Machine Learning Foundations
Calculus I: Limits & Derivatives

Using Differentiation,
including AutoDiff, in Python to
Optimize Learning Algorithms

Jon Krohn, Ph.D.

jonkrohn.com/talks

github.com/jonkrohn/ML-foundations

Machine Learning Foundations

Calculus I: Limits & Derivatives

Slides: jonkrohn.com/talks

Code: github.com/jonkrohn/ML-foundations

Stay in Touch:

jonkrohn.com to sign up for email newsletter

linkedin.com/in/jonkrohn

jonkrohn.com/youtube

twitter.com/JonKrohnLearns

The Pomodoro Technique

Rounds of:

- 25 minutes of work
- with 5 minute breaks

Questions best handled at breaks, so save questions until then.

When people ask questions that have already been answered, do me a favor and let them know, politely providing response if appropriate.

Except during breaks, I recommend attending to this lecture only as topics are not discrete: Later material builds on earlier material.

POLL

Where are you?

- The Americas
- Europe / Middle East / Africa
- Asia-Pacific
- Extra-Terrestrial Space

POLL

What are you?

- Developer / Engineer
- Scientist / Analyst / Statistician / Mathematician
- Combination of the Above
- Other

POLL

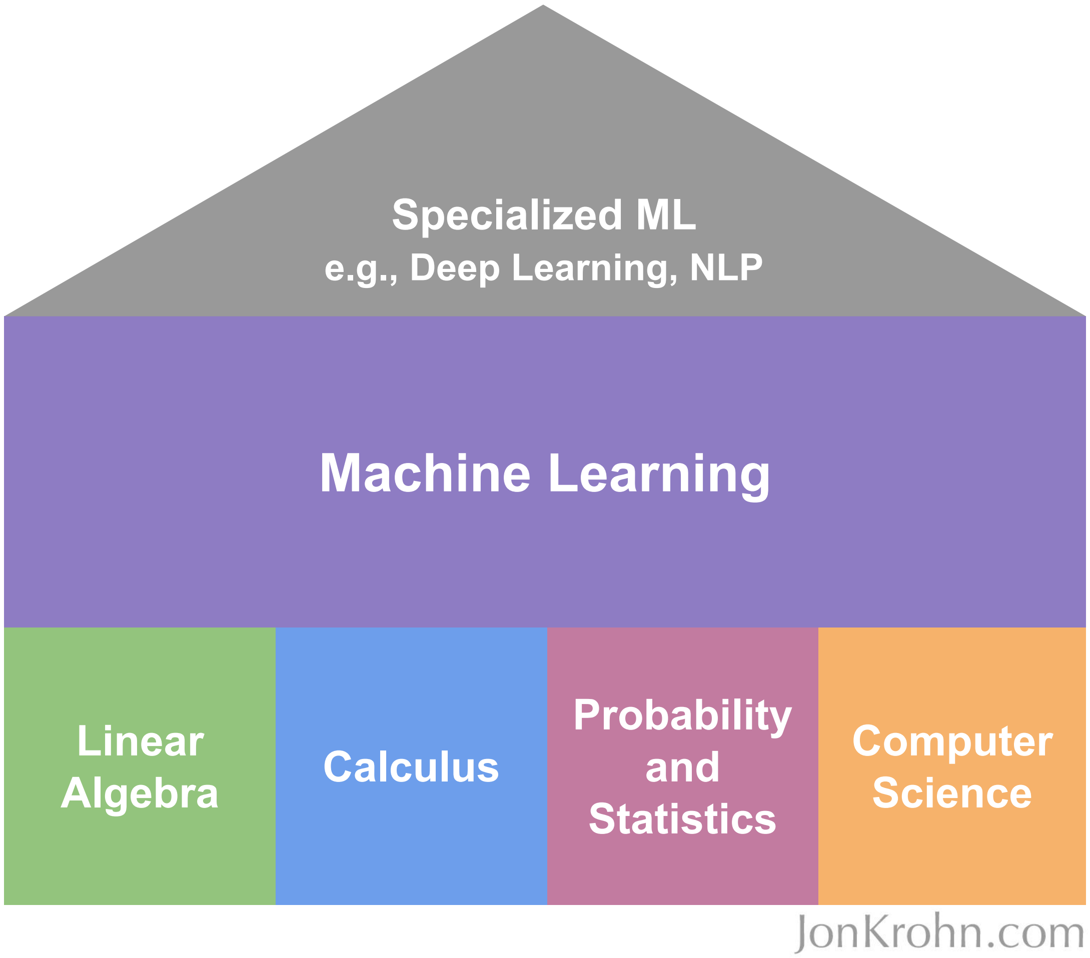
What is your level of familiarity with Calculus?

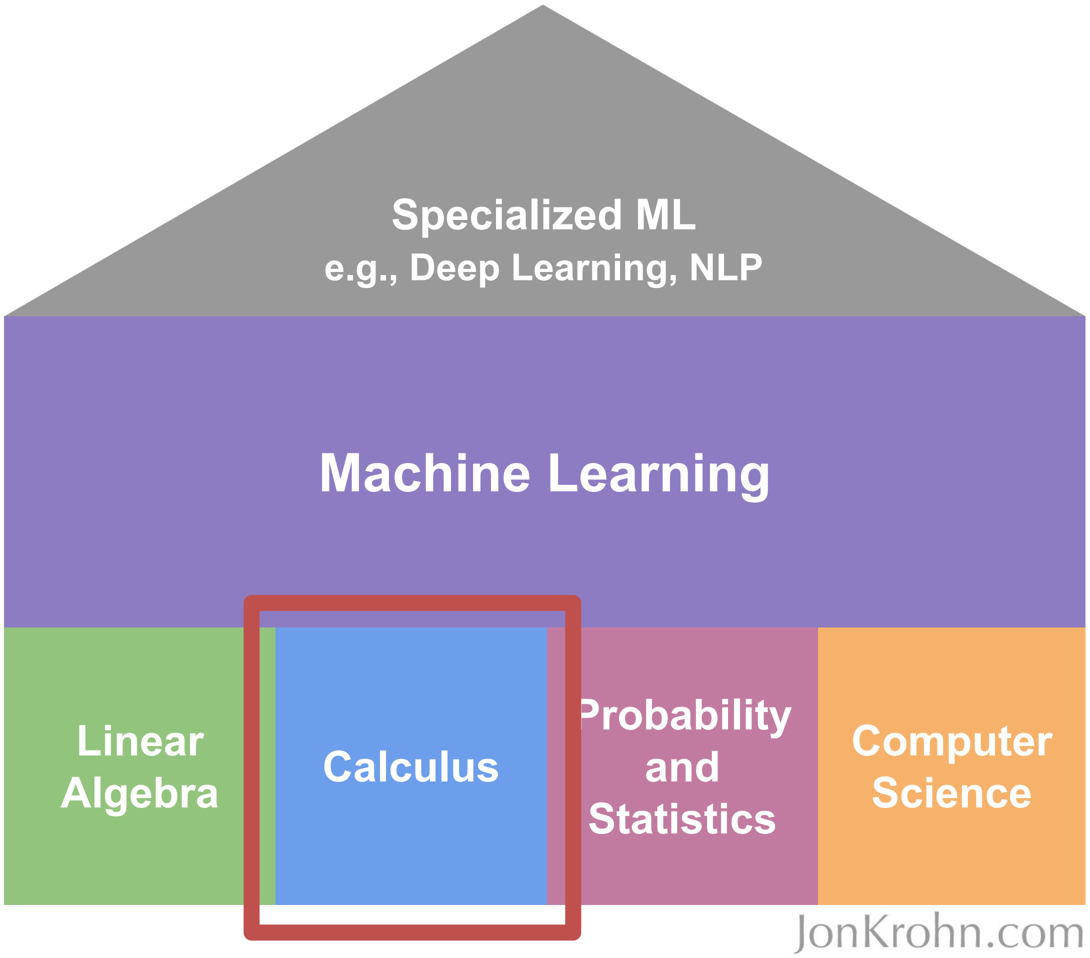
- Little to no exposure
- Some understanding of the theory
- Deep understanding of the theory
- Deep understanding of the theory and experience applying calculus operations (e.g., differentiation) with code

POLL

What is your level of familiarity with Machine Learning?

- Little to no exposure, or exposure to theory only
- Experience applying machine learning with code
- Experience applying machine learning with code and some understanding of the underlying theory
- Experience applying machine learning with code and strong understanding of the underlying theory


Specialized ML
e.g., Deep Learning, NLP
Machine Learning
Linear Algebra
Calculus
Probability and Statistics
Computer Science
JonKrohn.com


Specialized ML
e.g., Deep Learning, NLP
Machine Learning
Linear Algebra
Calculus
Probability and Statistics
Computer Science
JonKrohn.com

ML Foundations Series

Calculus I **builds upon** and is **foundational for**:

1. Intro to Linear Algebra
2. Linear Algebra II: Matrix Operations
3. **Calculus I: Limits & Derivatives**
4. **Calculus II: Partial Derivatives & Integrals**
5. Probability & Information Theory
6. Intro to Statistics
7. Algorithms & Data Structures
8. **Optimization**

Calculus I: Limits & Derivatives

1. Limits
2. Computing Derivatives with Differentiation
3. Automatic Differentiation

Calculus I: Limits & Derivatives

1. **Limits**
2. Computing Derivatives with Differentiation
3. Automatic Differentiation

Segment 1: Limits

- What Calculus Is
- A Brief History of Calculus
- The Method of Exhaustion
- Calculating Limits

What Calculus Is

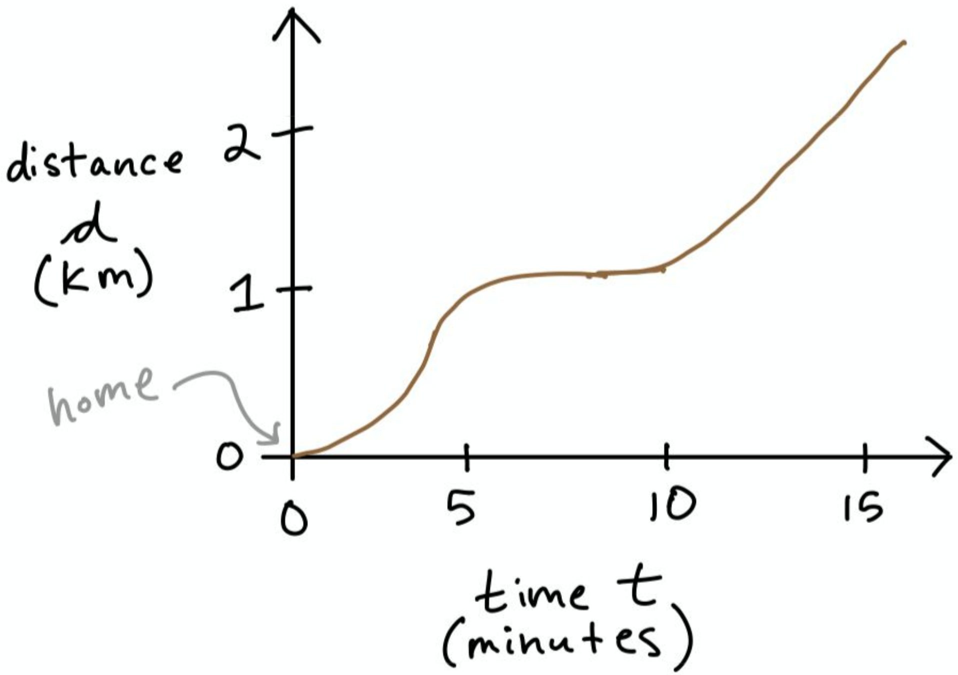
- **Mathematical study of continuous change**
- Two branches:
  a. **Differential calculus**: focus of *Calculus I*
  b. **Integral calculus**: a focus of *Calculus II* subject

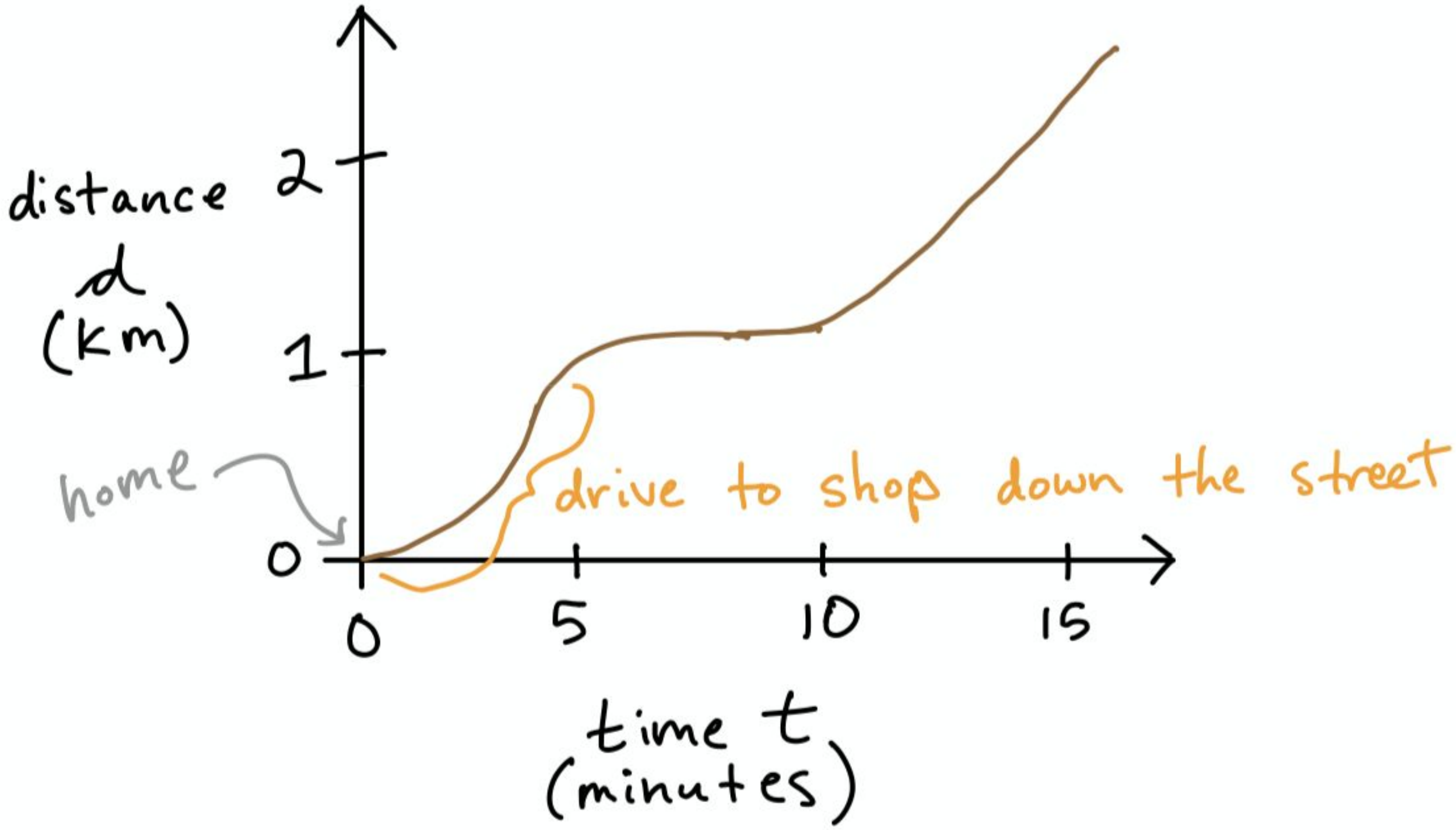
What Calculus Is

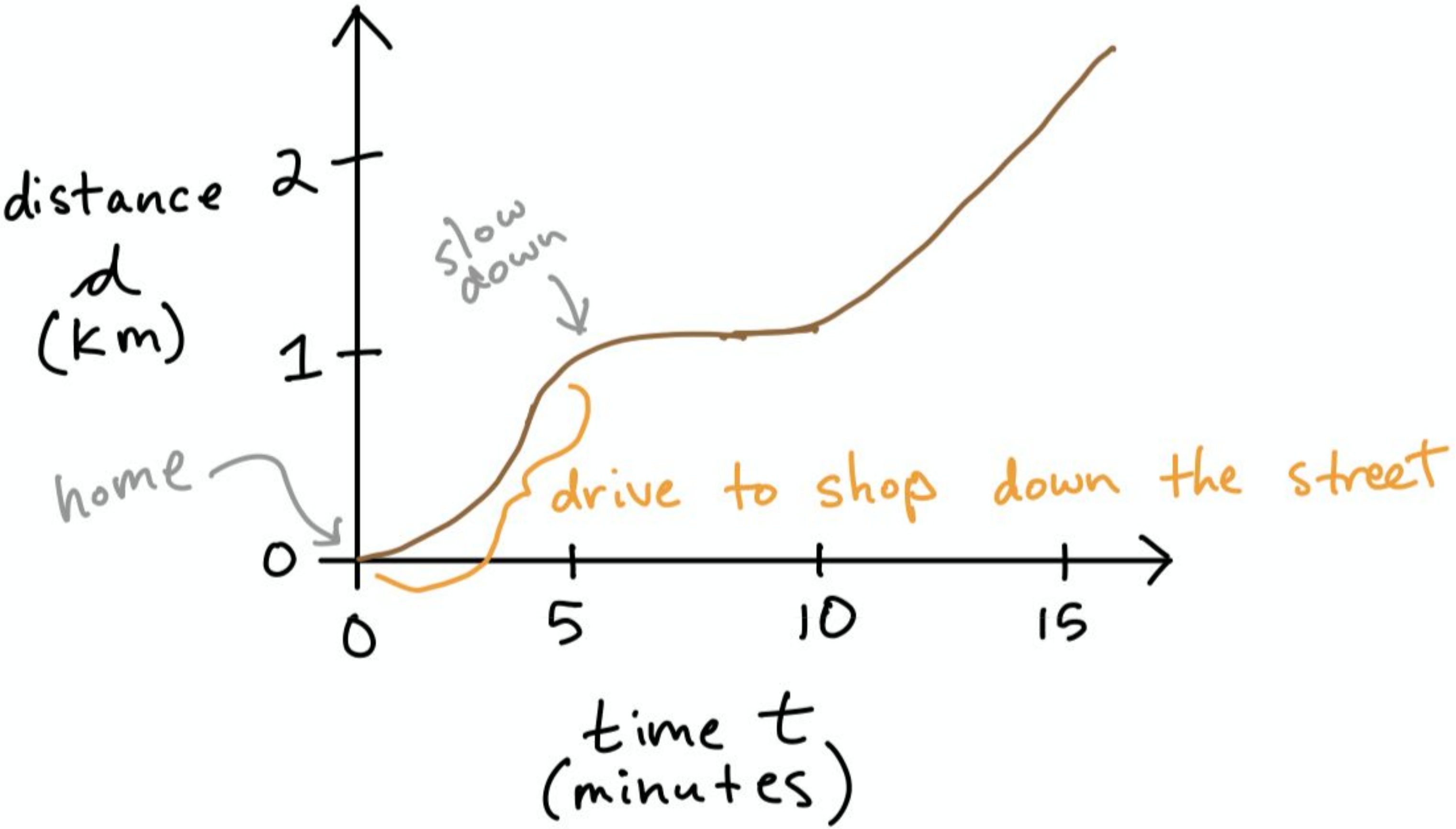
- **Mathematical study of continuous change**
- Two branches:
  a. **Differential calculus**: focus of *Calculus I*
  b. **Integral calculus**: a focus of *Calculus II* subject

