points directly greater than *max* are decreasing away from *max*, so the derivative is negative there. On the other hand, for a local minimum, the points directly less than *min* are decreasing towards *min*, so the derivative is negative at those values. Furthermore, the points directly greater than *max* are decreasing away from *max*, so the derivative is positive there. For a saddle point, the derivative of the values directly before and after are the same sign. Using this result, one can classify local minima/maxima and saddle points.

So, in order to classify a point as a saddle point or extremum, the sign of the first derivative is used. If the derivative itself is a differentiable function, then it is possible to further simplify the process by using the **second derivative**, the derivative of the derivative.

Definition 0.17: The Second Derivative

Let f be a differentiable function. If f' is differentiable, then the **second derivative** of f is the derivative of f', often denoted as f'' or $\frac{d^2f}{dx^2}$.

When the point is a local maximum, the sign of the first derivative changes from positive to negative, so the first derivative must be decreasing; in other words, the second derivative must be less than 0. On the other hand, when the point is a local minimum, the sign of the first derivative changes from negative to positive, so the first derivative must be increasing; so, the second derivative must be greater than 0. And when the sign of the first derivative is constant, there is a saddle point. So, when the second derivative is equal to 0, there is a saddle point. Using this test, classification of points as minima/maxima/saddle points is possible.



Example 11. Let f be a function from $(-\infty, \infty)$ to $(-\infty, \infty)$ defined by

$$f(x) = -x^7 + x^5 + x^3.$$

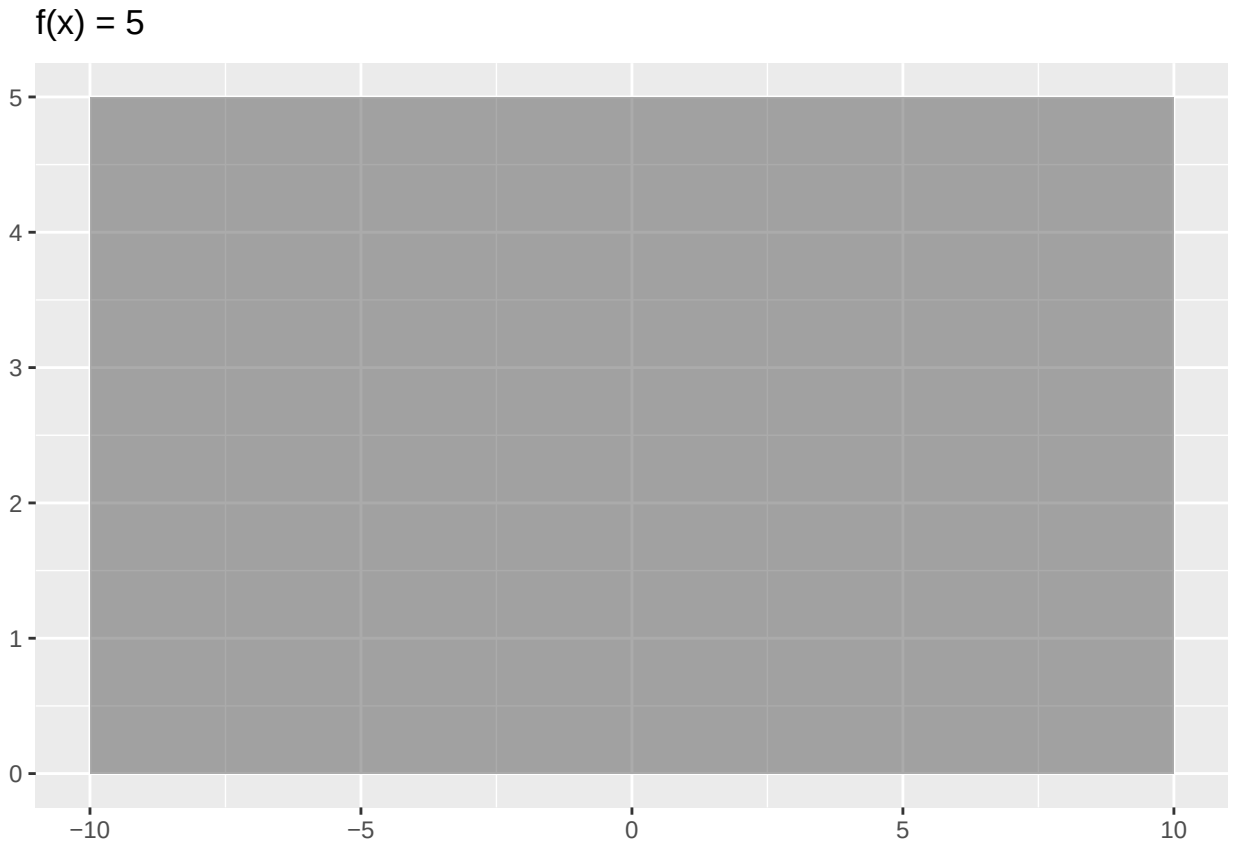
Find the local extrema and classify them as minima or maxima.

