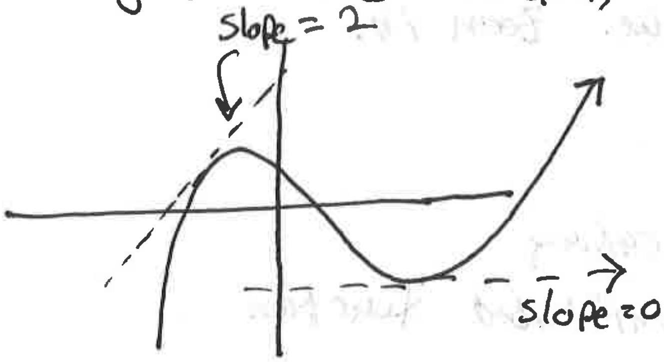
Part 1

Topic 0: Introduction

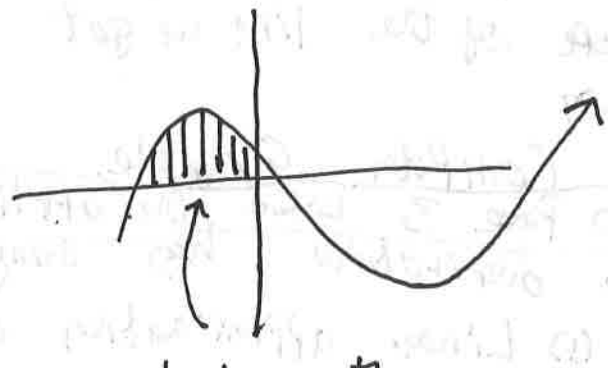
Calculus is traditionally motivated by 2 topics:

Tangent Line Problem



Q: How to compute these slopes given the function f?

Area under Curve Problem

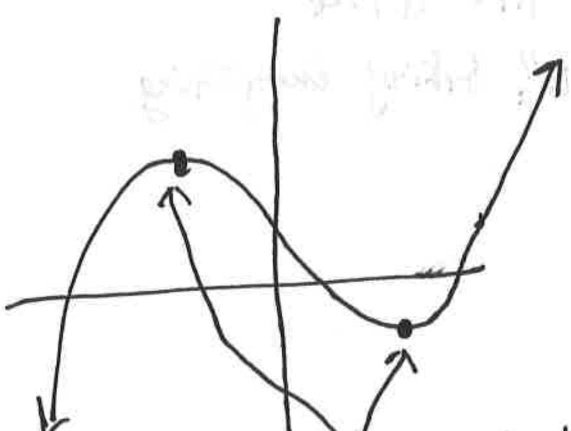


Q: What is the area of the shaded region?

Then, the climax of Calc I is the Fundamental Theoreom of Calculus that tells us ~~the know~~ that these 2 questions are intimately linked.

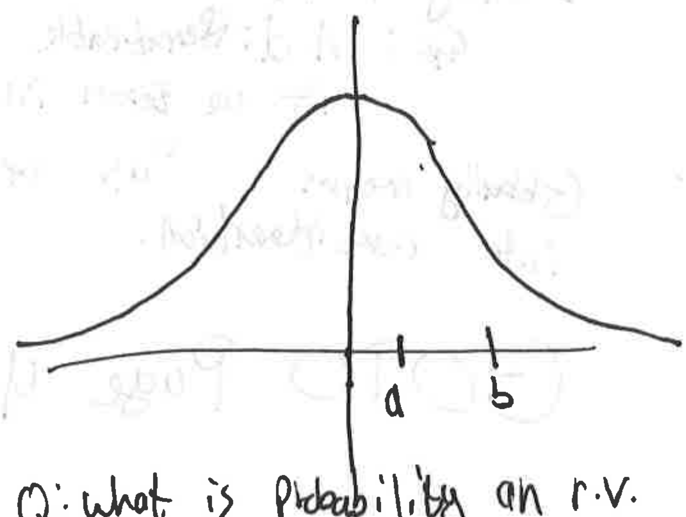
For us, we will focus on 2 different yet related questions

The optimization Problem



These points are special, how can we find them?

The continuous probability problem



Q: what is probability an r.v. is between a & b?

Those questions are essentially the same but in a different skin.