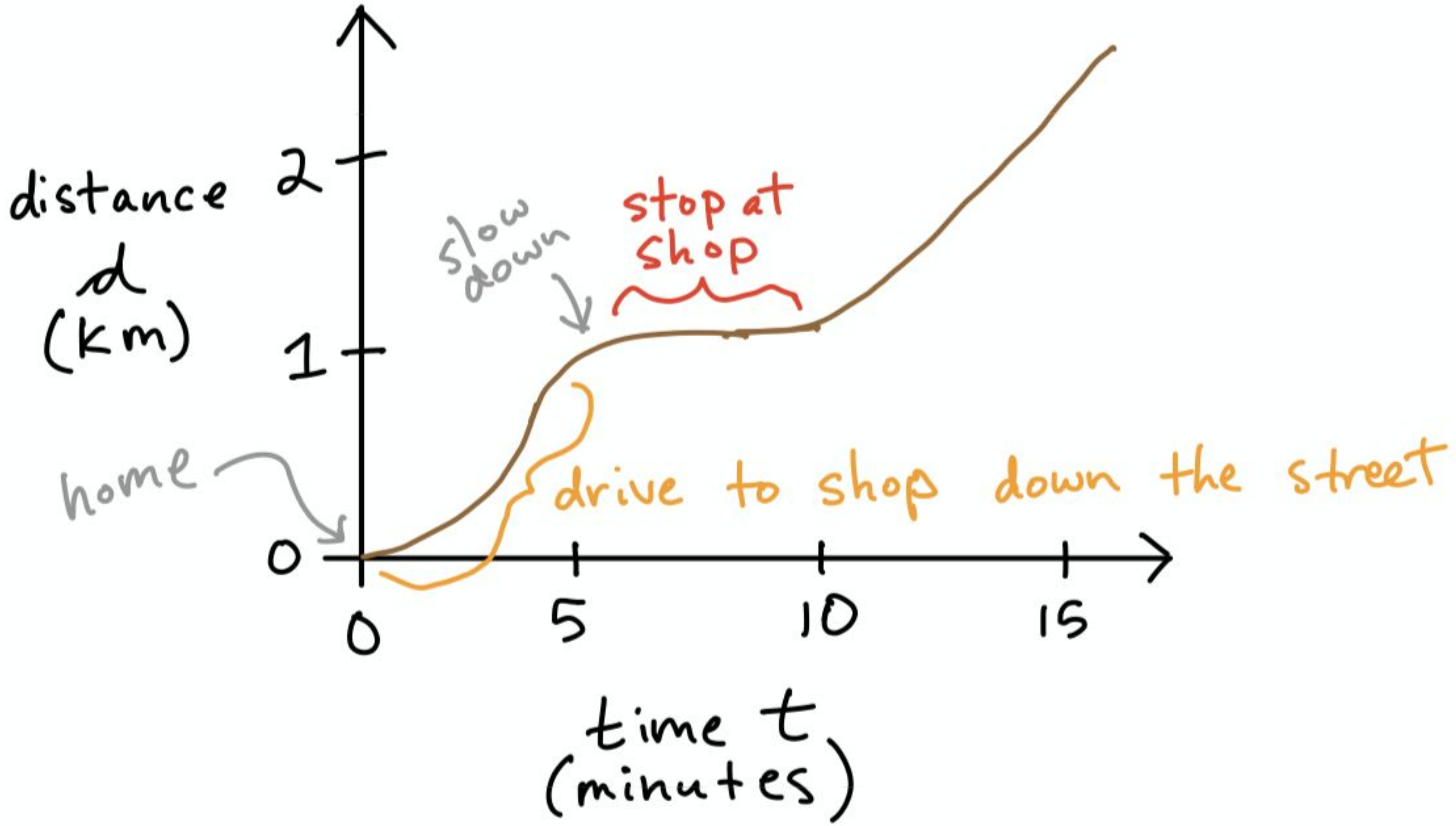
What Differential Calculus Is

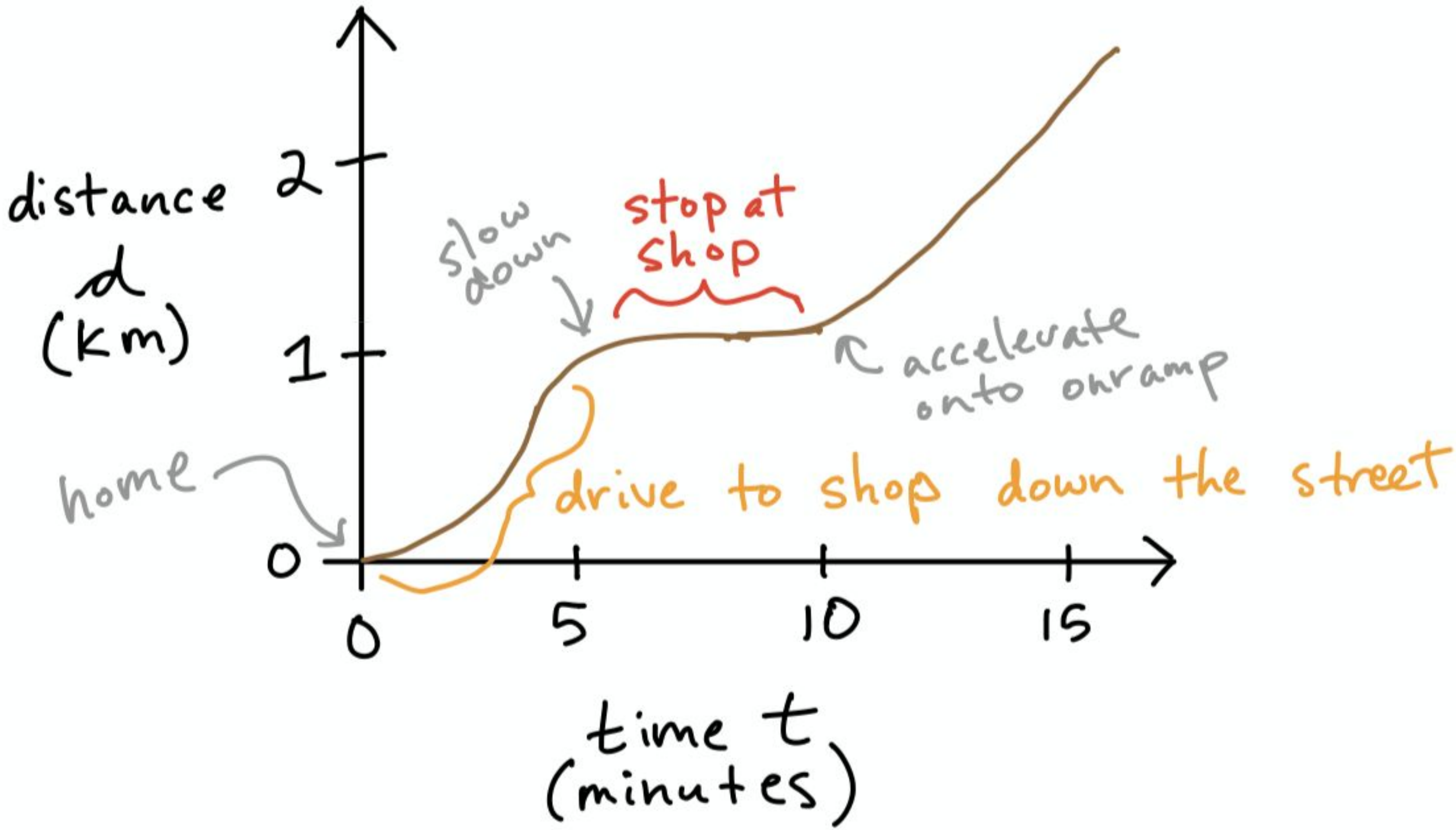
- **Study of rates of change**
- Consider a vehicle traveling some distance d over time t:

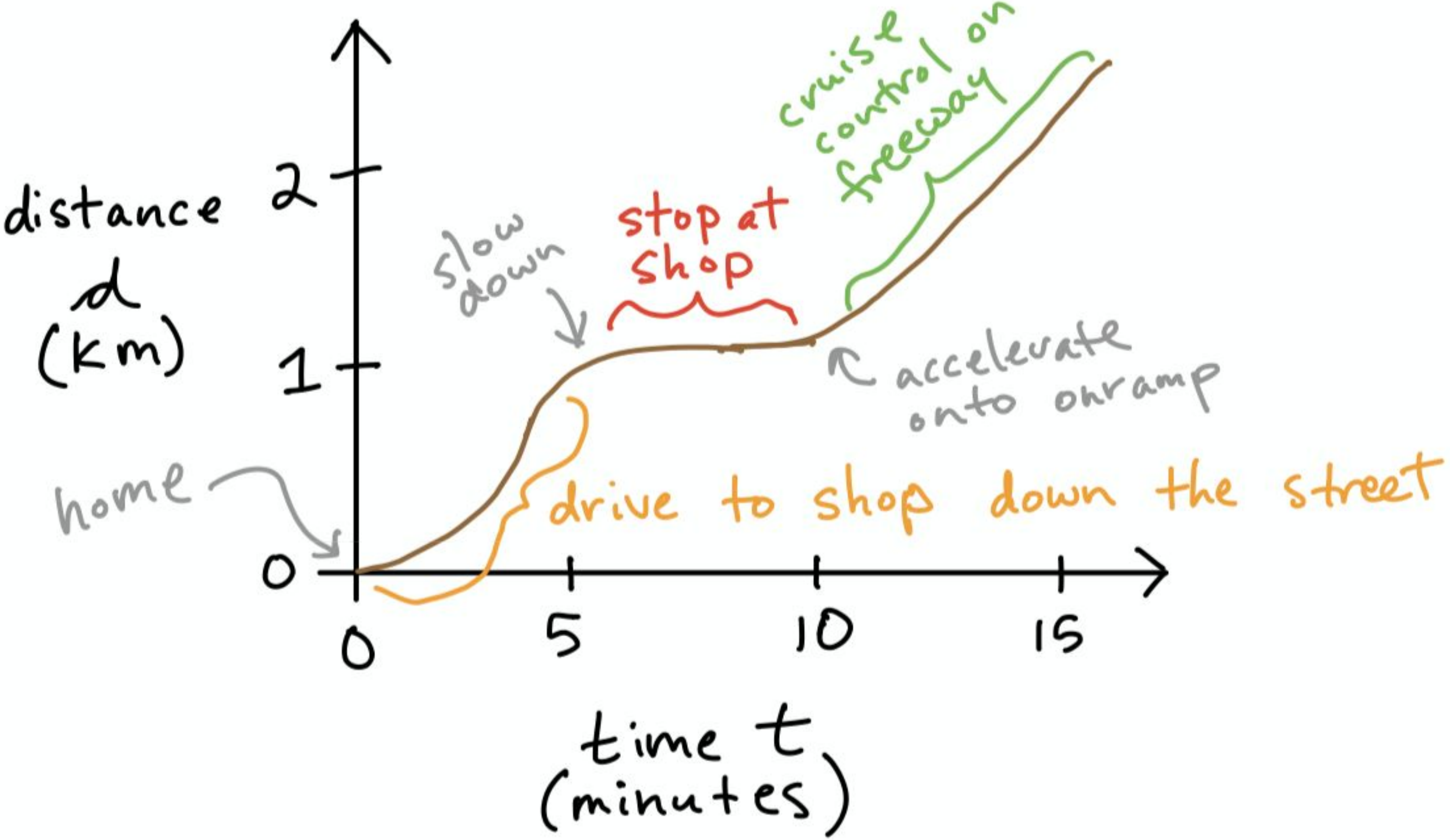


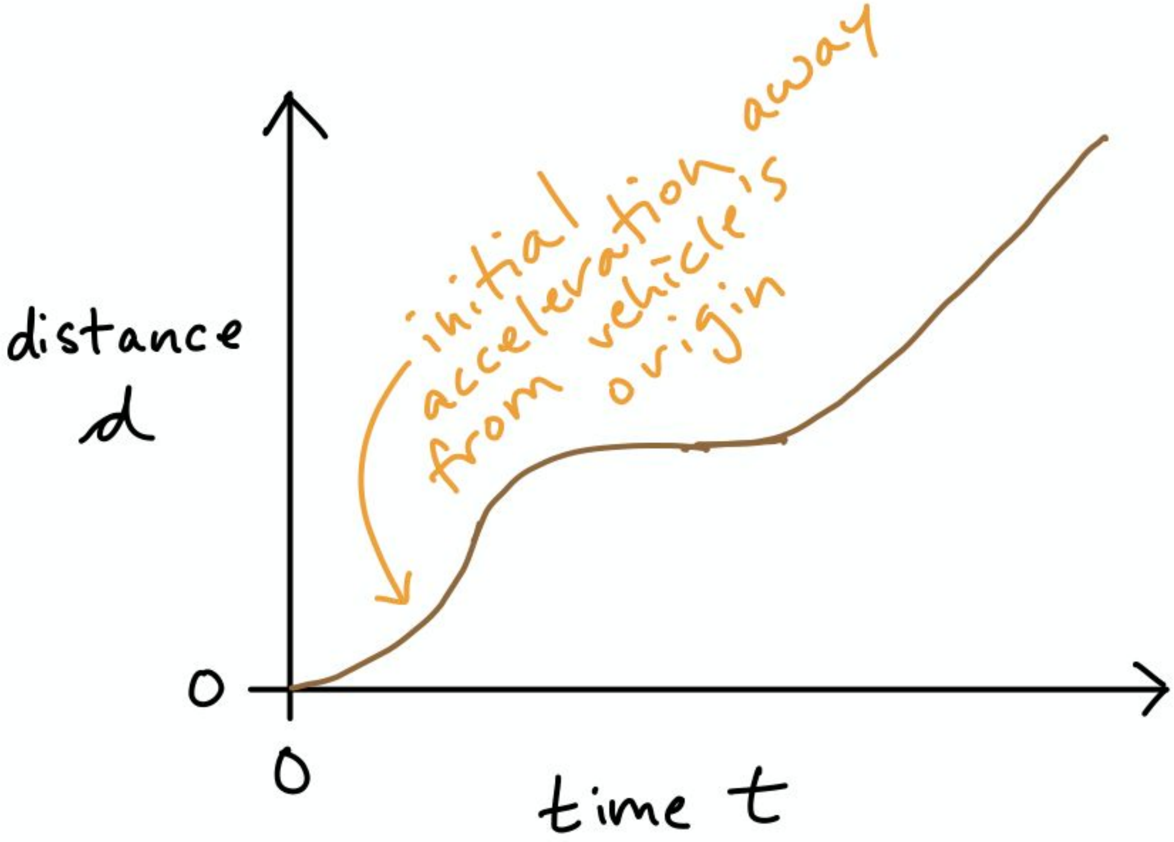

distance
d
(km)
2
1
0
home
drive to shop down the street
0
5
10
15
time t
(minutes)

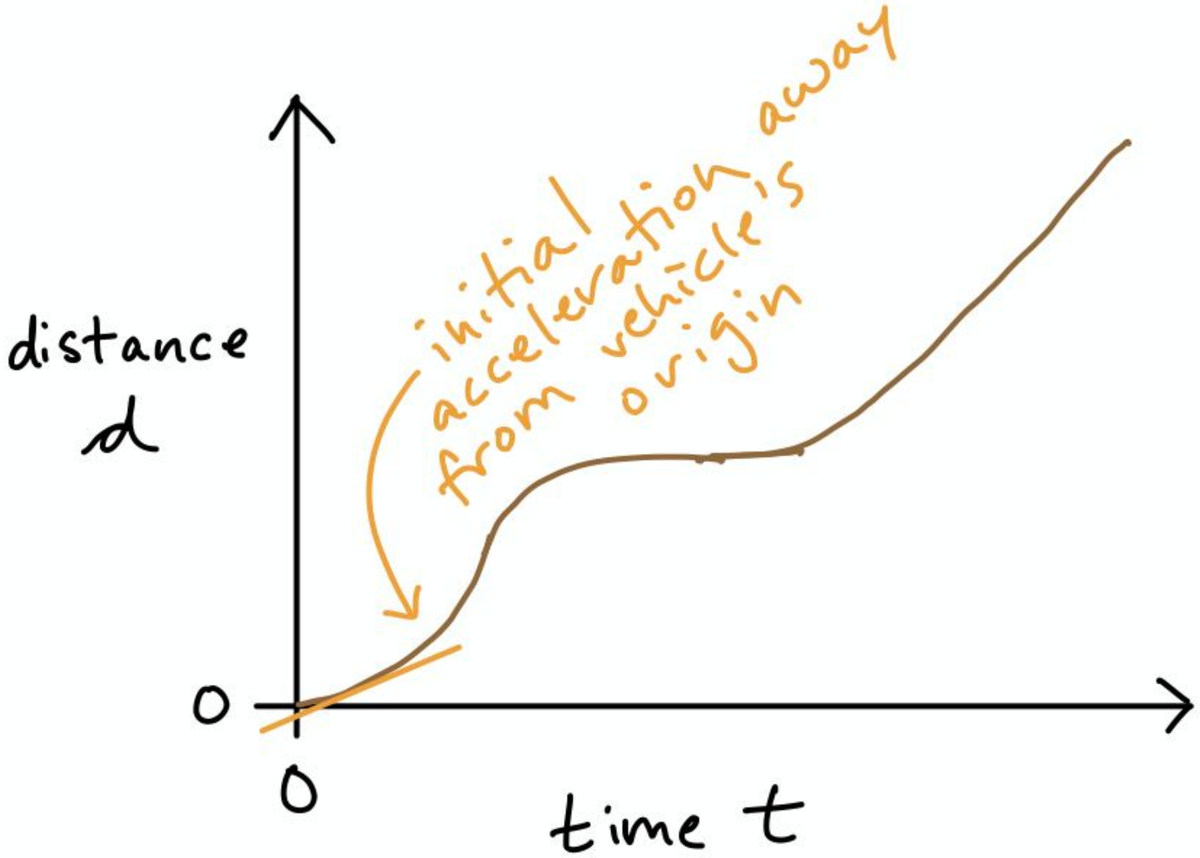

distance
d
(km)
2
1
0
slow down
home
drive to shop down the street
0
5
10
15
time t
(minutes)

