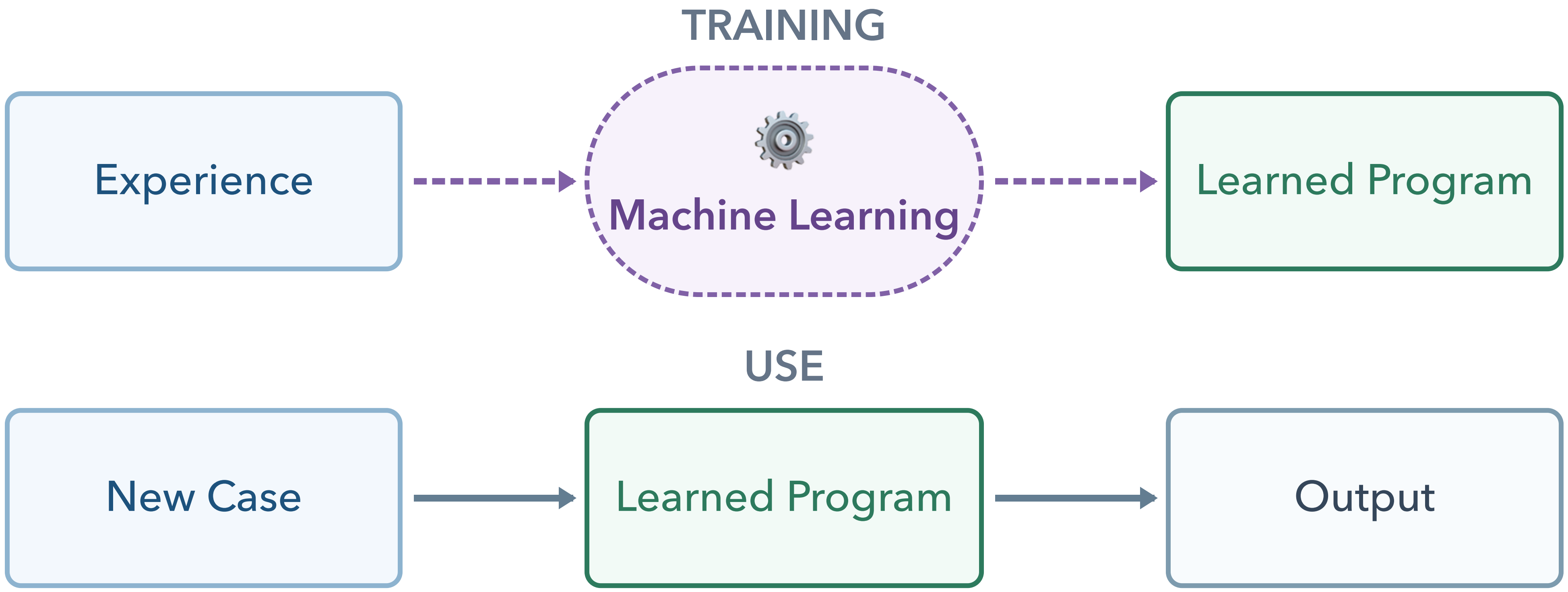
CMPUT 267: Machine Learning I · Fall 2026