Topic 1: Derivatives

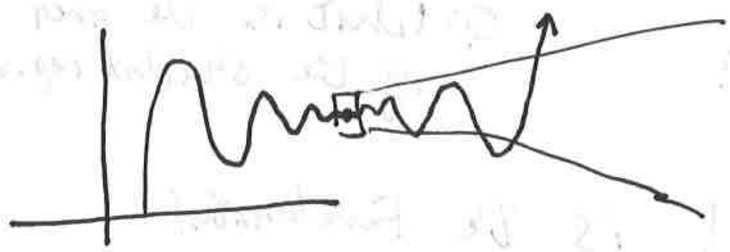
The key insight for derivatives is that, if we zoom in far enough on almost** any function, it looks like a line. The derivative is the slope of the line we get when we zoom in.

Computer Example.

Do Page 3 before the applications

The derivative has many applications

(1) Linear approximation of a complicated function



derivative is 0, at function is approximated well by a flat line.

(2) optimization:

☆ Throughout this class I will use the words locally & globally when talking about functions.

Locally means "as we zoom in".

Ex: A differentiable function is locally linear, if we zoom in it looks like a line

Globally means "as we zoom out", taking everything into consideration.

GOTO Page 4.

Part 1, Topic 1 cont.: Derivatives

This class is computational so we will not talk about limits formally. Here is how a derivative can be approximated.

Recall derivative is slope of line after we zoom in.

$$\text{Slope: } \cancel{\frac{f(x)-f(y)}{x-y}} \qquad \frac{f(y)-f(x)}{y-x}$$

Zooming in means x, y are very close. Generically, assume $y > x$ in the above. Then write $y = x+h$

$$\text{Slope: } \frac{f(x+h)-f(x)}{(x+h)-x} = \frac{f(x+h)-f(x)}{h}$$

Then, picking h very small (context dependent) can give us a good <u>estimate</u> for the derivative.

In a proper calc course we would discuss computing the actual derivative which is achieved by taking the "limit as $h \to 0$" which is "what the line looks like as we zoom in infinitely".

Simple differentiation rules to look up for those interested:

- Power Rule
- Product Rule
- chain Rule.

GOTO Page 2

If we look at the minimum or maximum of a function locally:

This is among the most common optimization methods: find somewhere that the derivative is $= 0$.

Topic 2: The first derivative test & 2nd derivative test

~~It turns out this is a theorem (something we can 100% prove is true).~~

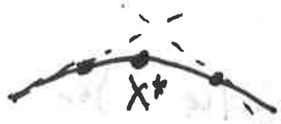
~~Theorem~~

It turns out this is always true, it is a "theorem" (something we can 100% prove is true)

Theorem If x^* is a local max/min of a differentiable function f, then the derivative of f at x^* is 0.

Notice that we ~~may be~~ cannot differentiate between max/min by knowing the derivative is 0.

1st derivative test:

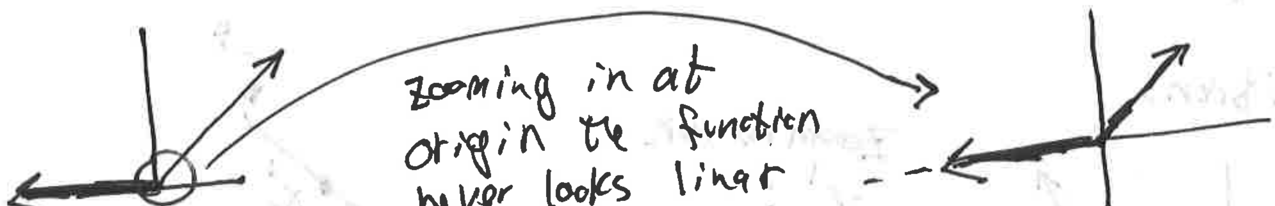
max:  slope on left is >0, slope on right is <0

min:  slope on left is <0, slope on right is >0.