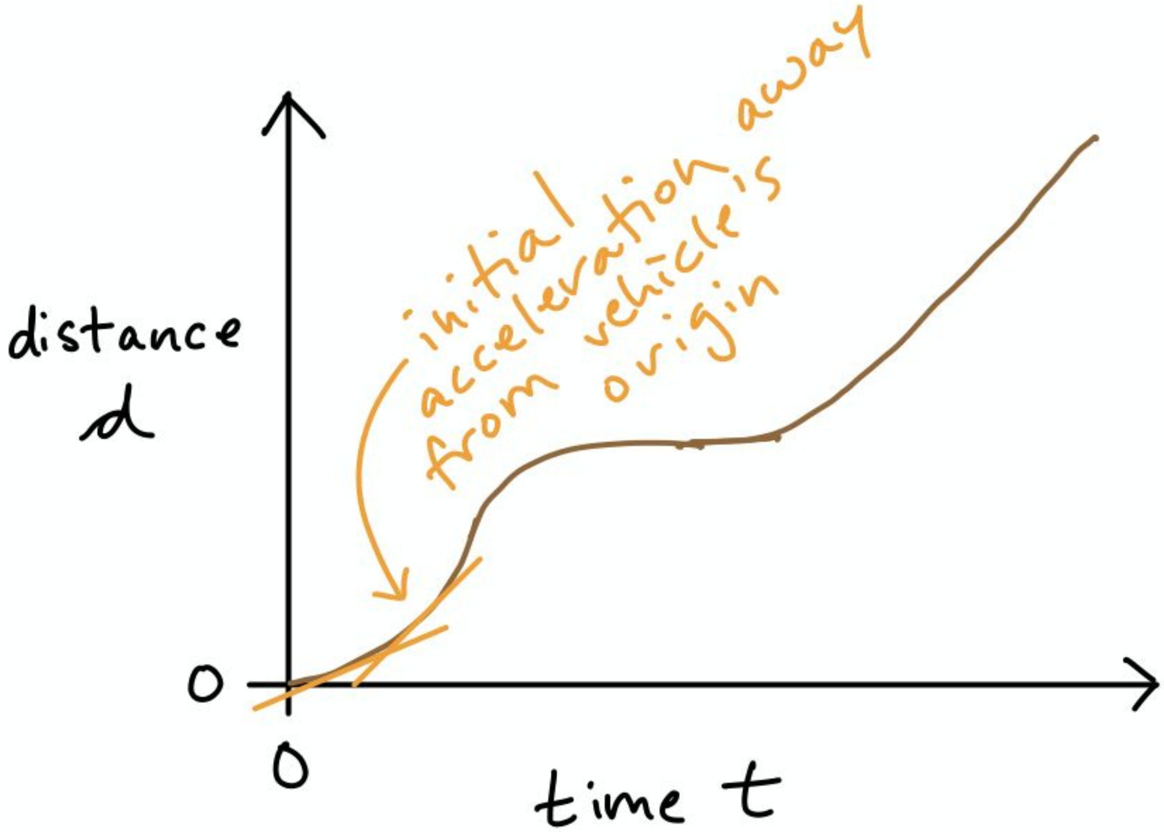

distance
d
(km)
2
1
0
slow down
stop at shop
home
drive to shop down the street
0
5
10
15
time t
(minutes)


distance
d
(km)
2
1
0
slow down
stop at shop
accelerate onto onramp
home
drive to shop down the street
0
5
10
15
time t
(minutes)

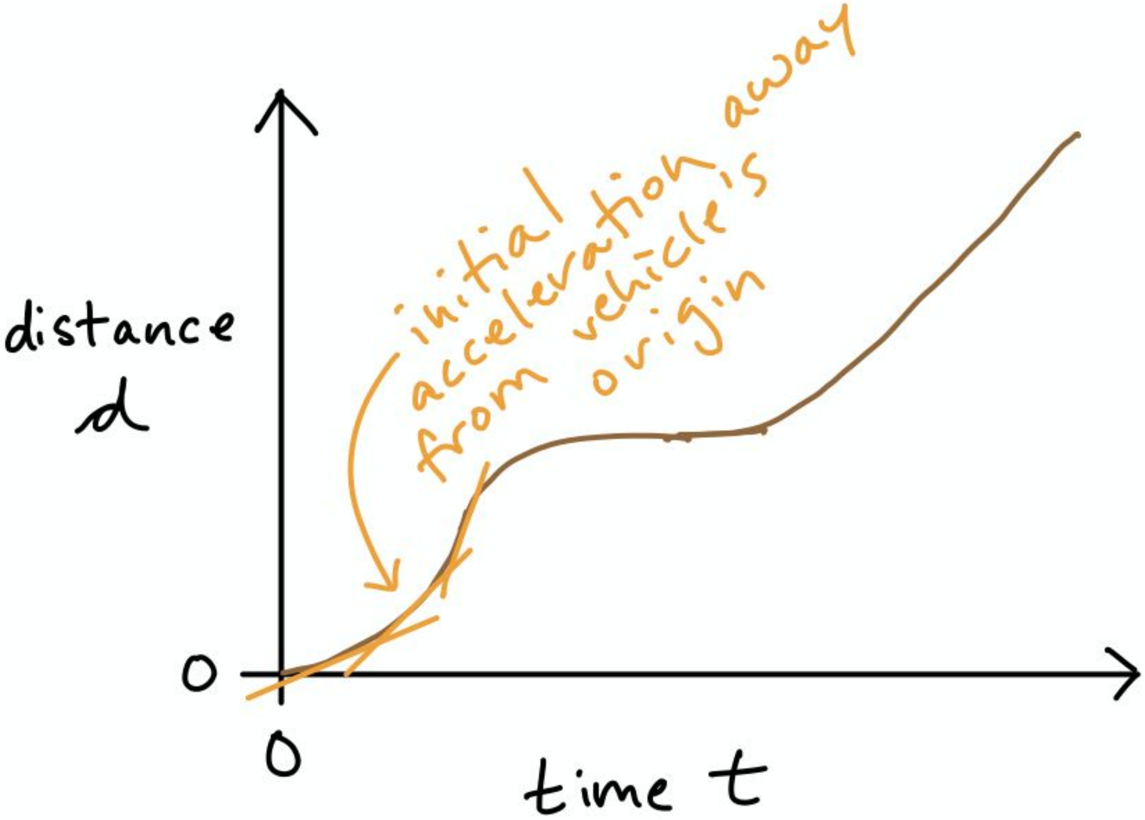

distance
d
(km)
2
1
0
0
5
10
15
time t
(minutes)
home
drive to shop down the street
slow down
stop at shop
accelerate onto onramp
cruise control on freeway