Notice that f is the sum of polynomials, $-x^7$, x^5, and x^3. So by combining the addition and polynomial arithmetic derivative rules, $f'(x) = -7x^6 + 5x^4 + 3x^2$. It is possible to find the local extrema by finding the zeros of the first derivative. It is expected that reader understands how to isolate the different values of x at which $f'(x) = 0$, so these steps are not provided. It is then found that $f'(x) = 0$ when x is one of $-1.05, 0, 1.05$. Because f' is a polynomial, it is continuous and un-kinked and is therefore differentiable. So classifying these values can utilize the second derivative. Notice, $f''(x) = \frac{d}{dx}(-7x^6) + \frac{d}{dx}(5x^4) + \frac{d}{dx}(3x^2)$. Each of these calculations utilize the fact that a constant term can be pulled out of a derivative and the polynomial rule. So, $f''(x) = -42x^5 + 20x^3 + 6x^2$. Notice, $f''(1.05) \approx -24.15 < 0$, so 1.05 is a local maximum. Additionally, $f''(-1.05) \approx 24.15 > 0$, so -1.05 is a local minimum. And $f''(0) = 0$, so 0 is a saddle point. Now, the last step is to understand whether 1.05 and -1.05 are local or global extrema; the only way they would not be global extrema is if the function tends towards $-\infty$ and ∞. So, $\lim_{x\to-\infty} f(x) \to \infty$, so -1.05 is not a global maximum. And $\lim_{x\to\infty} f(x) \to -\infty$, so 1..05 is not a global minimum. Thus, 1.05 and -1.05 are local minimum and maximum respectively.