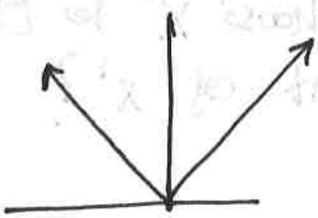
$\boxed{\text{Be Careful!!}}$

Some functions are not differentiable at every point:

Ex 1: $ReLU(x) = \max(0, x)$

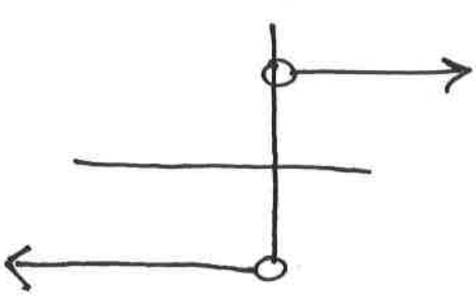


Ex 2: $f(x) = |x|$



Again, this map is not differentiable at $x=0$ because no matter how far in we zoom it never looks linear.

If we graph the derivative of $|x|$:



we see it is never 0 but our original function, $|x|$ still has a minimum at $x=0$

~~Topic 3: 1-dimensional gd~~

Related to Topic 2 Further reading
- 2^{nd} derivative test
- critical points

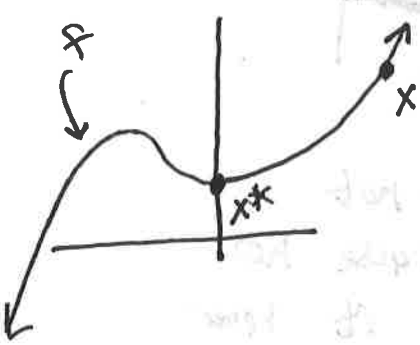
Topic 3: 1-dimensional gradient descent.

Sometimes (almost always actually) we can't solve for when the derivative $= 0$. We need another method.

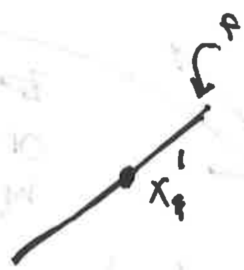
An iterative algorithm is one which repeats a number of steps to compute or approximate a solution.

An iterative optimization algorithm is one which generates a series of points $x^1, x^2, \ldots x^T$ that get closer to minimizing a function f.

Intuition:



Zoom in on x^1:

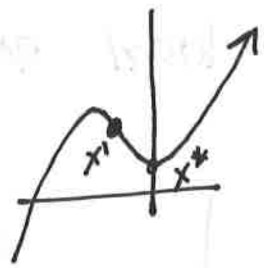


Our goal is to minimize f, should we choose x^2 to the left or right of x^1?

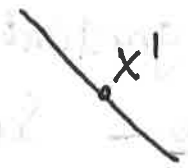
A: Set $x^2 < x^1$
or $x^2 = x^1 - d$ for some $d > 0$

Topic 3 cont.

Similarly:



Zooming in:



we will set $x^2 > x^1$

so $x^2 = x^1 + r$, $r > 0$

$x^2 = x^1 - (-r)$

The differences in the scenarios is the slope at x^1

- negative slope → go right
- positive slope → go left

The slope at x^1 is denoted $f'(x^1)$. We can set in both cases $x^2 = x^1 - \eta f'(x^1)$ (Check both cases yourself here!) where η determines how far we go!

Computer Example

This is all that gradient descent is. In higher dimension we need a more sophisticated notion of direction than left / right. More Later

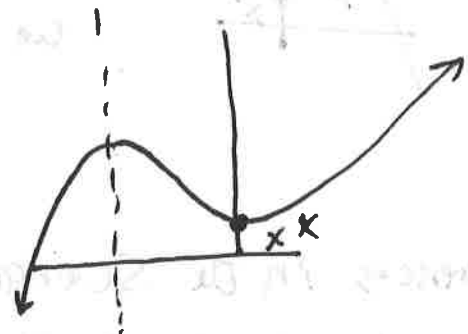
Topic 4: Uniqueness, Convexity, local v. global optimums

When using derivatives we are only looking at a function locally. We cannot tell what happens far away from our current point.

Thus, any gradient (derivative) based method is ~~assumed to~~ said to find only local extrema (maximums / minimums).

Prev. Example

starting anywhere to the right of the dotted line will converge to x^*



which is not the minimum of the function (it is a local min).