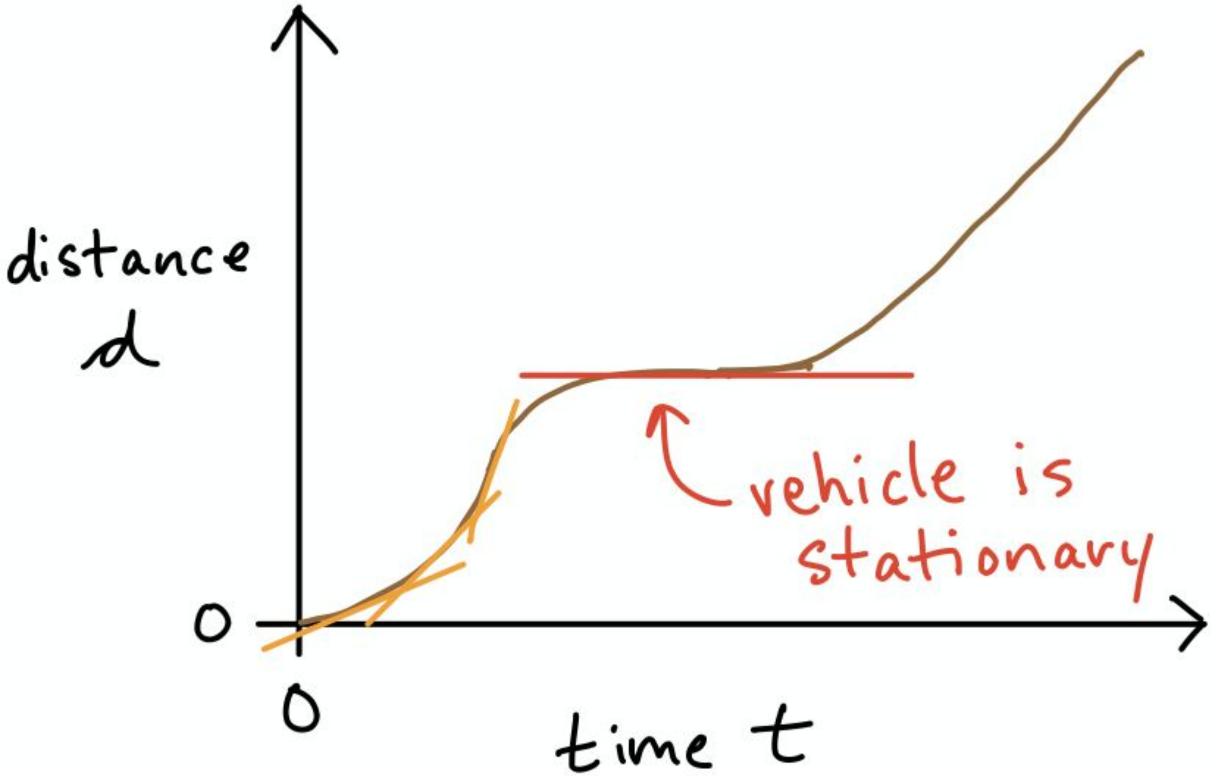

initial acceleration away from vehicle's origin
distance
d
0
0
time t

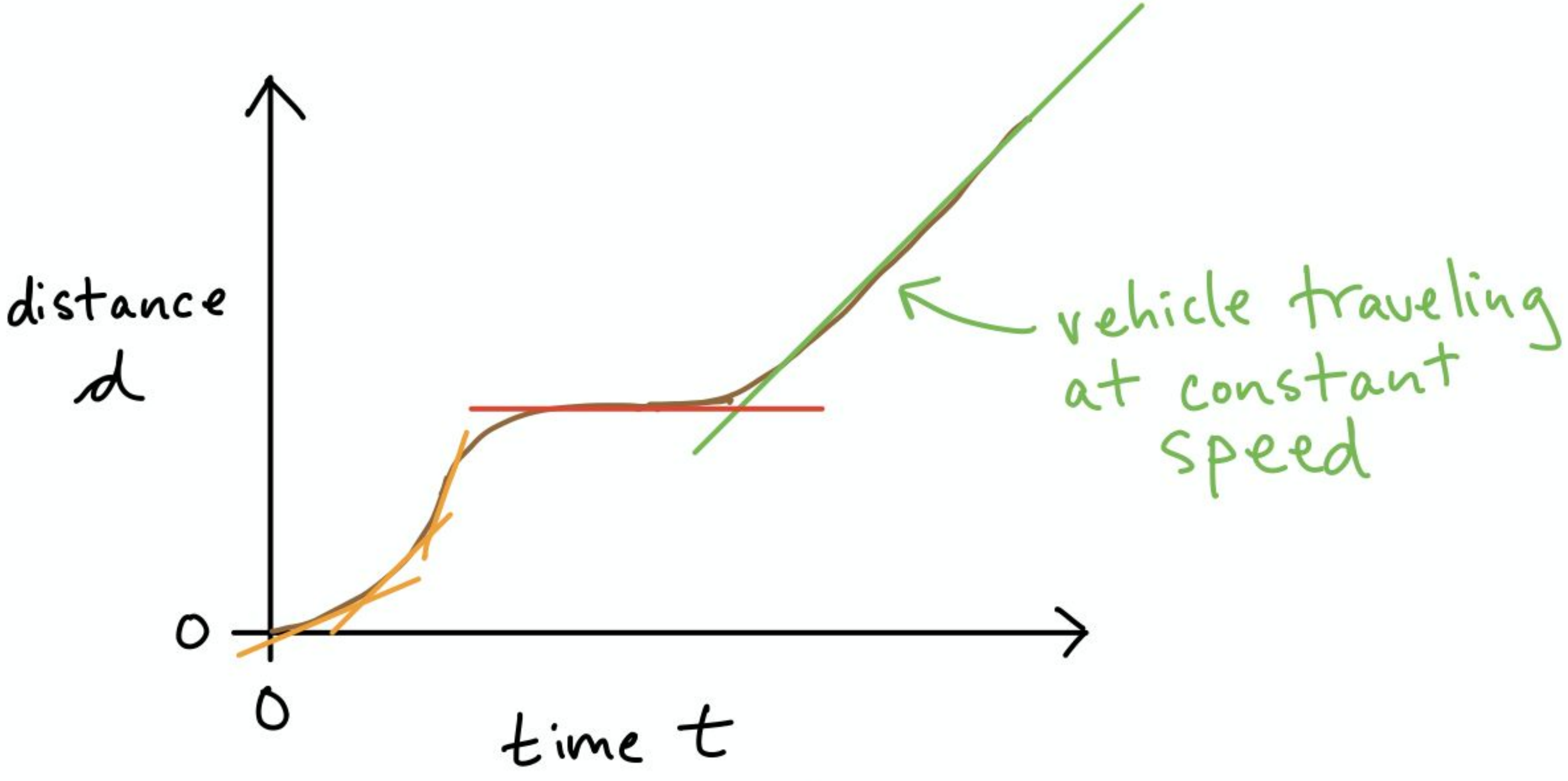

initial acceleration away from vehicle's origin
distance
d
0
0
time t

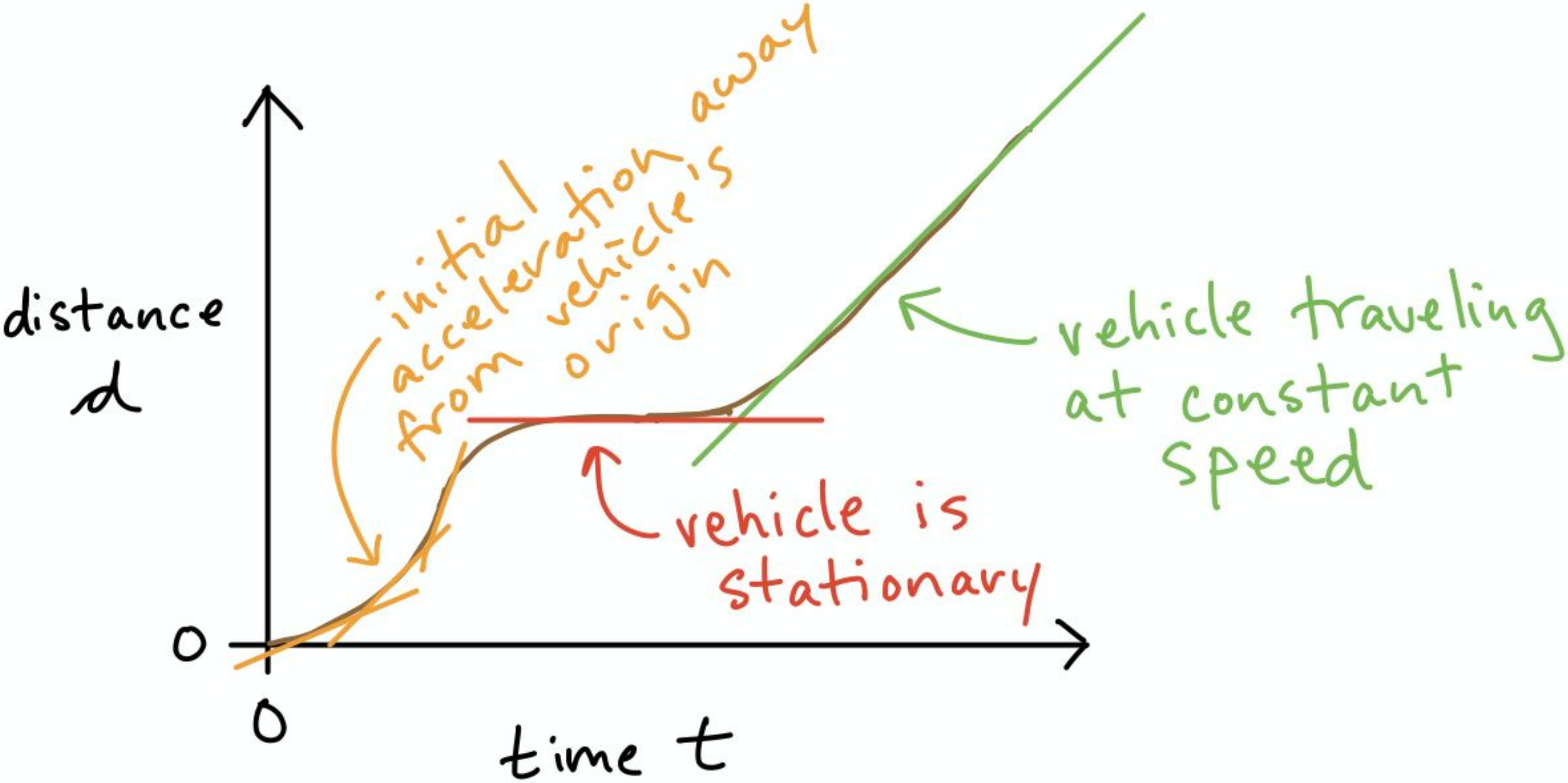


distance
d
0
0
time t
initial acceleration away from vehicle's origin

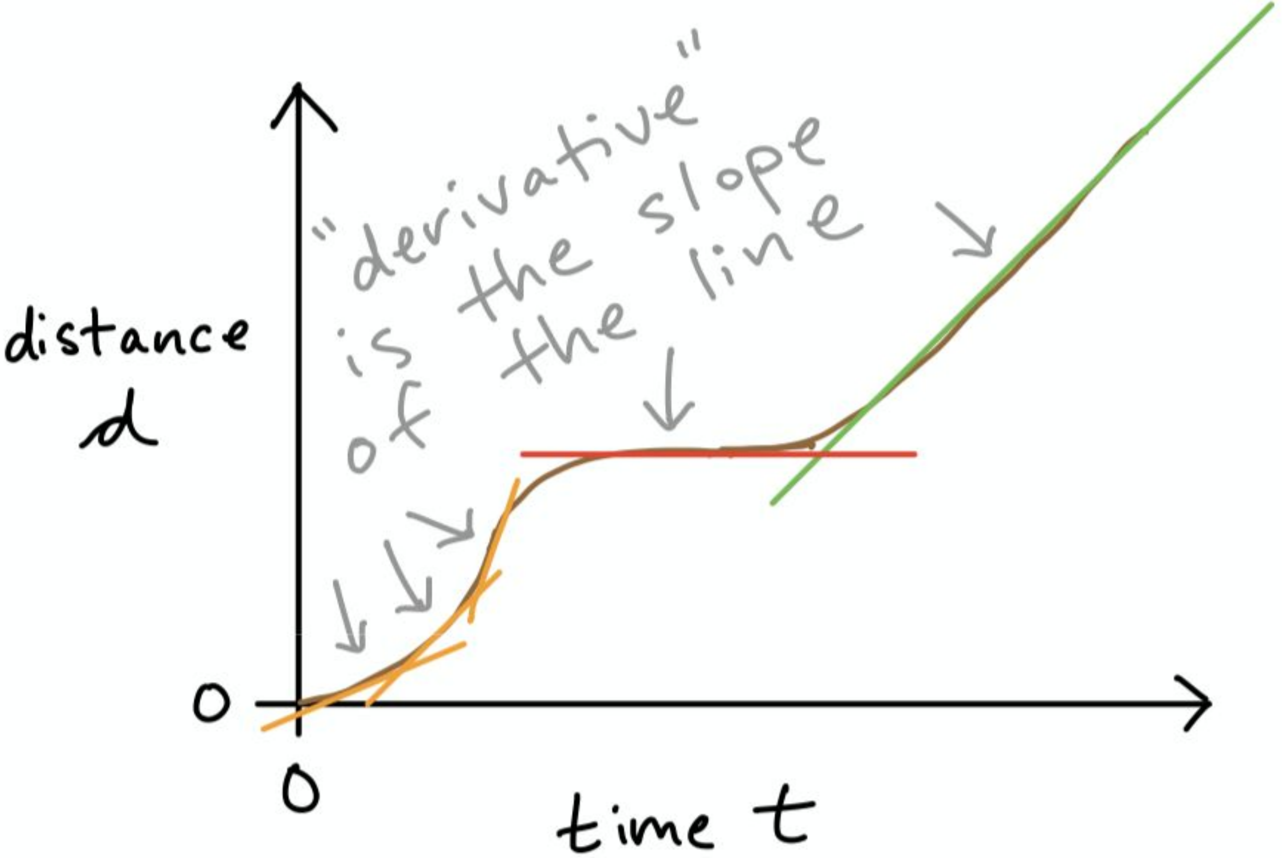

distance
d
0
0
time t
initial acceleration away from vehicle's origin

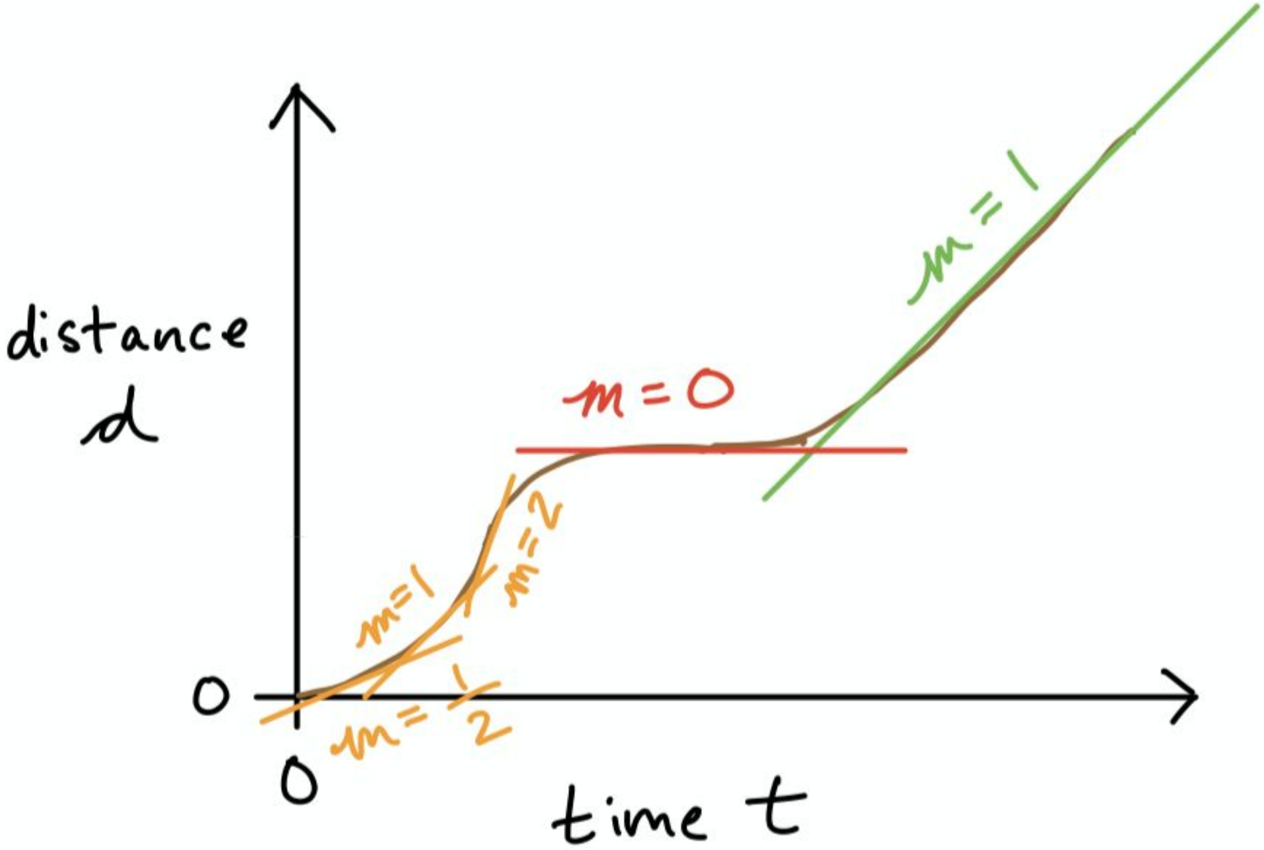

distance
d
0
0
time t
vehicle is stationary


distance
d
0
0
time t
vehicle traveling
at constant
speed

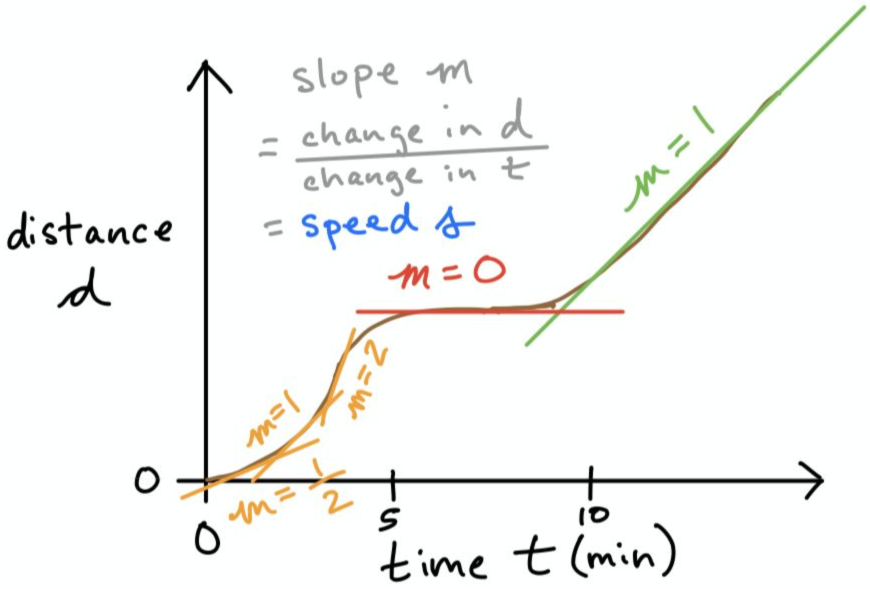


initial acceleration away from vehicle's origin
distance
d
0
0
time t
vehicle is stationary
vehicle traveling at constant speed

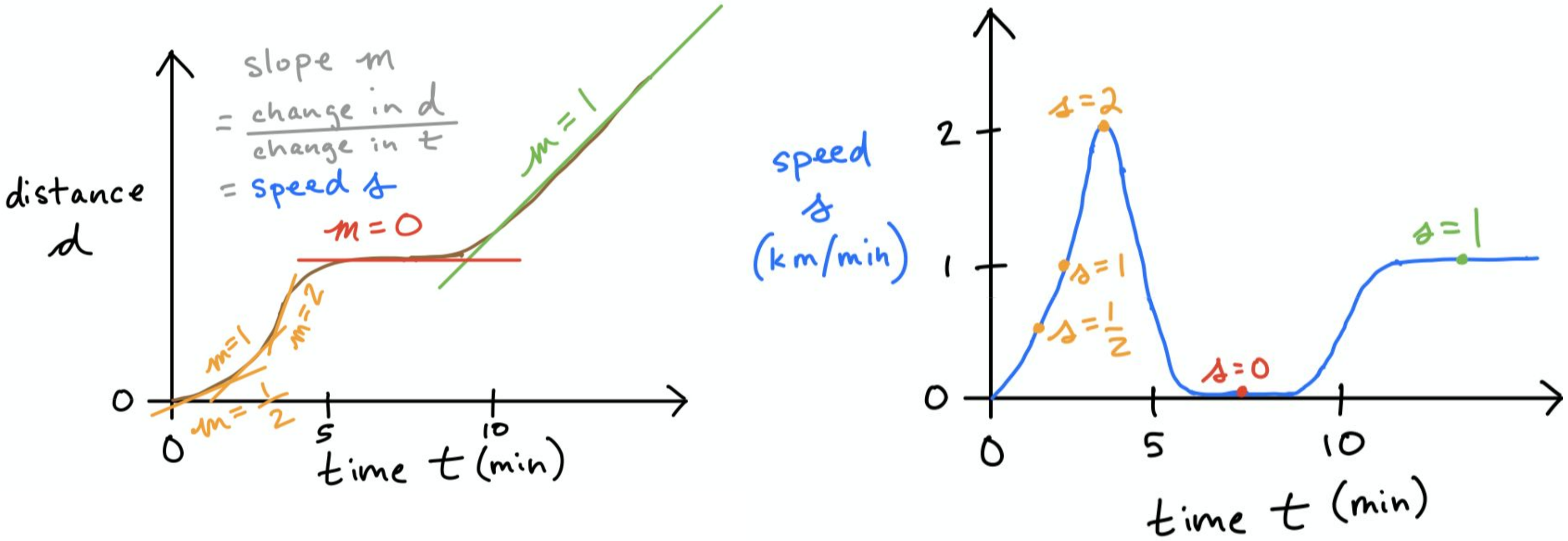

"derivative" is the slope of the line
distance d
time t
0
0

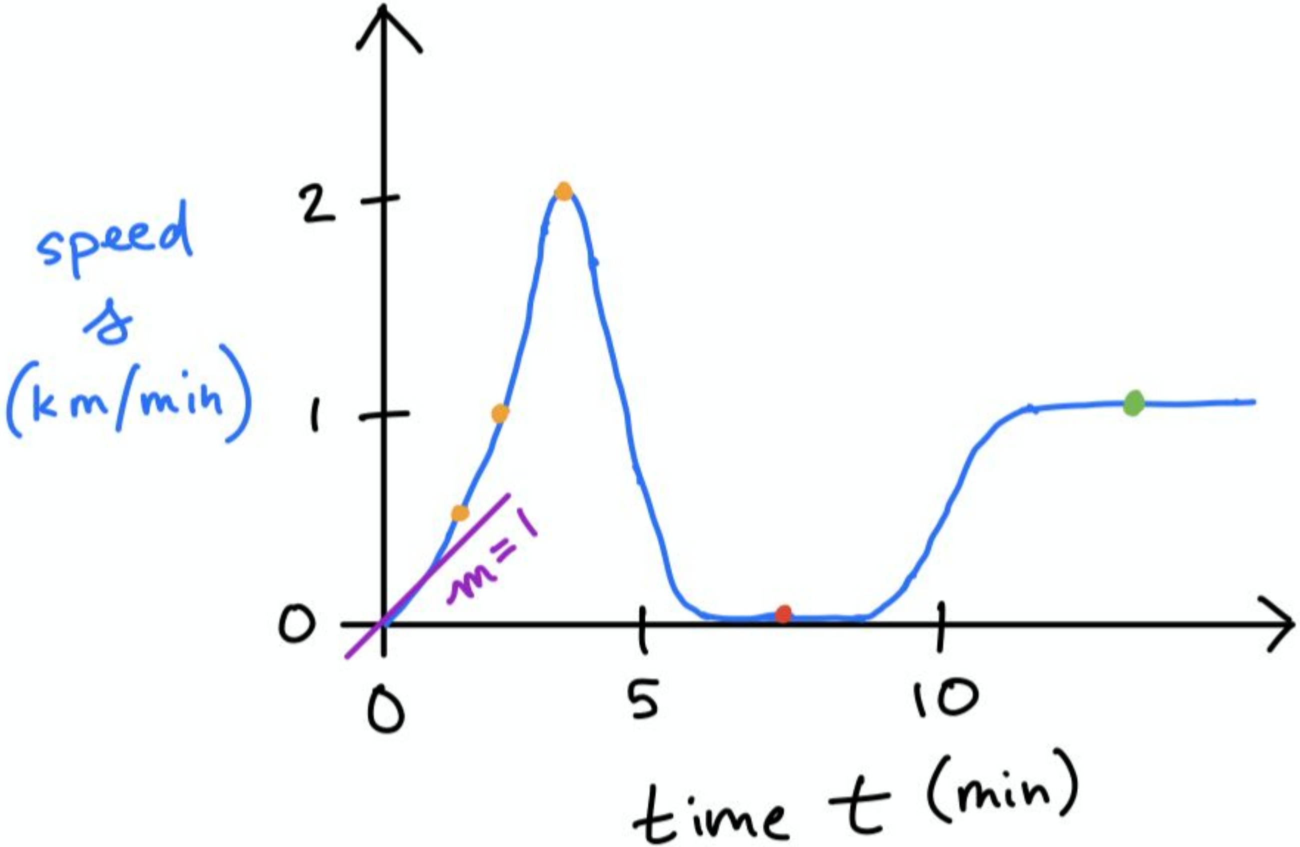


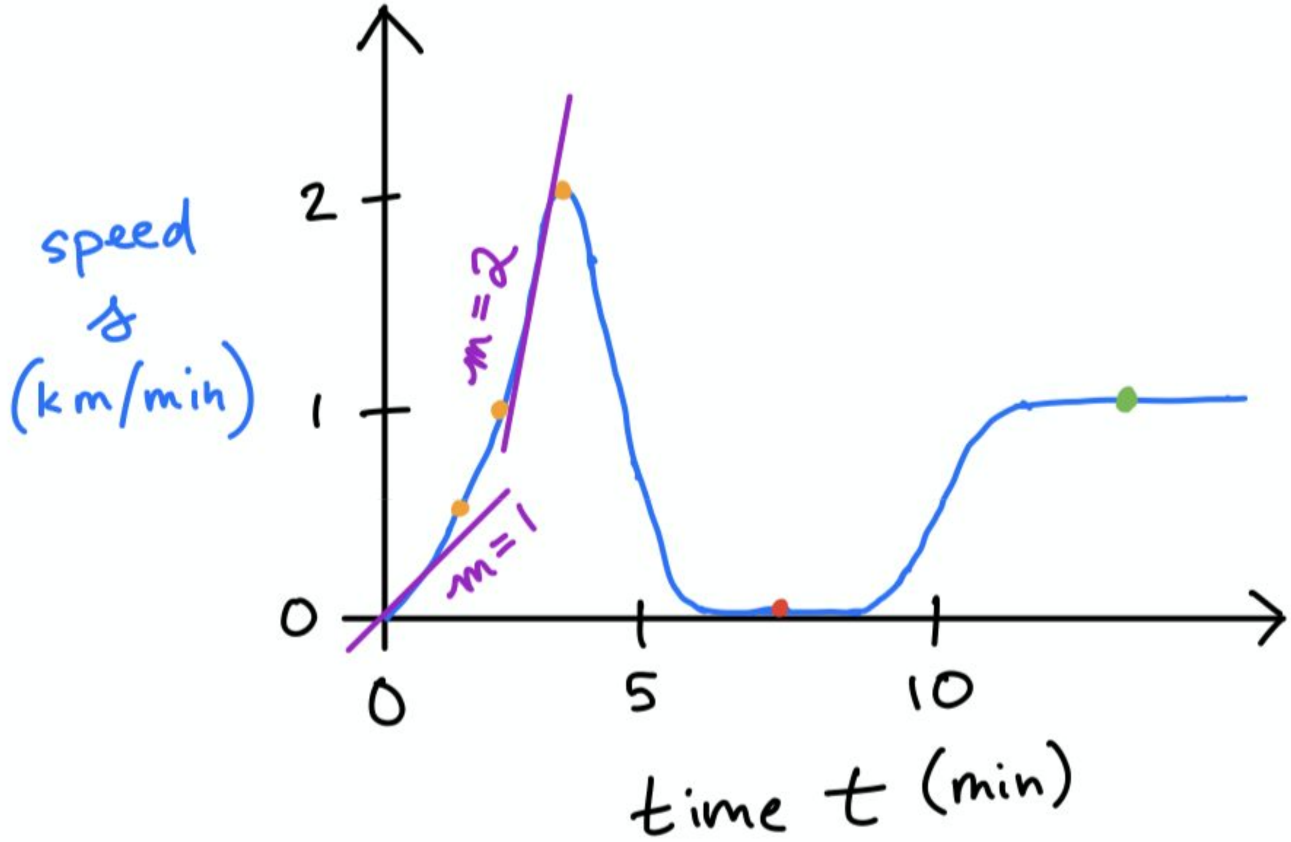
distance
d
time t
0
0
m = 0
m = 1
m=1
m=2
m = 1/2