Sums and Integration

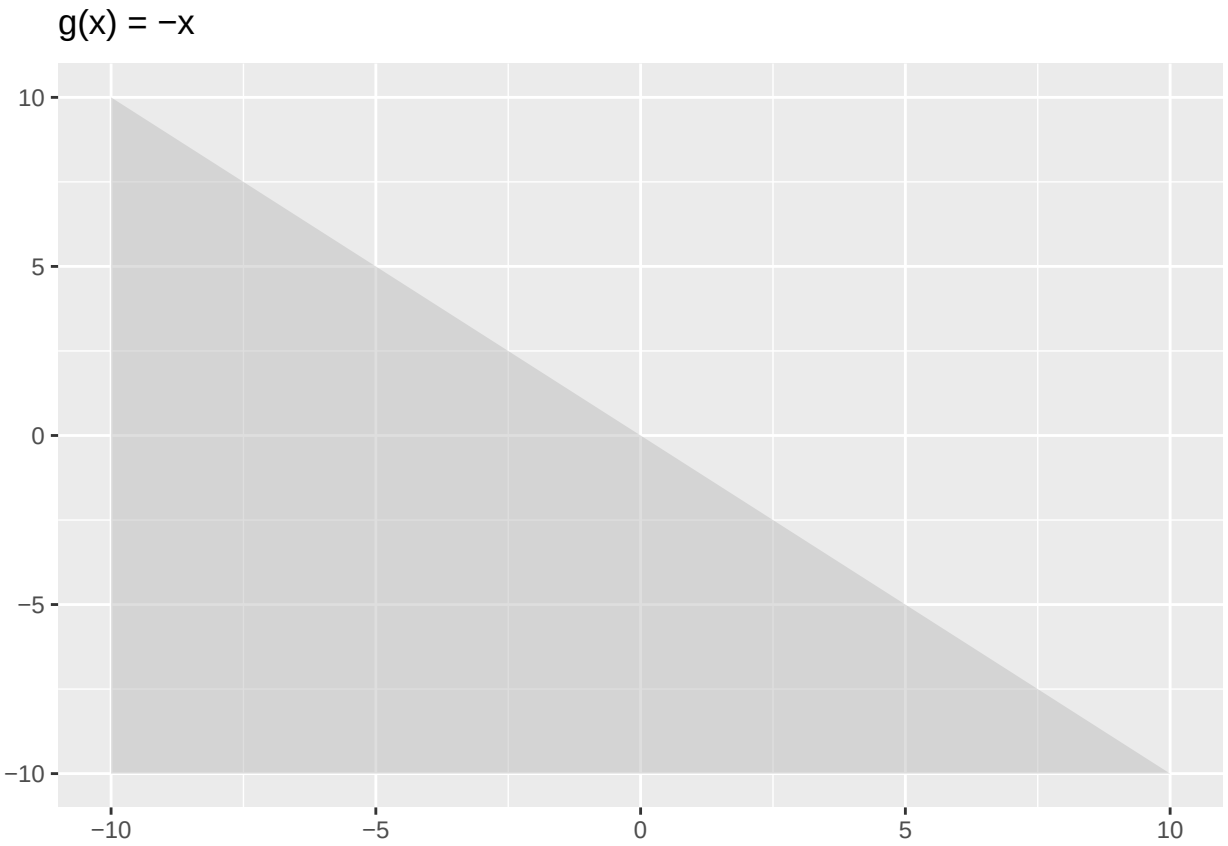
In addition to calculating rates of change, we often care about understanding areas under functions. % insert integration examples

We can start thinking about this intuitively: suppose we wanted to approximate the area under the function $f : [-10, 10] \to (-\infty, \infty)$ defined by $f(x) = 5$:

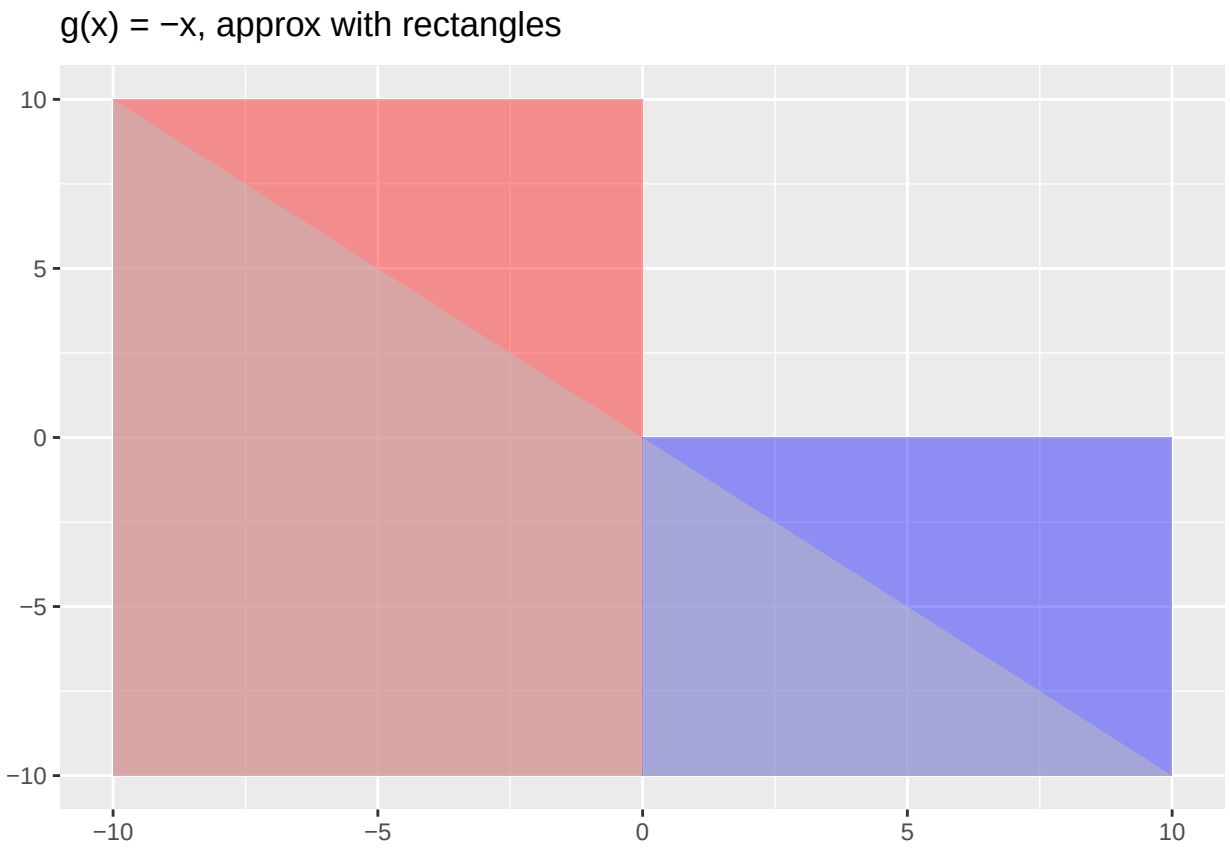


As we know from basic geometry, this is a rectangle and its area can be calculated as $base * height$. In this case, $base = 10 - (-10) = 20$ and $height = 5 - 0 = 5$, so the area under the function is $20 * 5 = 100$.

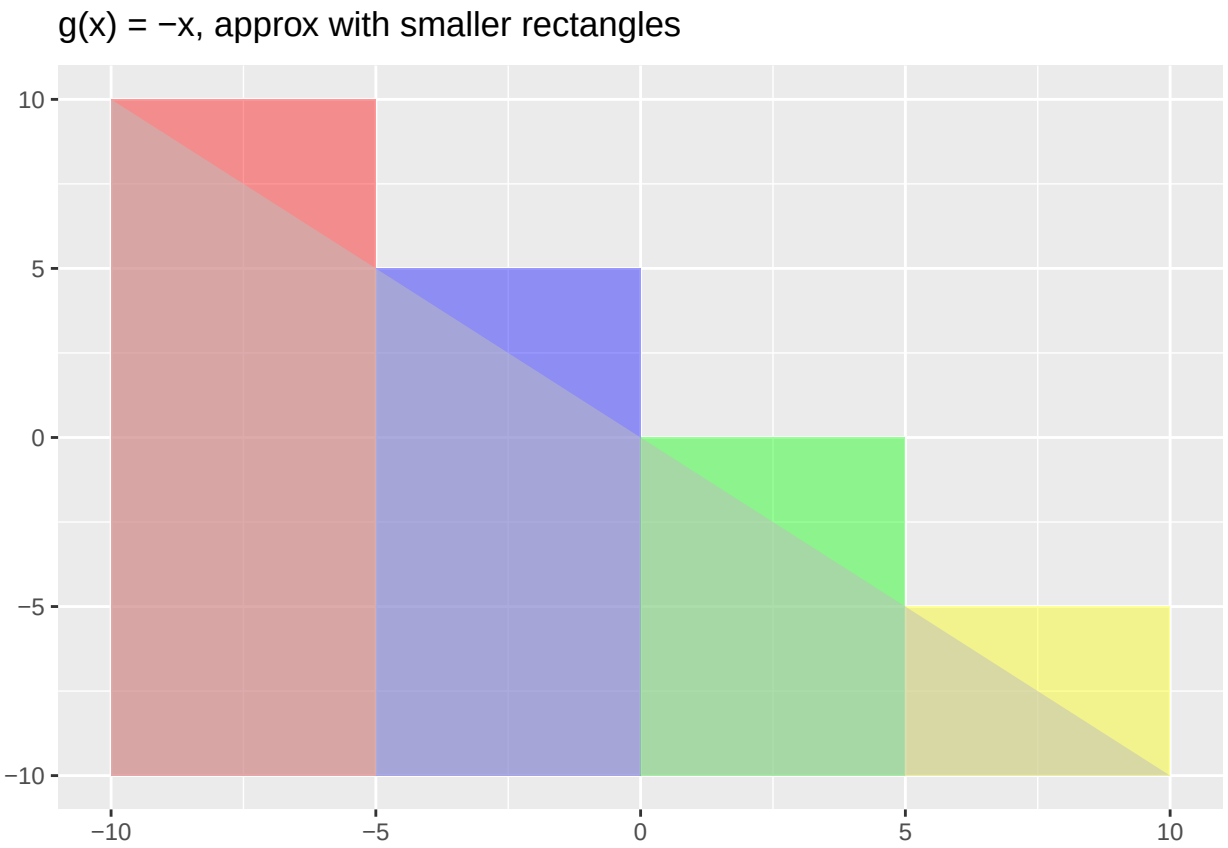
Now, consider the function $g : [-10, 10] \to (-\infty, \infty)$ defined by $g(x) = -x$:



While we know we can calculate this function's area because it is a triangle ($base * height$), we can also approximate the function's area using triangles:



The area of the red rectangle is $10 * 20 = 200$ and the area of the blue rectangle is $10 * 10 = 100$, so the approximated area of the triangle would be $200 + 100 = 300$, which is quite different than the true area of 200 However, we can get an even closer approximation by introducing more rectangles:



Here, the red rectangle has an area of $5 * 20 = 100$, the blue an area of $5 * 15 = 75$, the green an area of $5*10 = 50$, and the yellow an area of $5*5 = 25$; so the triangle's area approximation is $100+75+50+25 = 250$, which is closer to the true area of 200 than the earlier approximation with only two rectangles, suggesting that by reducing the width of the rectangles and introducing more, we can more closely approximate the area under the function.

However, let us more first define notation in order to allow us to write this information more compactly. Each rectangle will have the same width, or change in x, which we will refer to as Δx. Each rectangle's height is the value of g evaluated at the start of the rectangle. So the area of one rectangle that starts at the value a between -10 and 10 would be $\Delta a * g(a)$. To approximate the area of the function, we would then sum the areas of each individual rectangle. Suppose we are using natural number (i.e., 1, 2, 3, ...) n number of rectangles to approximate the function. Then the width of each rectangle would be $\frac{10-(-10)}{n} = \frac{20}{n}$.

Then, the area of the triangle's approximation would be,

$$\frac{20}{n}g(-10) + \frac{20}{n}g(-10 + \frac{20}{n}) + \frac{20}{n}g(-10 + 2\frac{20}{n}) \ldots \frac{20}{n}g(-10 + (n-2)\frac{20}{n}) + \frac{20}{n}g(-10 + (n-1)\frac{20}{n}).$$

Notice, we can actually factor out $\frac{20}{n}$ since it is common in each term, which becomes

$$\frac{20}{n}\left[g\left(-10\right) + g\left(-10 + \frac{20}{n}\right) + g\left(-10 + 2\frac{20}{n}\right) \ldots g\left(-10 + (n-2)\,\frac{20}{n}\right) + g\left(-10 + (n-1)\,\frac{20}{n}\right)\right].$$

In mathematics notation, we compress the summation using the $\sum$ symbol. $\sum_{i=1}^{n-1} i$ reads as "take the sum from i equals 1 to i equals n of i." So i becomes an indexing variable. The summation from the earlier equation now simplifies to

$$\frac{20}{n}\sum_{i=0}^{n-1} g\left(-10 + \frac{20i}{n}\right),$$

which reads as "multiply the quotient of twenty divided by n by the sum from i equals 1 to $n-1$ of g of negative ten plus twenty times i divided by n." As we discussed earlier, $\frac{20}{n}$ is simply the width of all of our rectangles, and because it is shared by all rectangle area calculations, we can factor it out; additionally, $-10+\frac{20i}{n}$ means that we shift the starting point of the rectangle over by $\frac{20}{n}$ at each calculation; and $g\left(-10+\frac{20i}{n}\right)$ is simply the height of each rectangle.