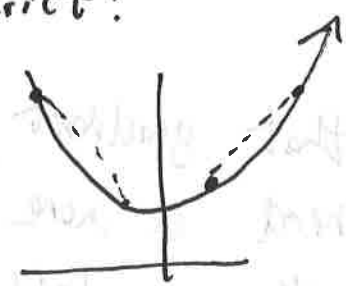
In general there is no way to combat this problem. However, if our function is $\boxed{\text{convex}}$ we will not have this problem.

Def A function f is said to be convex if every secant line to the graph lies above the graph itself. Strict?

connects 2 points on the graph

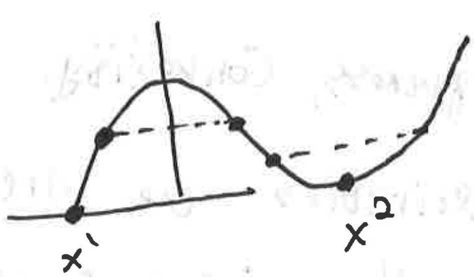
Example:

This is convex



Non-example

Notice: x^1, x^2 are both mins



Topic 4:

Time-Permitting Prove the following (it is Easy!)

Theorem If a function is convex, then Every local min is a global min as well. Moreover, if f is strictly convex this min is unique.

Proof hint: Secant line above graph if for Every x_1, x_2

$$\underbrace{f(tx_1 + (1-t)x_2)}_{\text{graph between } x_1, x_2} \leq \underbrace{t f(x_1) + (1-t) f(x_2)}_{\text{Equation for secant line.}}$$

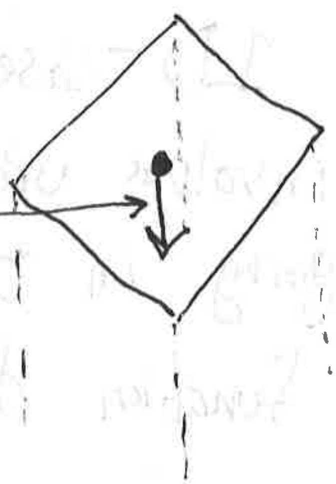
Topic 5: derivatives in higher dimensions:

when our graph is a surface it no longer makes sense to talk about tangent lines.

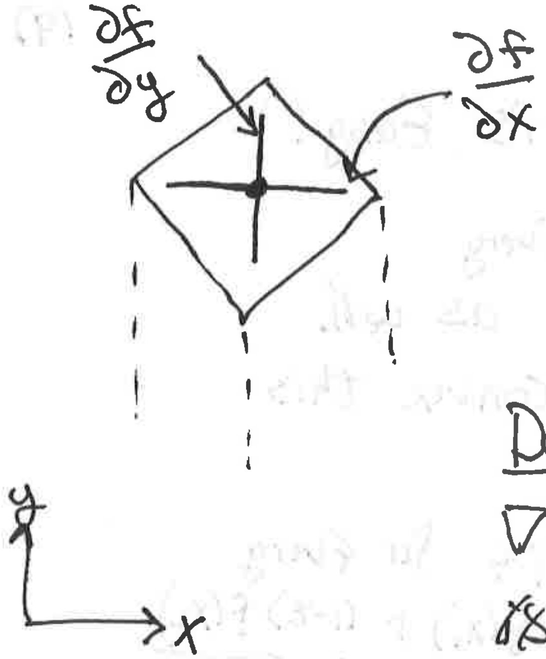
- Drawing in class, Computer Example? → matlab?
  ↳ ~~Neural Nets~~

For functions of 2 variables, differentiability is the same as zooming in & the function looking like a plane:

direction of "steepest descent"



we can talk about derivatives along a line (see next)



These <u>partial</u> <u>derivatives</u> tell us the slope in the x, y direction

<u>Def.</u> The gradient vector $\nabla f(x_0, y_0)$ of the function f ~~is the~~ at the point (x_0, y_0) is the vector $\left[\frac{\partial f}{\partial x}(x_0,y_0), \frac{\partial f}{\partial y}(x_0,y_0)\right]$

Understanding: The entries of the gradient tell us how "steep" our function is in a certain direction.

- Fact: The "direction" $-\nabla f(x_0, y_0)$ is the direction of "steepest descent" of the function f.