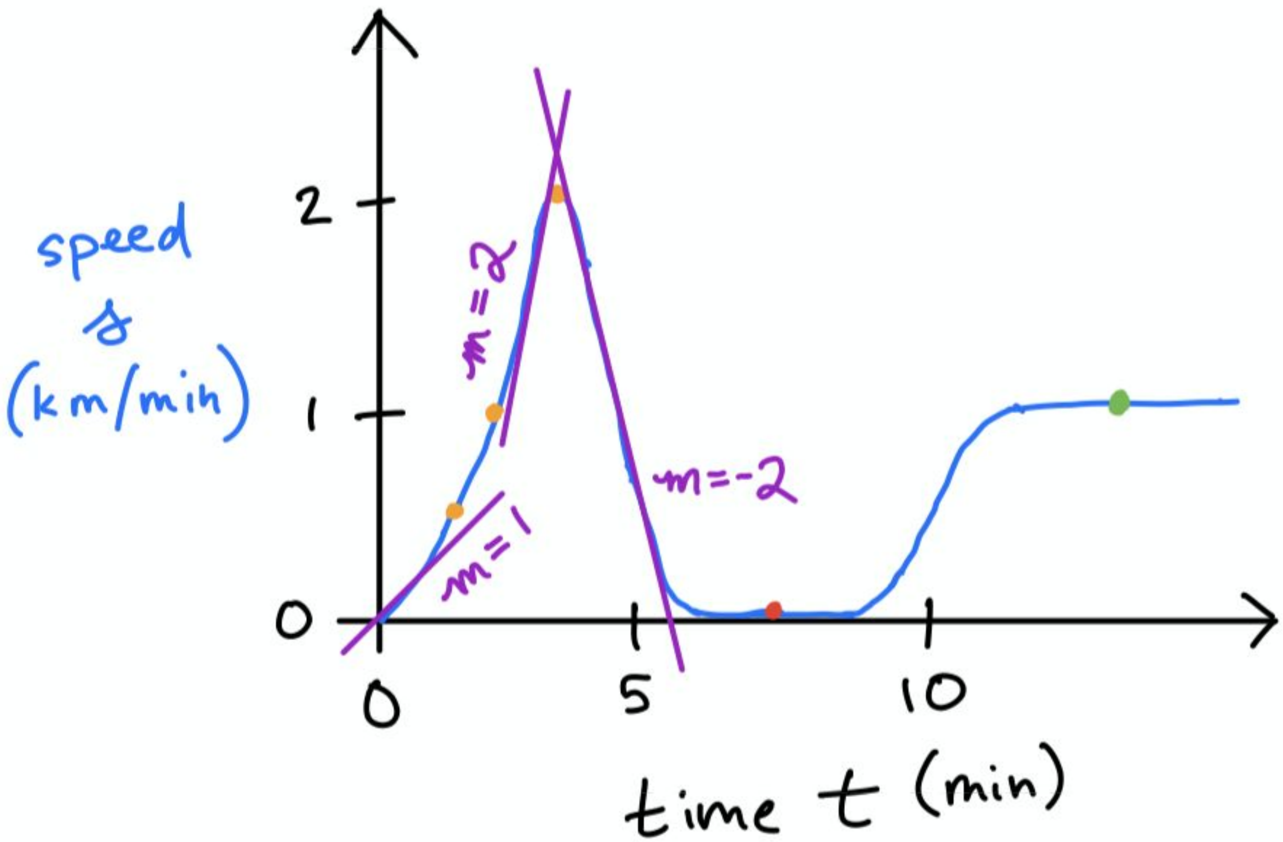


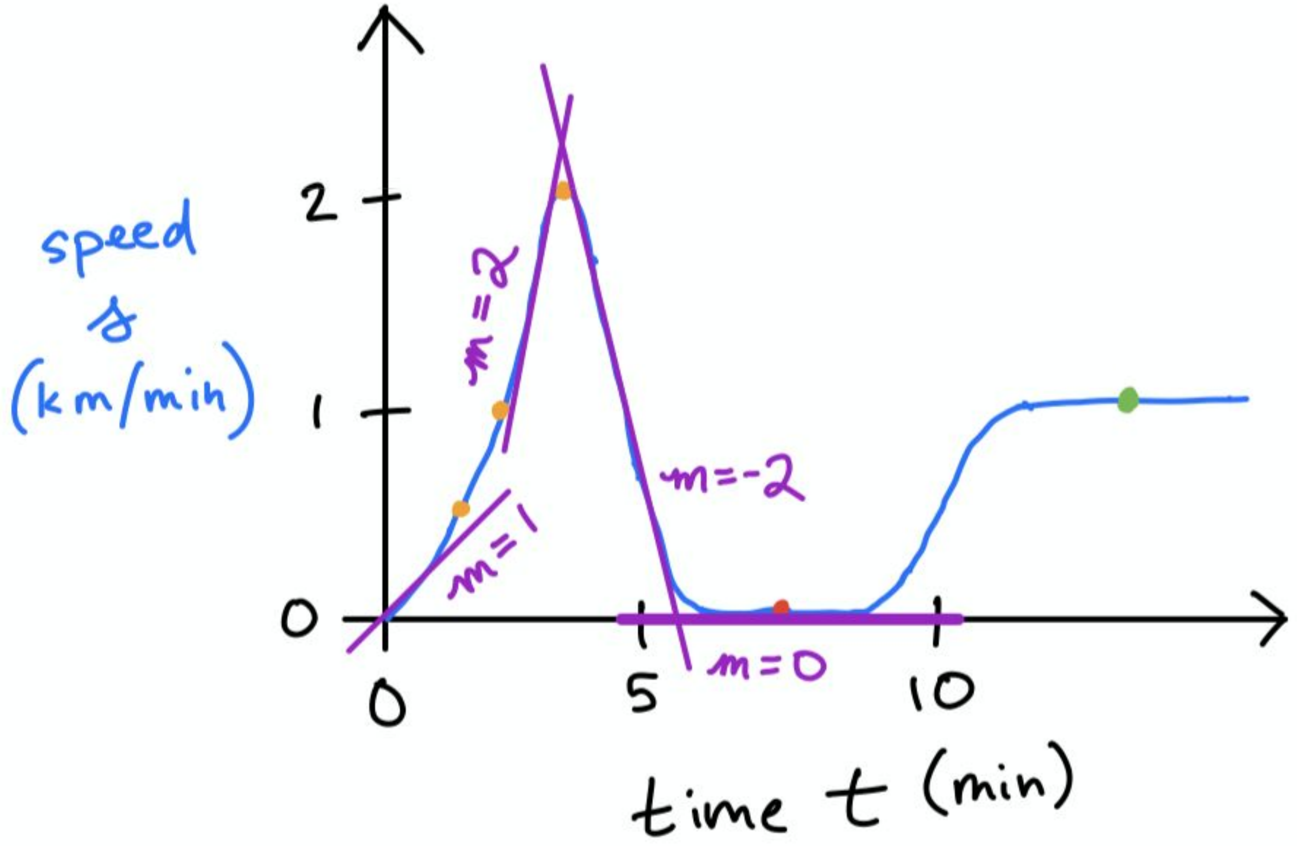
slope m
= change in d / change in t
= speed s
distance
d
0
m = 0
m = 1
m = 1
m = 2
m = 1/2
0
5
10
time t (min)

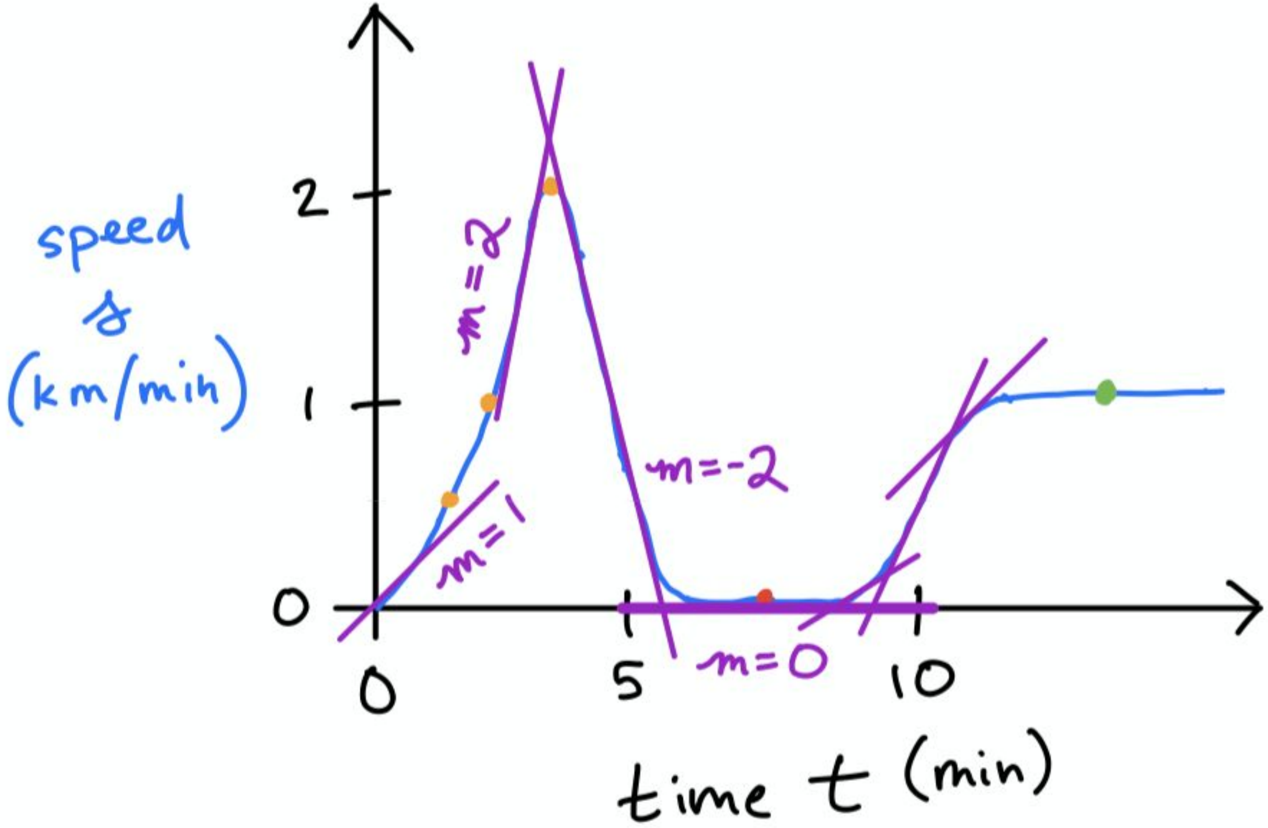

slope m
= change in d / change in t
= speed s
distance d
m = 0
m = 1
m = 2
m = 1
m = 1/2
0
0
5
10
time t (min)
speed s (km/min)
s = 2
s = 1
s = 1/2
s = 0
s = 1
2
1
0
0
5
10
time t (min)


speed
y
(km/min)
2
1
0
m=1
0
5
10
time t (min)

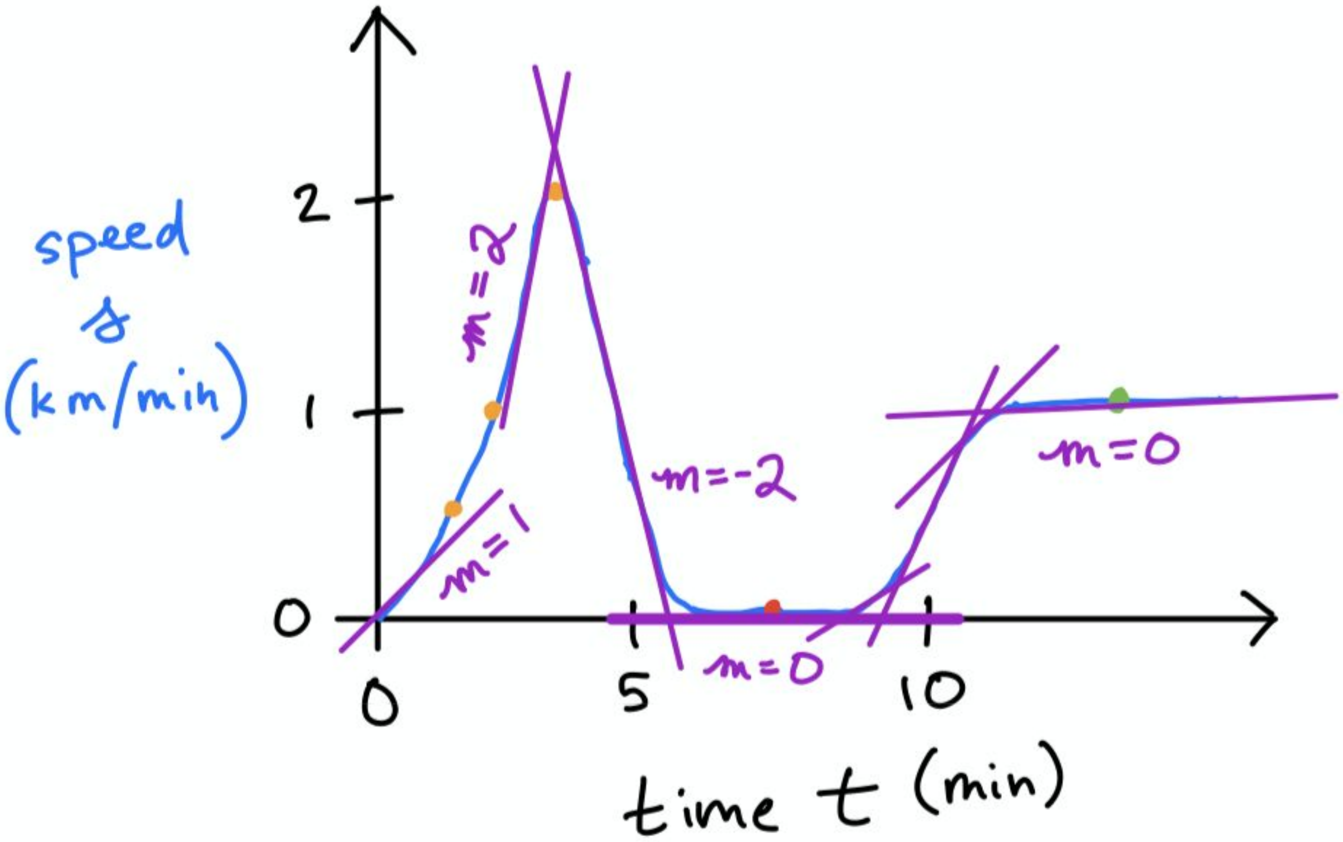

speed
y
(km/min)
2
1
0
m=2
m=1
0
5
10
time t (min)

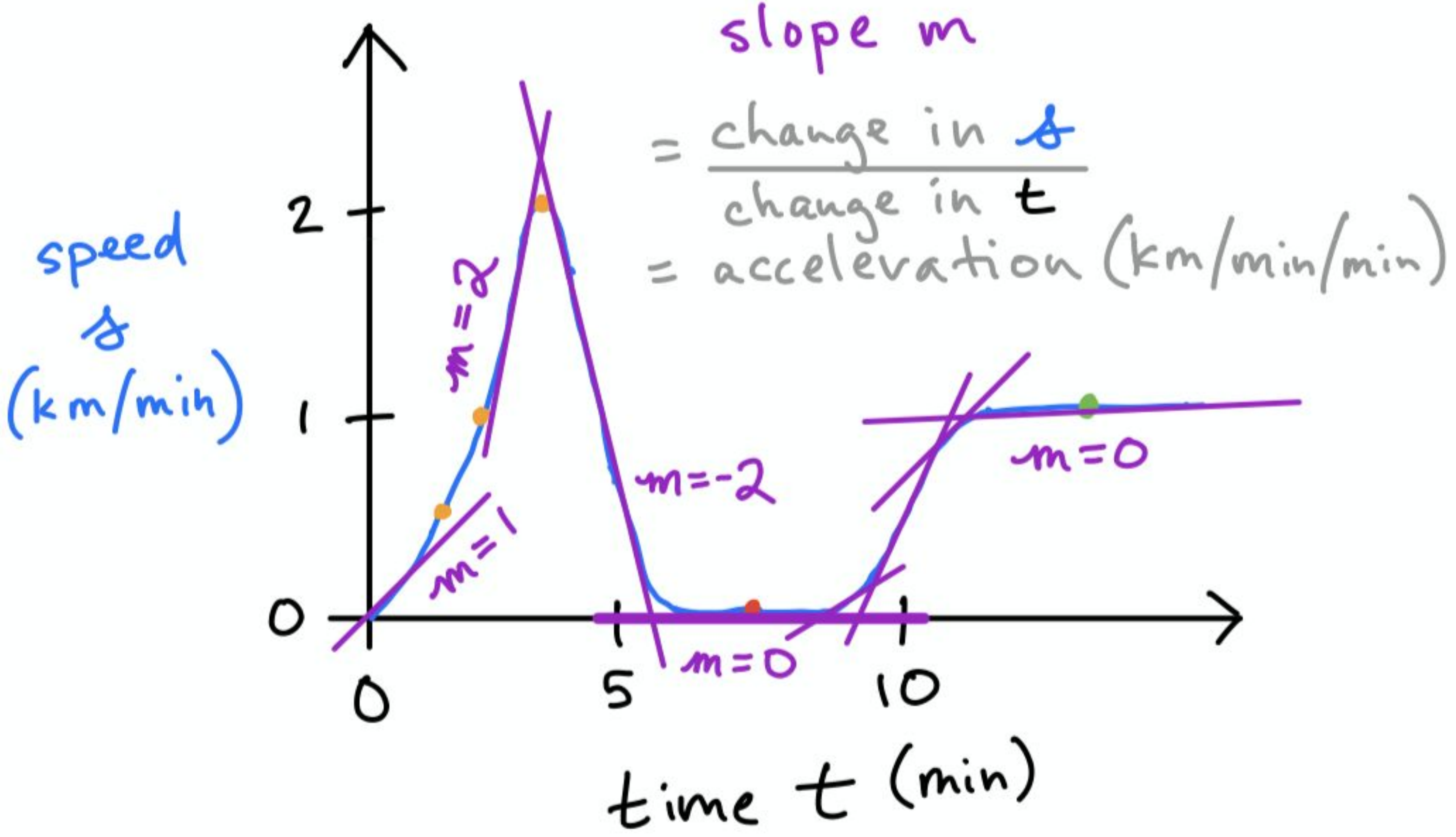

speed
y
(km/min)
2
1
0
m=2
m=1
m=-2
0
5
10
time t (min)


speed
s
(km/min)
2
1
0
m=2
m=1
m=-2
m=0
0
5
10
time t (min)



speed
y
(km/min)
2
1
0
0
5
10
m=2
m=-2
m=1
m=0
time t (min)


speed
y
(km/min)
2
1
0
m=2
m=-2
m=1
m=0
m=0
0
5
10
time t (min)

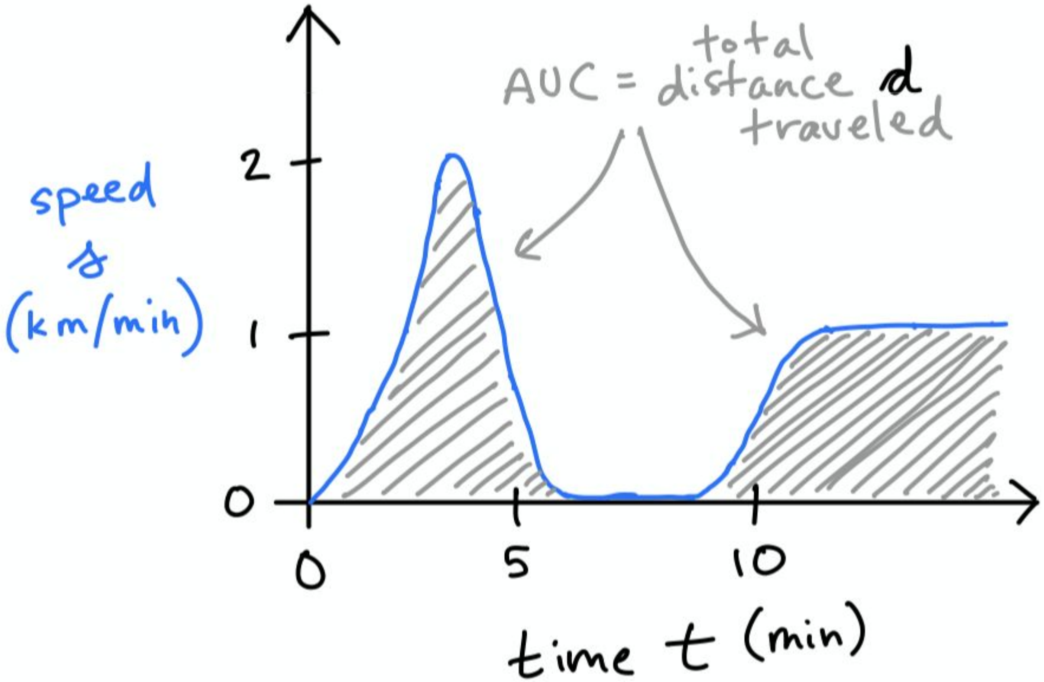

slope m
= change in s / change in t
= acceleration (km/min/min)
speed
s
(km/min)
2
1
0
m=2
m=-2
m=1
m=0
m=0
m=0
0
5
10
time t (min)

What Calculus Is

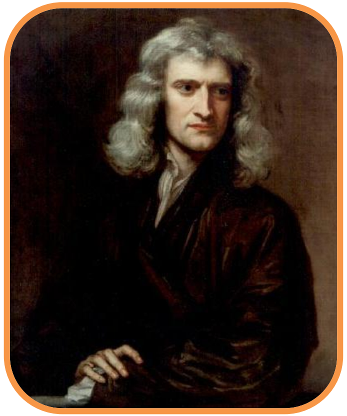
- **Mathematical study of continuous change**
- Two branches:
  a. **Differential calculus**: focus of *Calculus I*
  b. **Integral calculus**: a focus of *Calculus II* subject

What Integral Calculus Is

- **Study of areas under curves**
- Facilitates the inverse of differential calculus:



A Brief History of Calculus

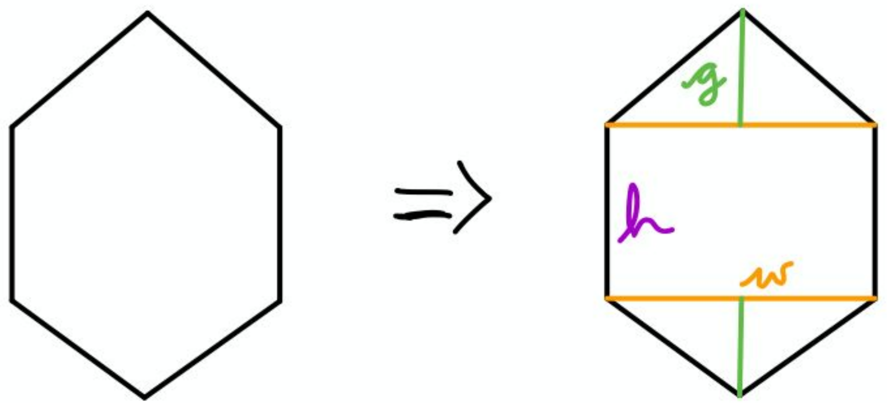


- ~1800 BCE: Egyptian papyrus with area calculations
- "Method of exhaustion" *(covered next)* developed by:
  - Greeks Eudoxus (~400 BCE) and Archimedes (~250 BCE)
  - Independently by Chinese Liu Hui (~250 CE)
- Arab al-Haytham (11th c. CE) used integrals to calculate volumes
- Indians (14th c.) had differentiation-like methods
- 17th c.: **Gottfried Leibniz** (Germany) & **Isaac Newton** (England)
  - Independently developed modern calculus
  - Incl. higher-order differentiation and integration
  - Incl. rules like "product rule" & "chain rule" *(in Calc I)*
  - Leibniz named it and devised notation (*we mostly use his!*)
  - Newton applied it first to physics, e.g.:
    - Laws of motion
    - Gravity

Images from public domain

The Method of Exhaustion

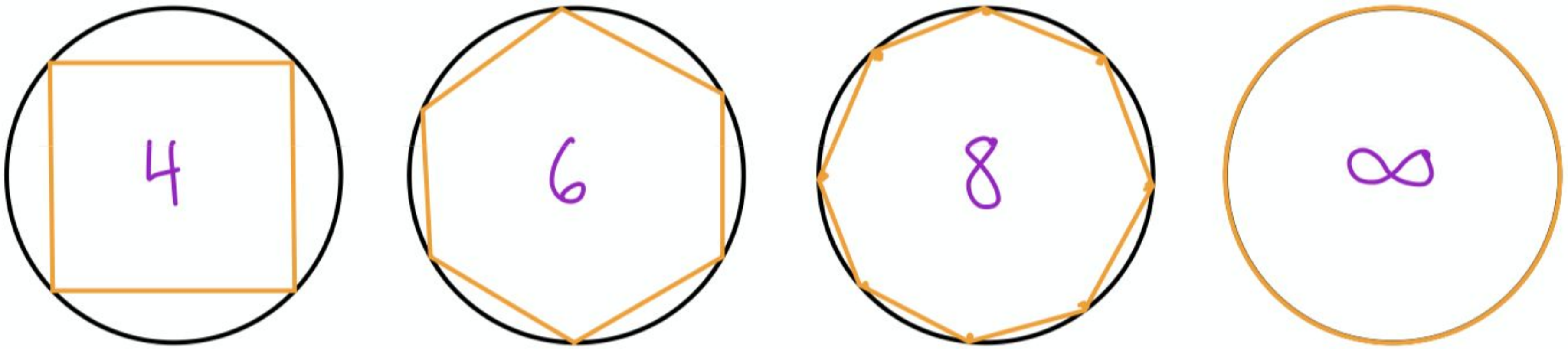
- **Polygon**: shape with multiple straight sides (3+; typically 5+)
  - Greeks learned to find area by filling with triangles



$$\text{area} = \frac{1}{2}(gw) + hw + \frac{1}{2}(gw)$$

The Method of Exhaustion

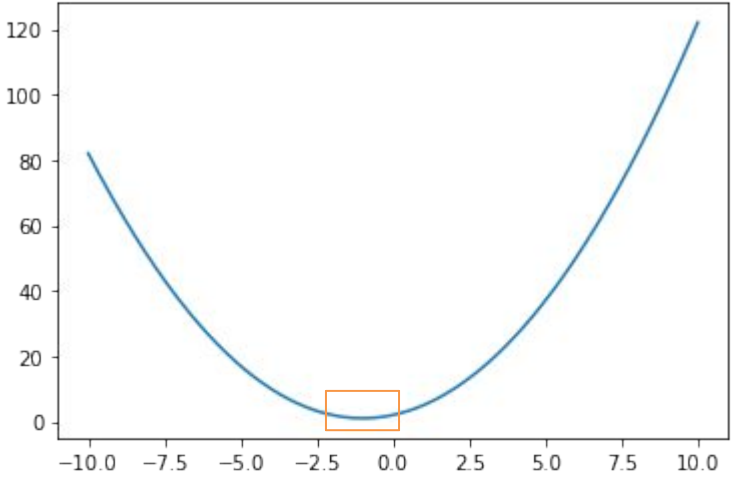
...the **Method of Exhaustion** builds on this to find area of curved shapes:

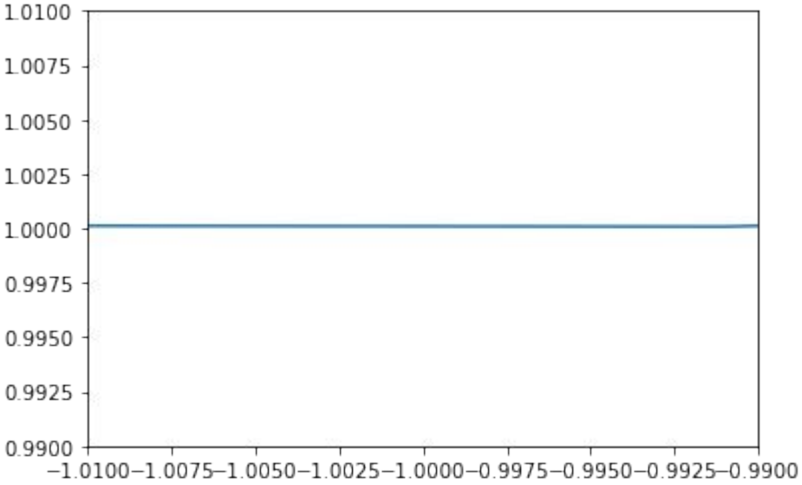


- Integral calculus is directly based on this method.
- Calculus in general relies on the idea of approaching infinity.
  - Leibniz gave field the name "calculus of the infinitesimals"

Calculus of the Infinitesimals

As integral accuracy improves as we approach an infinite-sided polygon, so too does differential accuracy improve as we approach a curve infinitely closely:



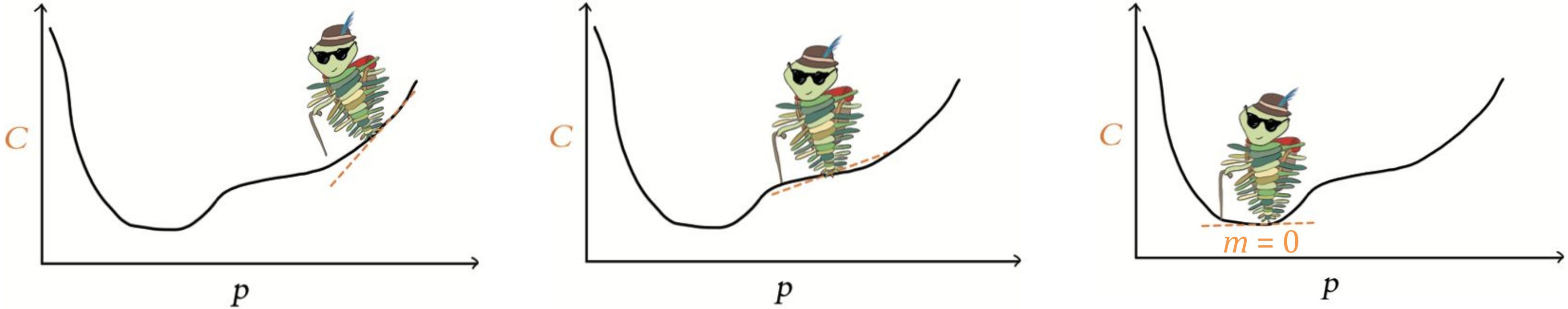


Hands-on code demo

Modern Calculus Applications

Differentials: optimize by finding minima and maxima of curves

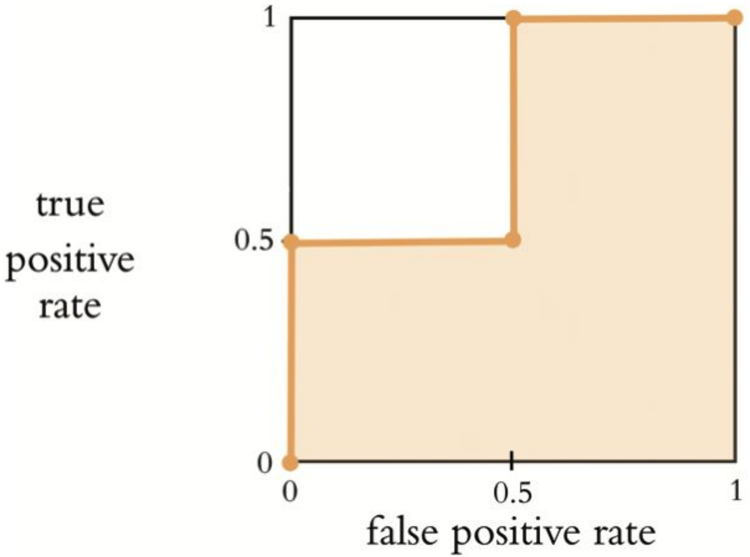
- Used, e.g., in engineering (max strength) and finance (min cost)
- Widely used in machine learning, incl. deep learning
    - **Gradient descent** to minimize cost (see `jonkrohn.com/deepTF1`)
    - Gradient ascent to maximize reward (see `jonkrohn.com/DQN`)
- Higher-order derivatives used in "fancy" optimizers *(Optimization)*



Modern Calculus Applications

Integrals: find area under the curve

- **Receiver operating characteristic** *(Calc II)*
- Probability theory's "expectation" of random variable is widely used in machine learning, incl. deep learning *(Prob. & Info. Thy.)*



Limits

Trivially easy to calculate for a continuous function, e.g.:

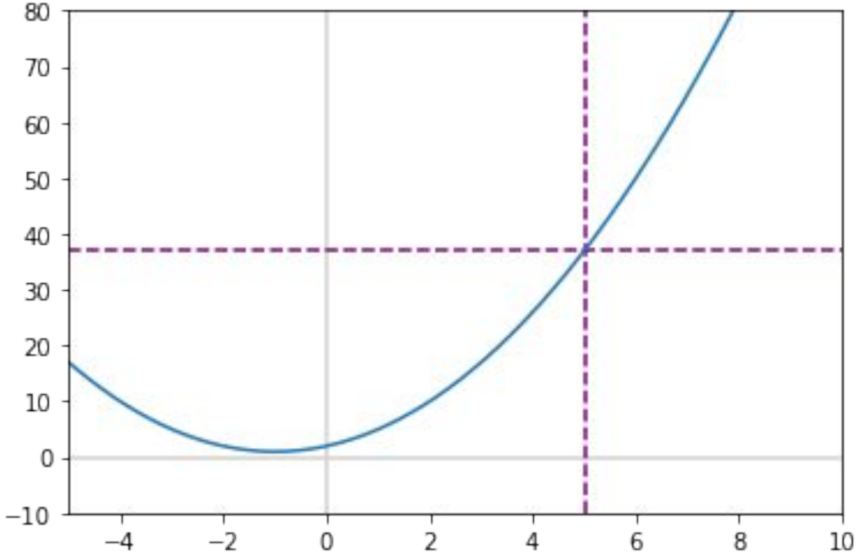
What is the limit as x approaches 5 in the expression $x^2 + 2x + 2$?

$$\lim_{x \to 5} (x^2 + 2x + 2)$$

$$= 5^2 + 2 \cdot 5 + 2$$

$$= 25 + 10 + 2$$

$$= 37$$

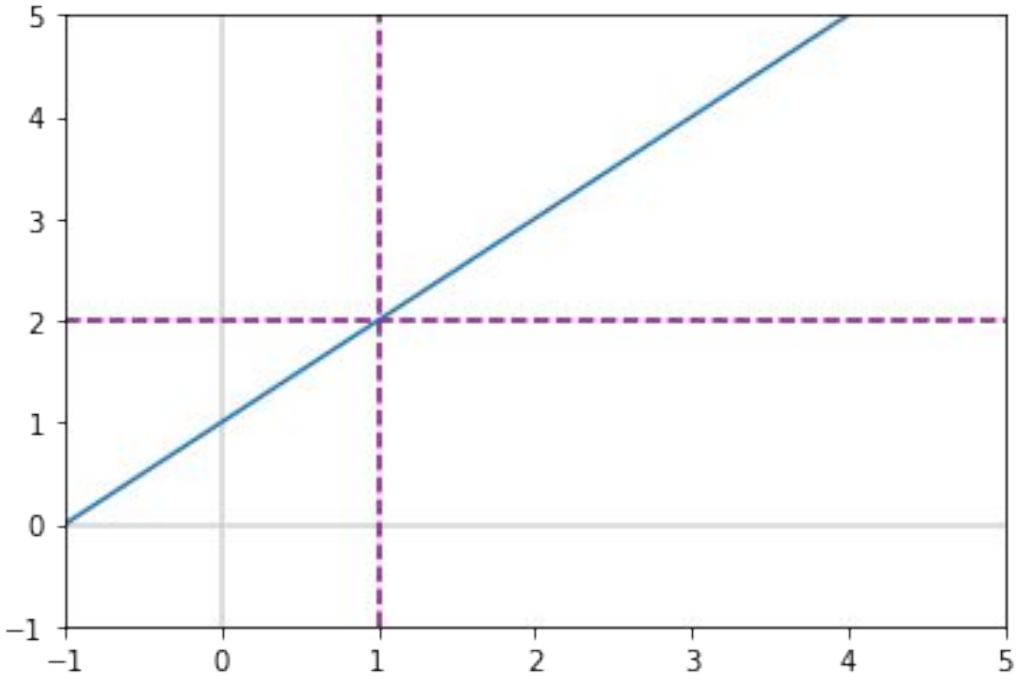


Limits

However, some functions are not continuous:

$$\lim_{x \to 1} \frac{x^2 - 1}{x - 1}$$

Hands-on code demo



Limits

In some cases, we can solve the limit through algebra, e.g., factoring:

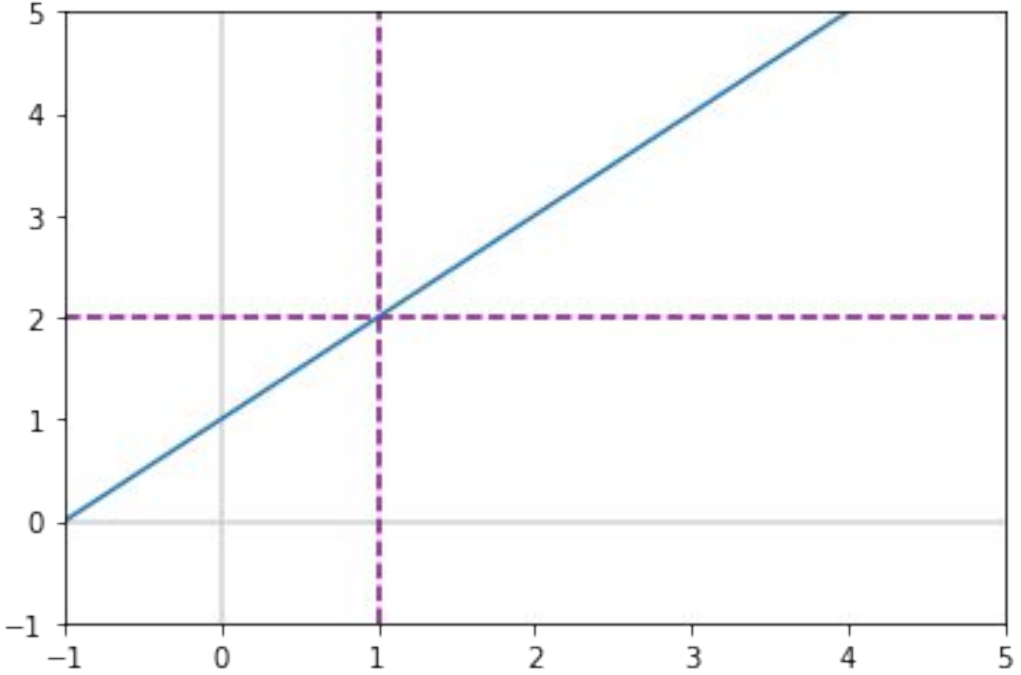
$$\lim_{x \to 1} \frac{x^2 - 1}{x - 1}$$

$$= \lim_{x \to 1} \frac{(x-1)(x+1)}{(x-1)}$$

$$= \lim_{x \to 1} \frac{\cancel{(x-1)}(x+1)}{\cancel{(x-1)}}$$

$$= \lim_{x \to 1} (x+1)$$

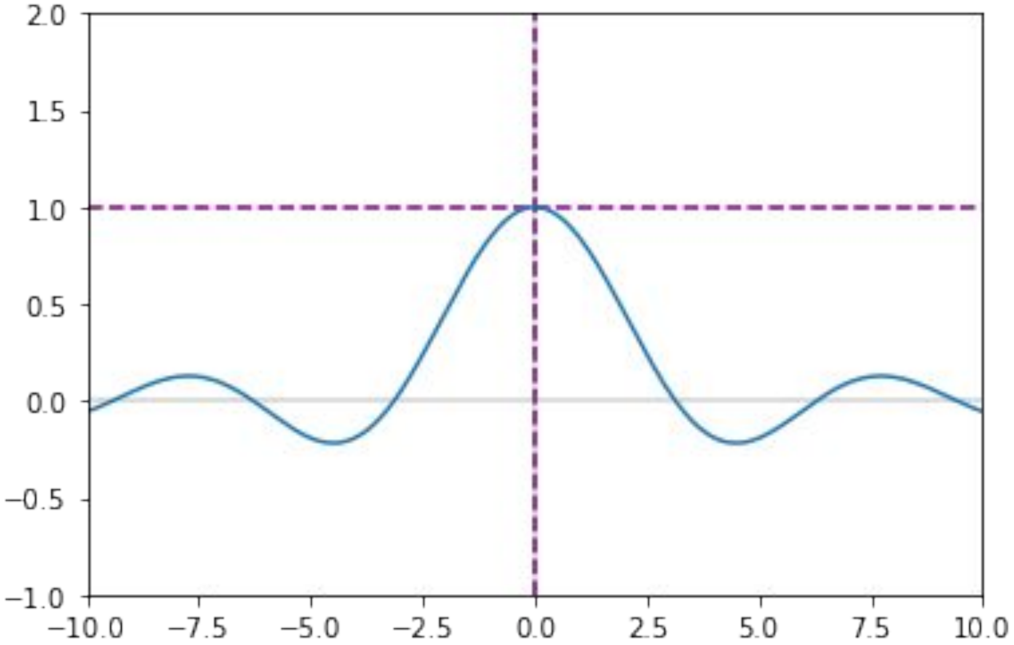
$$= 1 + 1$$



Limits

In other cases, we can't use algebra, but approaching the limit still works:

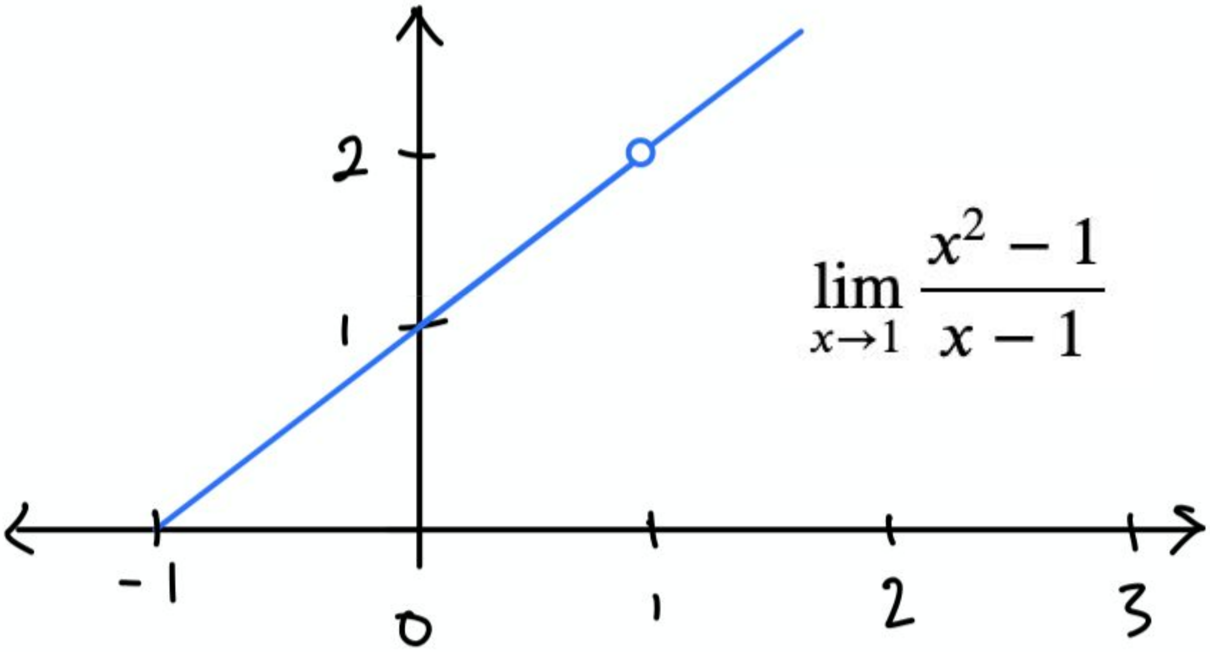
$$\lim_{x \to 0} \frac{\sin x}{x}$$



Hands-on code demo

Limits

Technically, charts should illustrate undefined points, e.g.:

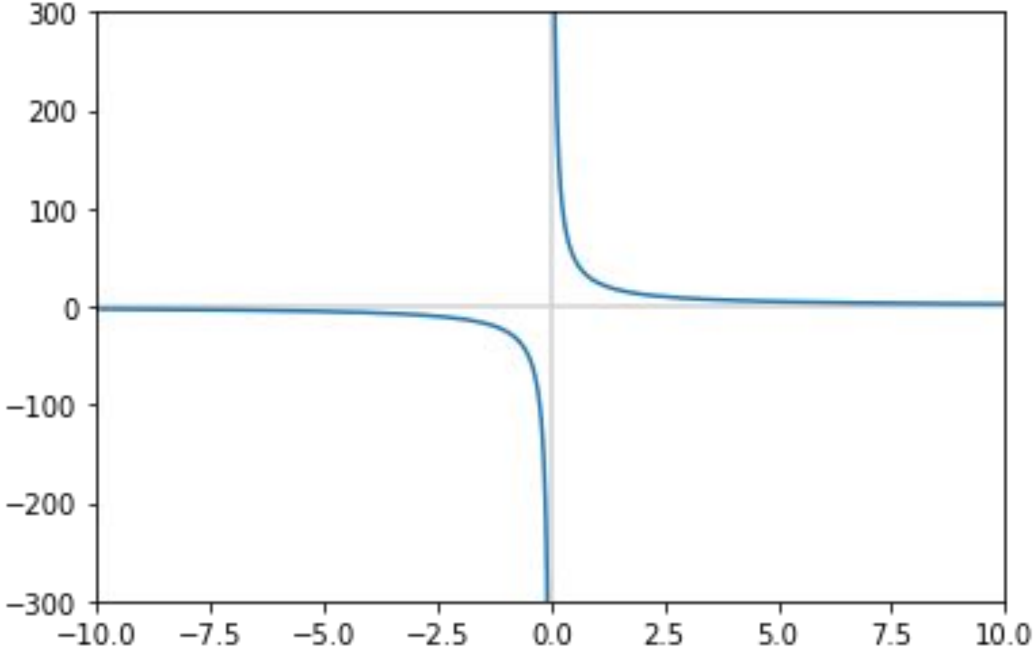


$$\lim_{x \to 1} \frac{x^2 - 1}{x - 1}$$

Limits

Finally, it is common for limits to approach infinity:

$$\lim_{x \to \infty} \frac{25}{x}$$



Hands-on code demo

Solutions

① $\lim_{x \to 0} \frac{x^2-1}{x-1}$

$= \frac{0^2-1}{0-1}$

$= \frac{-1}{-1}$

$= 1$

② $\lim_{x \to -5} \frac{x^2-25}{x+5}$

$= \lim_{x \to -5} \frac{\cancel{(x+5)}(x-5)}{\cancel{x+5}}$

$= \lim_{x \to -5} x-5$

$= -5-5 = -10$

Solutions

3

$$\lim_{x \to 4} \frac{x^2 - 2x - 8}{x - 4}$$

$$= \lim_{x \to 4} \frac{\cancel{(x-4)}(x+2)}{\cancel{x-4}}$$

$$= \lim_{x \to 4} x + 2$$

$$= 4 + 2 = 6$$

Solutions

4. 0

5. $\lim_{x \to 0^+} f(x) = \infty$ (right-handed limit)

$\lim_{x \to 0^-} f(x) = -\infty$ (left-handed limit)

Calculus I: Limits & Derivatives

1. Limits
2. **Computing Derivatives with Differentiation**
3. Automatic Differentiation

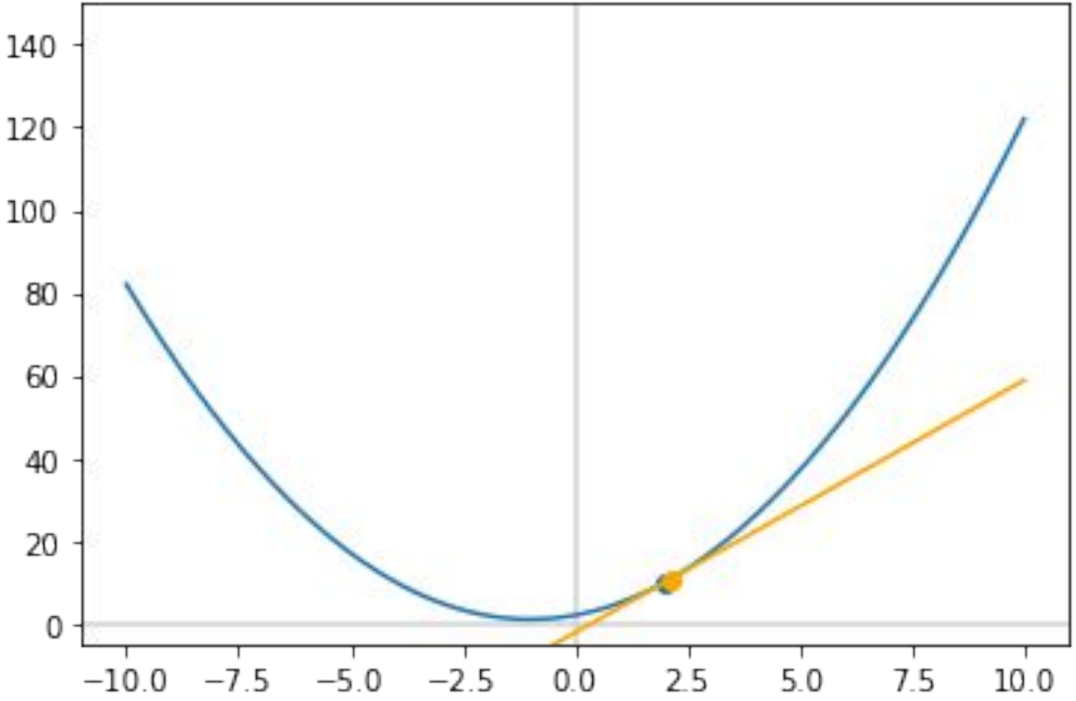
Segment 2: Derivatives & Differentiation

- The Delta Method
- The Differentiation Equation
- Differentiation Notation
- The Power Rule
- The Constant Multiple Rule
- The Sum Rule
- The Product Rule
- The Quotient Rule
- The Chain Rule

The Delta Method

Hands-on code demo

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$



Derivative Notation

$$y = f(x)$$

differentiation operator:					$\frac{d}{dx}$	D_x
first derivative of y w.r.t. x:	$\dot{y}$	y'	$f'(x)$	$\frac{dy}{dx}$	$\frac{d}{dx}f(x)$	$D_x f$
second derivative of y w.r.t. x:	$\ddot{y}$	y''	$f''(x)$	$\frac{d^2y}{dx^2}$	$\frac{d^2}{dx^2}f(x)$	$D_x^2 f$

Derivative of a Constant

Assuming c is constant: $\frac{d}{dx} c = 0$

The Power Rule

$$\frac{d}{dx} x^n = n x^{n-1}$$

The Constant Multiple Rule

$$\frac{d}{dx}(cy) = c\frac{d}{dx}(y) = c\frac{dy}{dx}$$

The Constant Multiple Rule *Simplified*

$$y = 2x^4$$

$$\frac{dy}{dx} = 2(4x^3) = 8x^3$$

The Sum Rule

$$\frac{d(y+w)}{dx} = \frac{dy}{dx} + \frac{dw}{dx}$$

Exercises

1. Differentiate $y = -5x^3$
2. Differentiate $y = 2x^2 + 2x + 2$
3. Differentiate $y = 10x^5 - 6x^3 - x - 1$
4. Use paper and pencil (i.e., the derivative rules) to find the slope of $y = x^2 + 2x + 2$ where $x = 2$
5. Repeat Exercise 4 but where $x = -1$

Solutions

① Differentiate $y = -5x^3$

$$\frac{dy}{dx} = -5(3x^2) = -15x^2$$

② Differentiate $y = 2x^2 + 2x + 2$

$$\frac{dy}{dx} = 2(2x^1) + 2(1x^0) + 0$$
$$= 2(2x) + 2(1)$$
$$= 4x + 2$$

Solutions

③ Differentiate $y = 10x^5 - 6x^3 - x - 1$

$$\frac{dy}{dx} = 10(5x^4) - 6(3x^2) - 1x^0 - 0$$

$$= 50x^4 - 18x^2 - 1$$

Solutions

(4) Using paper & pencil, find the slope of the curve $y = x^2 + 2x + 2$ where $x = 2$ as well as where $x = -1$

$$\frac{dy}{dx} = 2x + 2$$

$$\therefore m_1 = 2(2) + 2 = 4 + 2 = 6$$

$$m_2 = 2(-1) + 2 = -2 + 2 = 0$$

The Product Rule

$$\frac{d(wz)}{dx} = w\frac{dz}{dx} + z\frac{dw}{dx}$$

Useful:

- During calculations involving several rules
- Whenever product is incalculable pre-diff.

The Quotient (i.e., Fraction) Rule

$$\frac{d}{dx}\left(\frac{w}{z}\right) = \frac{z\frac{dw}{dx} - w\frac{dz}{dx}}{z^2}$$

The Chain Rule

- Many applications within ML
  - Gradient descent in general
  - Critical for backpropagation algo used to train neural nets
- Based on *nested functions* (a.k.a. *composite functions*):
  - Let's say $y = (5x + 25)^3$
  - We can let $u = 5x + 25$
  - In that case, $y = u^3$
  - y is a function of u, and u is a function of x
- *Chain rule* is easy way to find derivative of nested function:

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

The Chain Rule

$$y = (2x^2 + 8)^2$$

$$y = u^2$$

$$\frac{dy}{du} = 2u$$

$$= 2(2x^2 + 8)$$

$$= 4x^2 + 16$$

$$u = 2x^2 + 8$$

$$\frac{du}{dx} = 4x$$

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

$$= (4x^2 + 16)(4x)$$

$$= 16x^3 + 64x$$

Exercises

1. Use the product rule to find $f'(x)$ where $y = (2x^2 + 6x)(2x^3 + 5x^2)$
2. Use the quotient rule to find y' where $y = 6x^2/(2 - x)$
3. Differentiate $y = (3x + 1)^2$
4. Find y' where $y = (x^2 + 5x)^6$
5. Differentiate $f(x) = 1/((\mathrm{x}^4 + 1)^5 + 7)$

Solutions

1.

$$\frac{d(wz)}{dx} = w\frac{dz}{dx} + z\frac{dw}{dx}$$

$$y = (2x^2 + 6x)(2x^3 + 5x^2)$$

$$w = 2x^2 + 6x \qquad z = 2x^3 + 5x^2$$

$$\frac{dw}{dx} = 4x + 6 \qquad \frac{dz}{dx} = 6x^2 + 10x$$

$$\frac{d(wz)}{dx} = w\frac{dz}{dx} + z\frac{dw}{dx}$$

$$= (2x^2 + 6x)(6x^2 + 10x) + (2x^3 + 5x^2)(4x + 6)$$

$$= 12x^4 + 20x^3 + 36x^3 + 60x^2 + 8x^4 + 12x^3 + 20x^3 + 30x^2$$

$$= 20x^4 + 88x^3 + 90x^2$$

Solutions

2.

$$\frac{d}{dx}\left(\frac{w}{z}\right) = \frac{z\frac{dw}{dx} - w\frac{dz}{dx}}{z^2}$$

$$y = \frac{6x^2}{2-x} \qquad w = 6x^2 \qquad z = 2-x$$

$$\frac{dw}{dx} = 12x \qquad \frac{dz}{dx} = -1$$

$$y' = \frac{d}{dx}\left(\frac{w}{z}\right) = \frac{(2-x)(12x) - (6x^2)(-1)}{(2-x)^2}$$

$$= \frac{24x - 12x^2 + 6x^2}{(2-x)^2}$$

$$= \frac{24x - 6x^2}{(2-x)^2}$$

Solutions

3.

$$y = (3x+1)^2$$

$$y = u^2$$

$$\frac{dy}{du} = 2u$$
$$= 2(3x+1)$$
$$= 6x+2$$

$$u = 3x+1$$

$$\frac{du}{dx} = 3$$

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$
$$= (6x+2)(3) = 18x+6$$

Solutions

4.

$$y = (x^2 + 5x)^6$$

$$y = u^6$$

$$\frac{dy}{du} = 6u^5$$

$$= 6(x^2 + 5x)^5$$

$$u = x^2 + 5x$$

$$\frac{du}{dx} = 2x + 5$$

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

$$= \left(6(x^2 + 5x)^5\right)(2x + 5)$$

Solutions

5.

$$y = \frac{1}{(x^4+1)^5+7}$$

$$y = \frac{1}{u} = u^{-1}$$

$$\frac{dy}{du} = (-1)u^{-1-1}$$

$$= -u^{-2}$$

$$= -(t^5+7)^{-2}$$

$$= -((x^4+1)^5+7)^{-2}$$

$$u = t^5 + 7$$

$$\frac{du}{dt} = 5t^4$$

$$= 5(x^4+1)^4$$

$$t = x^4 + 1$$

$$\frac{dt}{dx} = 4x^3$$

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dt}\frac{dt}{dx}$$

$$= \left(-((x^4+1)^5+7)^{-2}\right)\left(5(x^4+1)^4\right)\left(4x^3\right)$$

Power Rule on a Function Chain

$$\frac{d}{dx} u^n = n u^{n-1} \frac{du}{dx}$$

Exercise

Repeat questions 4 and 5 using the power rule.

Calculus I: Limits & Derivatives

1. Limits
2. Computing Derivatives with Differentiation
3. **Automatic Differentiation**

Segment 3: Automatic Differentiation

- Autodiff with PyTorch
- Autodiff with TensorFlow
- Machine Learning via Differentiation
- Cost (or Loss) Functions
- The Future: Differentiable Programming

Automatic Differentiation

- A.K.A.:
  - Autodiff
  - Autograd
  - Computational diff.
  - Reverse mode diff.
  - Algorithmic diff.
- Distinct from classical methods:
  - Numerical diff. (delta method; introduces rounding errors)
  - Symbolic diff. (algebraic rules; computationally inefficient)
- Relative to classical methods, better handles:
  - Functions with many inputs (common in ML)
  - Higher-order derivatives

Automatic Differentiation

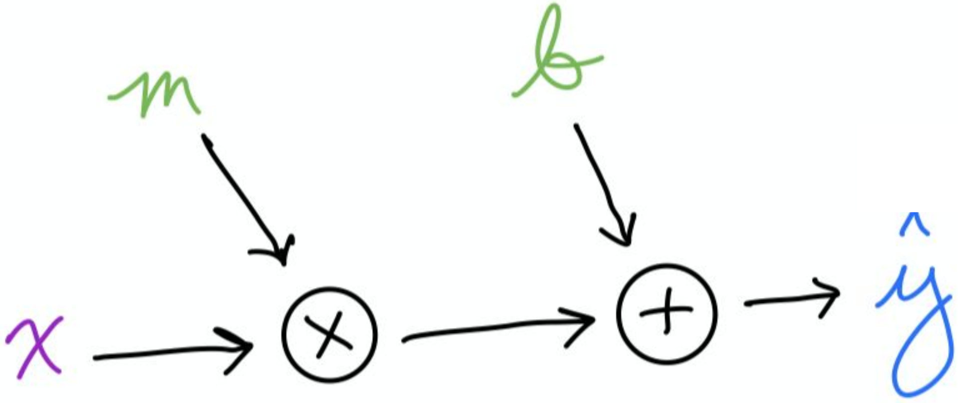
- Application of chain rule (typically *partial* derivative) to sequence (forward pass) of arithmetic operations
- Whereas chain rule by hand typically begins at most-nested function, autodiff proceeds from outermost function inward
- Small constant factor more compute than forward pass (at most)

$$\frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx}$$

Hands-on code demo

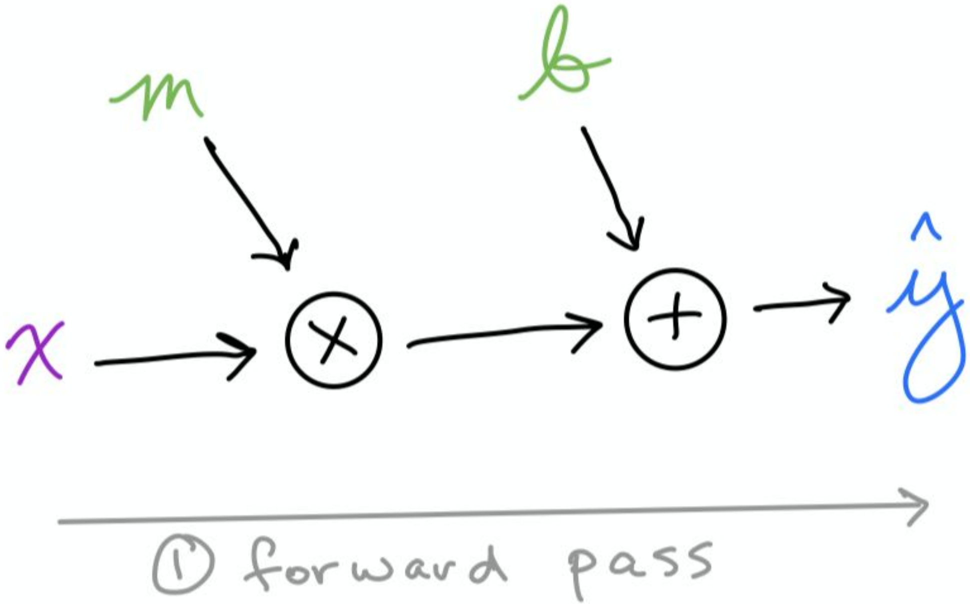
Fitting a Line with Machine Learning

- Line equation $y = mx + b$ as *directed acyclic graph* (DAG)
- Nodes are input, output, parameters, or operations
- Directed edges ("arrows") are tensors (N.B.: non-operation nodes can be tensors too)