As we increase n, we approximate the area of the triangle more and more closely. Since we discussed limits already, we can apply a similar idea and get an exact solution for the area by calculating the limit of the sum as $n \to \infty$:

$$\lim_{n\to\infty} \frac{20}{n} \sum_{i=0}^{n-1} g\left(-10+\frac{20i}{n}\right).$$

We can generalize the idea of approximating areas for any function with a bounded domain. Suppose we have a function $f : A \to B$ with sets A and B. Assume that A is bounded below by real number a and above by real number $\bar{a}$. Then the area under f will be

$$\lim_{n\to\infty} \frac{\bar{a}-a}{n} \sum_{i=0}^{n-1} f\left(a+\frac{\bar{a}-a}{n}i\right).$$

Recall that the goal of increasing n is to reduce the width of each of our rectangles, so we can actually rewrite this formula using Δx, the density of the rectangles:

$$\lim_{\Delta x\to 0} \Delta x \sum_{i=0}^{\frac{\bar{a}-a}{\Delta x}} f(a+i\Delta x).$$

The idea of using an infinite number of rectanlges to approximate the area under the curve of a bounded interval is known as a Reimann Integral.

Definition 0.18: Reimann Integral

The **Reimann integral** is an approximation of the area under the curve of a bounded function using the infinite sum of infinitesimally dense rectangles. Let $f : A \to B$, with sets A and B, be a function with A being bounded below by a and above by $\bar{a}$. Then, the integral is denoted by $\int$ and the area under the function is written as $\int_a^{\bar{a}} f(x)dx$, where dx represents the variable that should be used for the density of the rectangles.

One limitation of the Reimann integral, however, is that it can only find the area over a bounded set. There are often instances when we are interested in finding the area over infinite spaces. For instance, while we will review this in our discussions on probability, one of the basic axioms of probability is that the probability over the entire space of the probability distribution is 1. When we have a distribution whose space is $(-\infty, \infty)$, like the normal distribution, we cannot know that the probability of the entire space is 1 if we cannot calculate the area under the distribution's curve using the Reimann Integral.

To solve this problem, we often using the Lebesgue integral. While we will not cover its derivation or even the intuition because it requires its own textbook and background, it is important to understand that we are able to solve integrals for unbounded sets. This will be the integral we refer to throughout this textbook, but the notation will remain consistent with the Reimann Integral problems, as will any techniques for actually solving integrals.

Theorem 0.2: Integration Rules

Let f and g be integrable functions over a bounded interval $[a, b]$. Then,

1. $\int_a^b [f(x)+g(x)]dx = \int_a^b f(x)dx + \int_a^b g(x)dx$

2. $\int_a^b f(x)dx = -\int_b^a f(x)dx$
3. $\int_a^a f(x)dx = 0$
4. If f is a polynomial of order n (i.e., $f(x) = x^n$), then $\int_a^b f(x)dx = \frac{b^{n+1}-a^{n+1}}{n+1}$
5. If f is of the form $f(x) = e^x$, then $\int_a^b f(x)dx = e^b - e^a$
6. If f is of the form $f(x) = \frac{1}{x}$, then $\int_a^b f(x)dx = ln(|b|) - ln(|a|)$.

While derivatives and integrals may seem like disjoint ideas, they are actually importantly linked together. Intuitively, the integral represents the sum of the parts of some function. The derivative, on the other hand, represents the instantaneous changes in the function. So, the integral of the derivative will provide the sum of the minute changes of the original function, which is actually the total change in the total function. In other words, the integral and derivative can be seen as inverses of one another.

Mathematically, suppose there exists a function f that is differentiable over the interval $[a, b]$. Then,

$$\int_a^b f'(x)dx = f(b) - f(a).$$