- Just like in the 1D-case, our optimization algorithm involves updating the current point by going in the "best direction" to get a function decrease.

Topic 6: Gradient Descent

- This method is called gradient descent.

The update from iterate t to $t+1$ looks like

$$x^{t+1} = x^t - \eta \nabla f(x^t)$$

where η is some parameter called the learning rate.

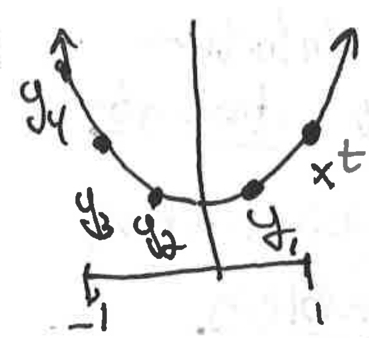
Remembering x^t & x^{t+1} are vectors that look like $(x_1^t, x_2^t, \dots x_n^t)$ & $(x_1^{t+1}, x_2^{t+1} \dots x_n^{t+1})$ then the update for each parameter individually looks like $x_i^{t+1} = x_i^t - \frac{\partial f}{\partial x_i}(x^t)$

(compare this yourself to the 1D-case).

Notes you should know about this algorithm:

1) The choice of η is extremely important & is mostly accomplished via trial & error.

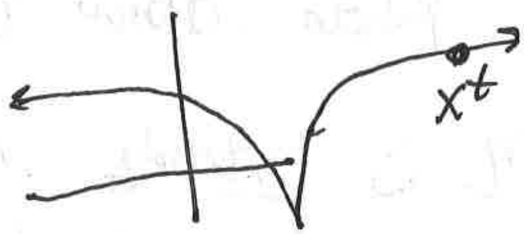
See:



starting at x^t, various values of η can result in $x^{t+1} = y_1, y_2, y_3, y_4$.

- y_1, y_2 are both "good" updates
- y_3 results in no function change
- y_4 actually gives a function increase

- We've seen a big η can be bad, but choosing η ~~too~~ too small can slow down the algorithm so much that it never gets near the minimum even though we get a "decrease" at every step.
- Since this is a derivative based method (read: local info only) we only expect to find local rather than global minima unless we assume something stronger, e.g. convexity.
- Even if you choose η well, certain functions perform very slowly using gradient descent. Since the update step size is dependent on the size of the gradient, small gradients result in slow progress. Example



Further reading for curious students:

- (1) For a "generally" better yet less intuitive method, see <u>conjugate gradient descent</u>.
- (2) For an instance of the last point above, see ~~a~~ <u>vanishing gradient problem</u>
- (3) For some functions (those whose values are fast to compute) we may use a <u>line search</u> method to optimize the choice of η at each step. See also, "<u>Hem-stitching problem</u>"

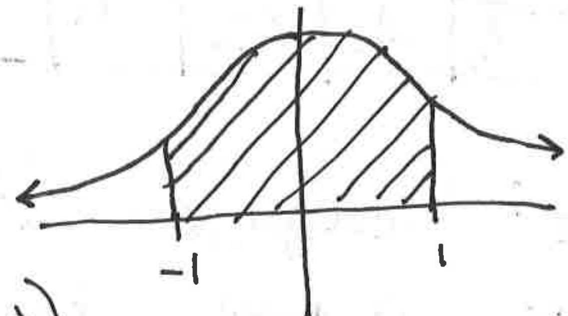
Part 2 Integration, or, Areas under curves

Topic 0: Why?

Simplest Motivation: Suppose the function $f(x)$ is a probability density function (pdf). Then, the area under the curve between a & b is the probability that X lies between a & b.