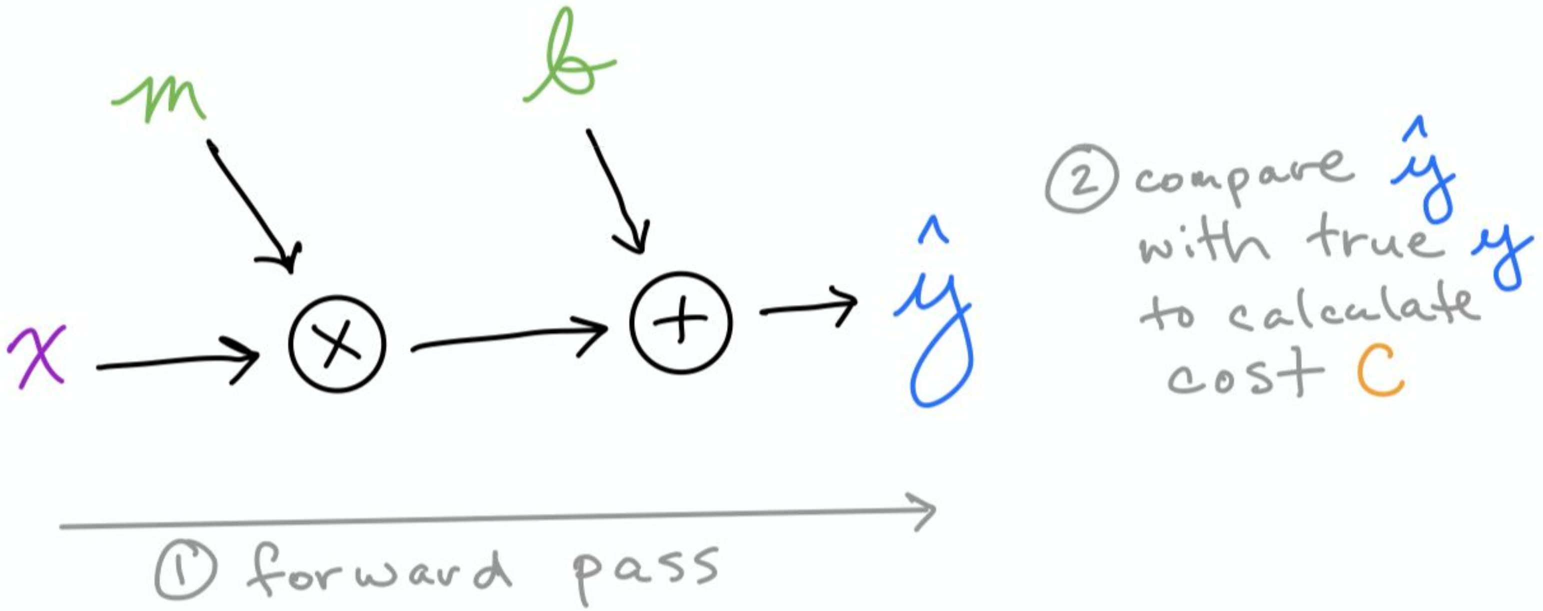


Hands-on code demo: `regression-in-pytorch.ipynb`

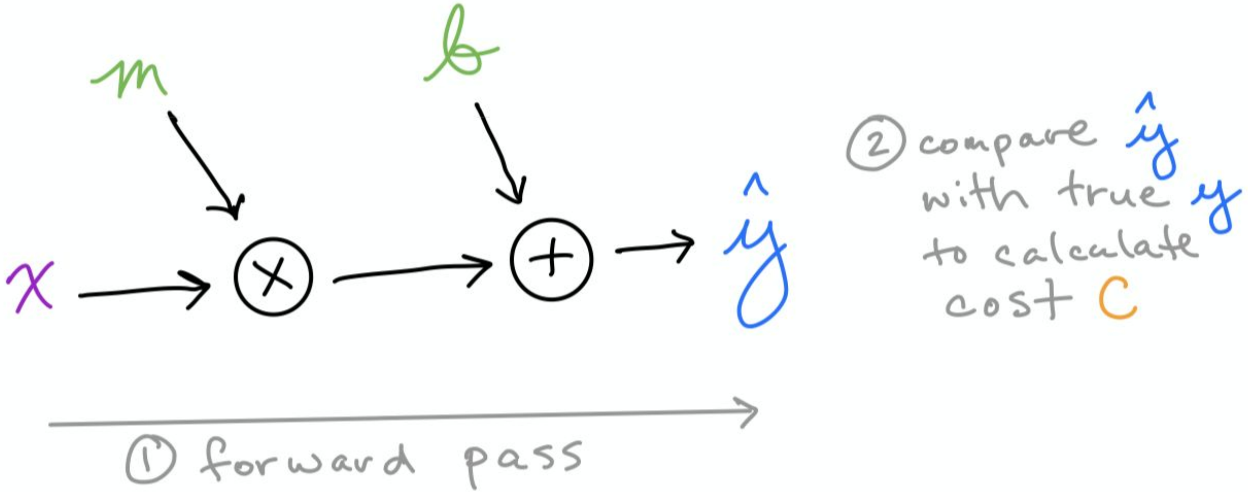
Machine Learning



Machine Learning



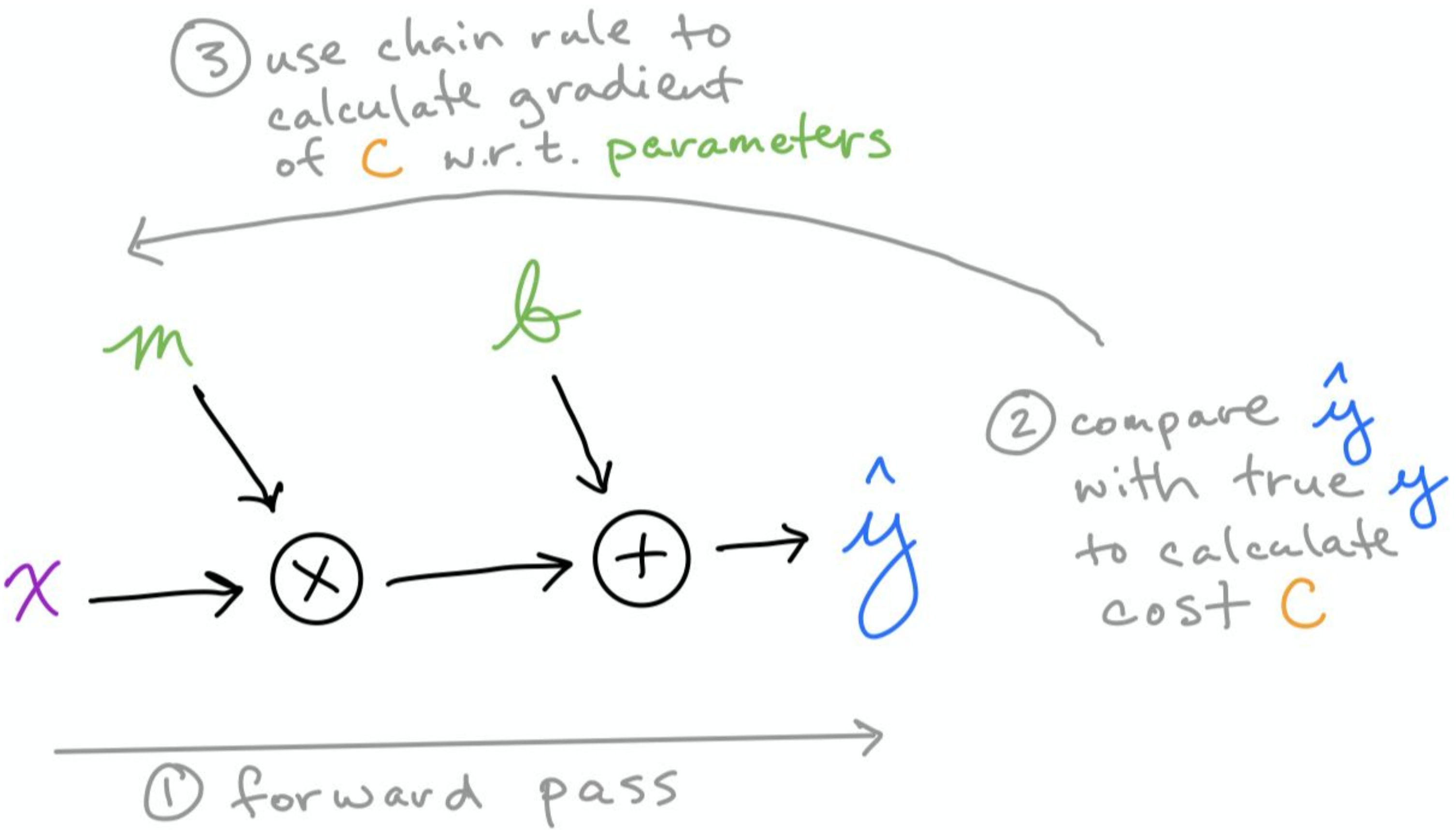
The Chain Rule’s Necessity



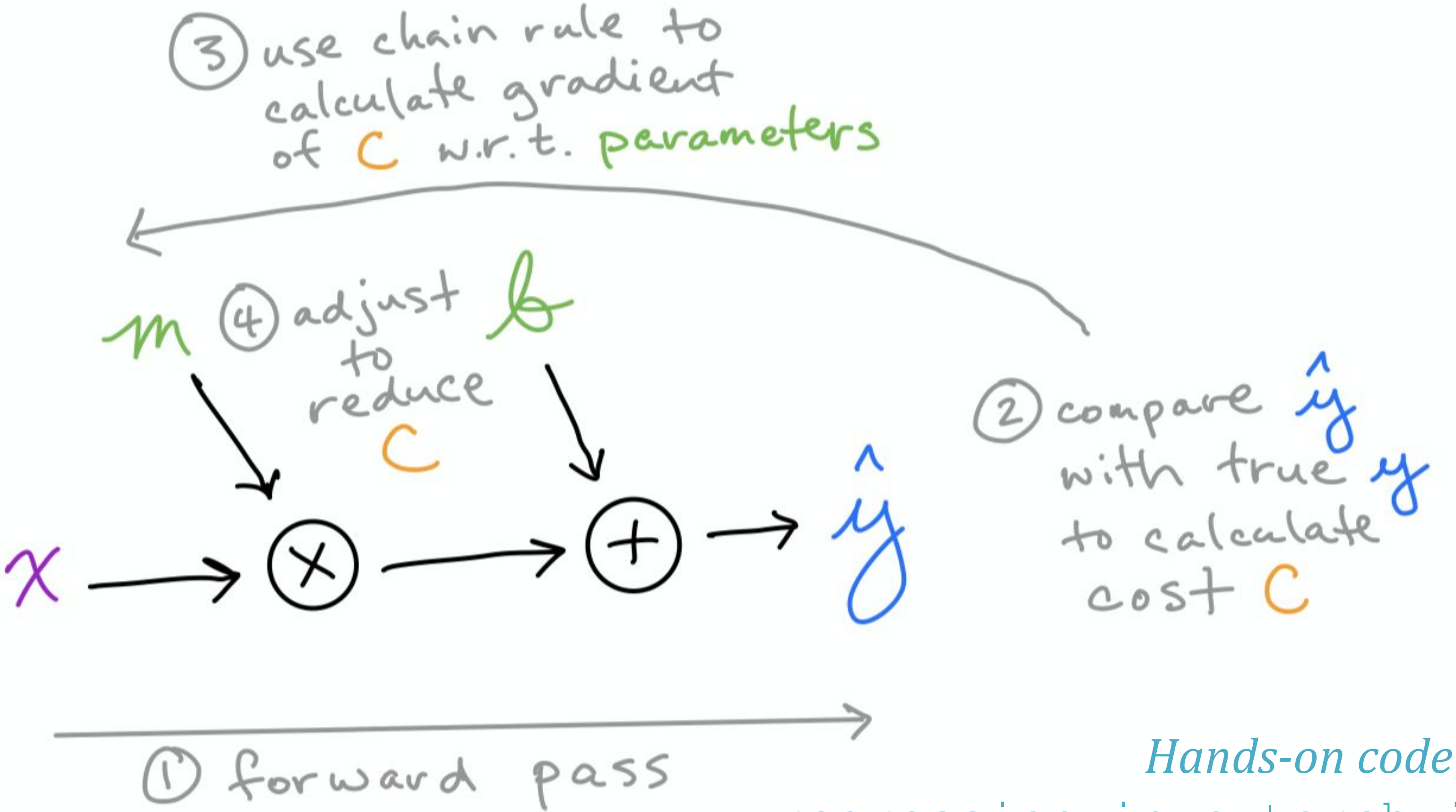
$$\hat{y} = f(x, m, b) \qquad C = g(\hat{y}, y)$$

$$C = g(f(x, m, b), y)$$

Machine Learning



Machine Learning



Hands-on code demo:
`regression-in-pytorch.ipynb`

Exercises

1. Use PyTorch (or TensorFlow, if you like) to find the slope of $y = x^2 + 2x + 2$ where $x = 2$.

2. Use the *Regression in PyTorch* notebook to simulate a new linear relationship between y and x, and then fit the parameters m and b.

3. Read about how *differential programming*, wherein computer programs can be differentiated, could be common soon (perhaps thanks to *quantum* ML; see `pennylane.ai`).

Next Subject: *Calculus II*

- Partial derivatives
  - Required for autodiff from regressions to deep neural nets
- Integrals
  - Calculate area under a curve, e.g., for ROC AUC metric

POLL *with Multiple Answers Possible*

What follow-up topics interest you most?

- Linear Algebra
- More Calculus
- Probability / Statistics
- Computer Science (e.g., algorithms, data structures)
- Machine Learning Basics
- Advanced Machine Learning, incl. Deep Learning
- Something Else

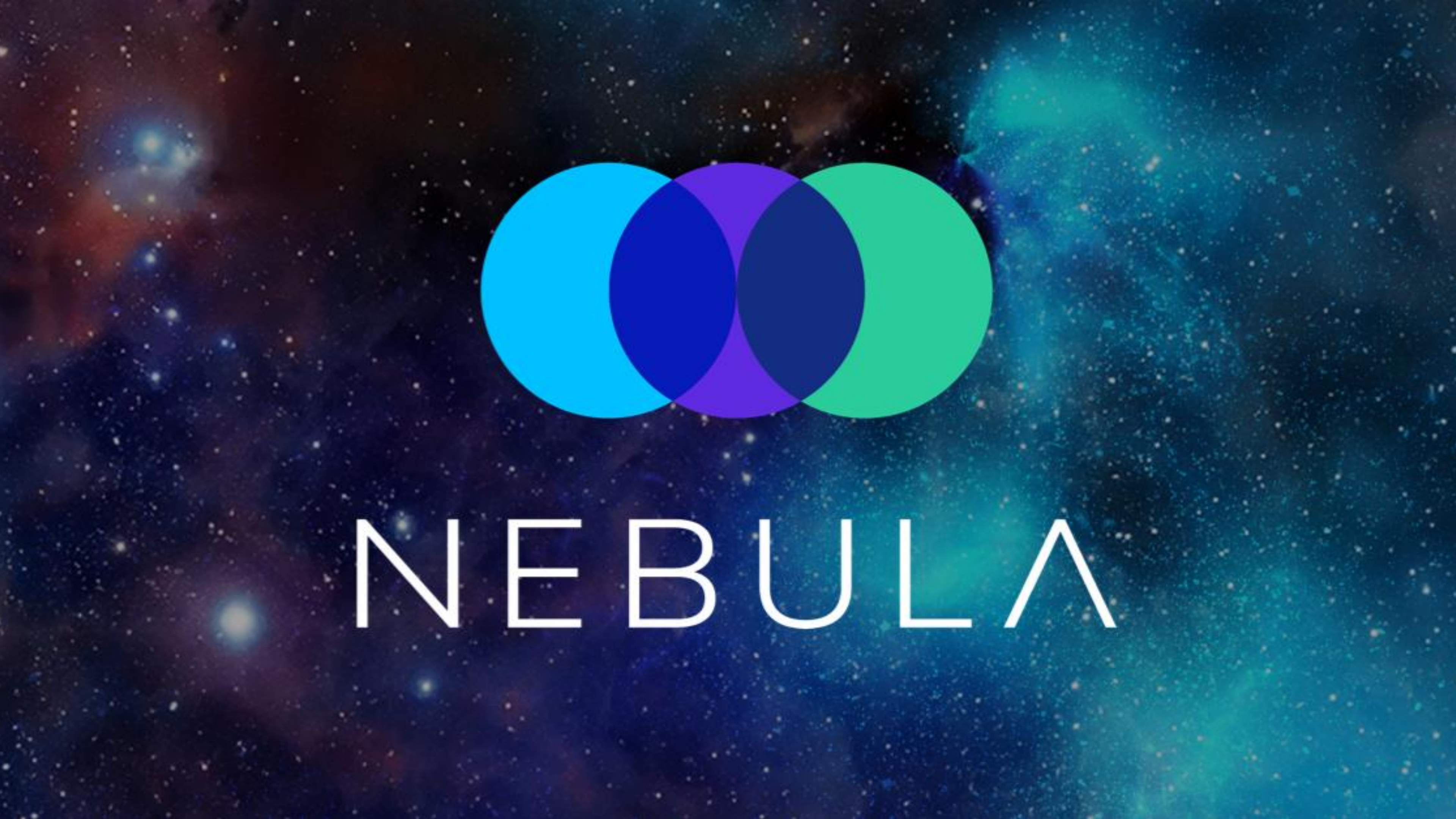
Stay in Touch

`jonkrohn.com` to sign up for email newsletter

`linkedin.com/in/jonkrohn`

`jonkrohn.com/youtube`

`twitter.com/JonKrohnLearns`


NEBULA

PLACEHOLDER
FOR:

5-Minute Timer

PLACEHOLDER
FOR:

10-Minute Timer

PLACEHOLDER
FOR:

15-Minute Timer