Ex. $\cancel{f(x) = \frac{1}{\sqrt{2\pi}}e}$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



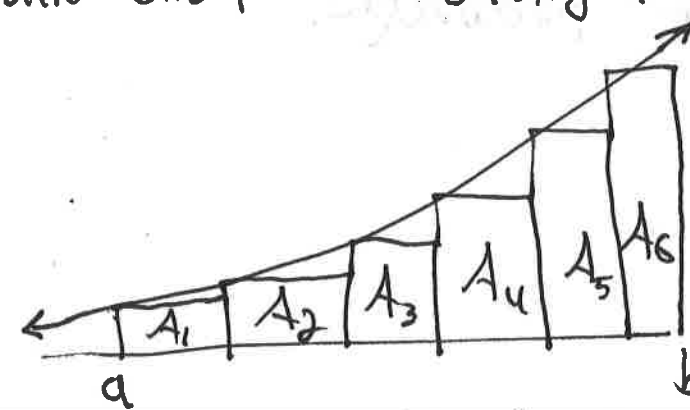
(The Standard Normal, $N(0,1)$)

The shaded area represents the probability of being within 1 s.d. of the mean (about $p = .68$)

Topic 1: First Principles

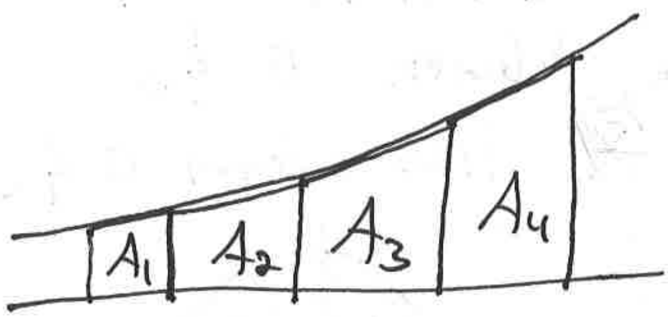
The technique used is actually quite simple, even obvious. Simply put, we approximate the area by rectangles (whose areas are easy to compute) and improve the approximation until the resulting area converges or stabilizes.

Picture:



Area under the curve is approximately $\sum_{i=1}^{6} A_i$.

- There are a variety of ways to "choose" which rectangles to use. See "Left/Right-hand Endpoints"
  ↳ Examples of both
- Other methods include Simpson's rule (more on that later) or Trapezoidal rule (uses trapezoids rather than rectangles)



$$\text{Area} \approx \sum_{i=1}^{4} A_i$$

Topic 1: Convergence (May be skipped in class for time)

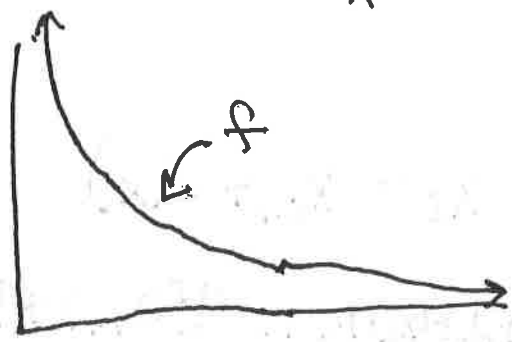
Care must, in general, be taken that the sums actually approach something (i.e. the area is finite or well-defined).

For PDF'S, none of these issues will show up but in general, we must take much care. Let us look at 3 examples. For ~~each~~ the first two, our numeric methods will be misleading, i.e. will give "areas" that may seem correct without the proper knowledge.

Topic 1, cont. : Convergence

Example 1:

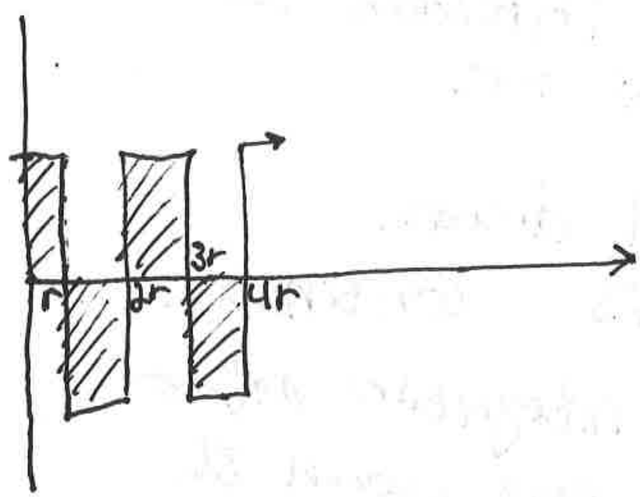
Consider $f(x) = \frac{1}{x}$ on the interval $(0,1)$.



For $0 < a < b < \infty$, the area under f between a & b is $\ln\left(\frac{b}{a}\right)$

what happens as $a \to 0$?

Example 2: $f_r(x) = \begin{cases} 1 & \text{if } \lfloor rx \rfloor \text{ is even} \\ -1 & \text{if } \lfloor rx \rfloor \text{ is odd} \end{cases}$

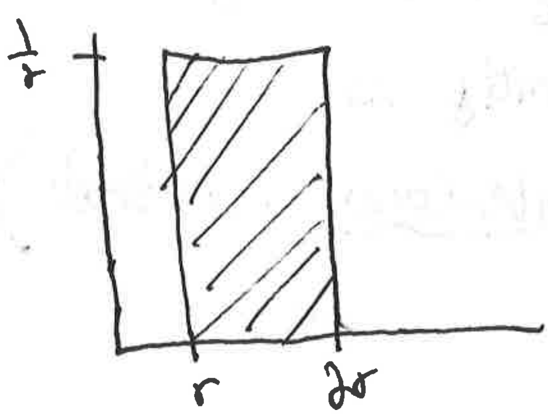


2 issues here:

1) If we take the area from 0 to ∞, what should the answer be?
2) Even on finite intervals, for small enough r, we may "accidentally" only pick rectangles with positive height

Example 3: $f_r(x) = \begin{cases} \frac{1}{r} & x \in [r, 2r] \\ 0 & \text{Else.} \end{cases}$

This $f_r(x)$ is a PDF for all $r \in (0, \infty)$



Suppose our rectangle widths are $w > 2r$ and we use left hand endpoints to determine rectangle height. Picture in class. Our estimated area will be 0.

Topic 2: Simpson's Rule ~~& Quadrature~~

Simpson's Rule

To approximate the area under $f(x)$ between a & b using n rectangles:

1) Set $\Delta x = \frac{b-a}{n}$, $x_0 = a$
2) For $i = 1$ to n

2a) Set $x_i = a + i\Delta x$ (or, $x_i = x_{i-1} + \Delta x$)

2b) Store $f(x_i)$

3) Output Area $\approx \frac{\Delta x}{3}\left(f(x_0) + 4f(x_1) + 2f(x_2) + \ldots\ 4f(x_{n-1}) + f(x_n)\right)$

This method can be seen as the average of the approximations obtained by using Trapezoidal rule and the midpoint rectangles rule.

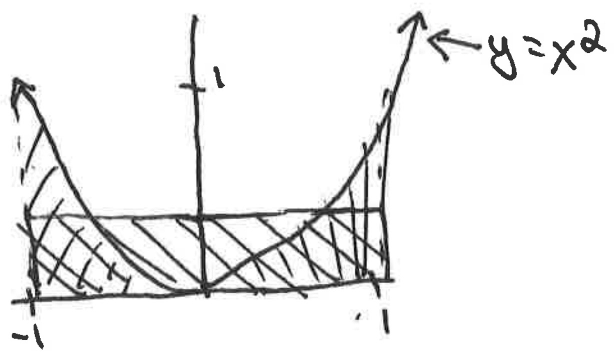
As $n \to \infty$ the error will decrease provided the function f is continuous.

Other more advanced numerical integration methods exist & the interested reader may search the term "numerical quadrature"

Topic 3: Monte-Carlo Integration

(Aside: This area is particularly interesting to me as my research currently is investigating <u>randomized</u> or <u>Monte-Carlo</u> Methods)

A common application or interpretation of integration is related to <u>average value</u> of a function. Consider: suppose the area under f between a & b is A. Then, if we place a single rectangle between a & b of height $\frac{A}{b-a}$, the area of the rectangle is A:



Here, A between -1, 1 is $\frac{2}{3}$. The shaded rectangle with height 1/3 has the same area.

We now know that $\frac{1}{3} = \frac{A}{b-a}$ is the <u>average height</u> or <u>average value</u> of f.

Another way to compute average value:

- Pick n points randomly between a & b
- Compute f at these points & take the average function value at these points

Computer Example, Everyone Try

$$f(x) = \sqrt{1-x^2} \text{ as } x \in [-1,1]$$

Answer should be ______?