Math Review

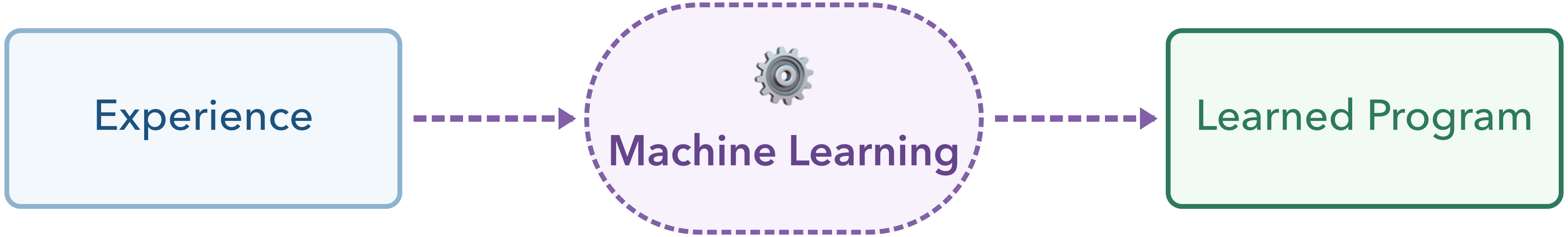
Xi Ye

2026-09-03

Recall: Machine Learning



Example: House-Price Prediction

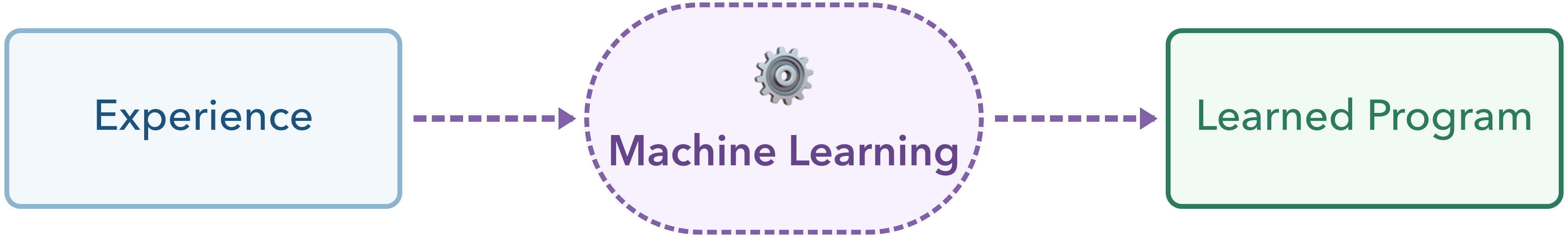


Experience: A dataset sampled from an (unknown) distribution.

Size (m^2)	**Age (years)**	**Rooms**	**Price ($)**
90	25	3	310,000
140	8	4	510,000
115	40	3	355,000

Each row is one data point. Size, Age, and Rooms are its **features**; Price is its **label**.

Example: House-Price Prediction



Learned program: A program (a function) that takes the size, age, and number of rooms of a house as input and predicts its price.

Machine learning: The process of using the observed house data to produce such a program.

This Lecture: Math Review

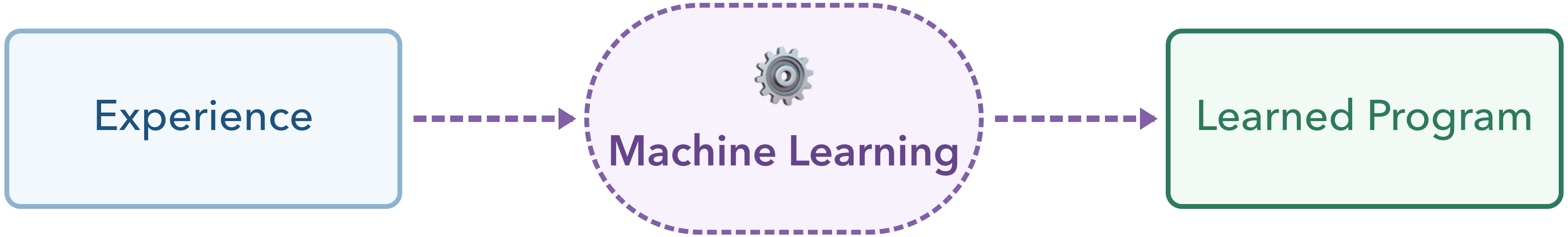
Today we will review:

1. Sets, tuples, and functions
2. Datasets and predictors
3. Summations, integrals, and derivatives

This is a **review lesson**. We will move quickly and assume the necessary prerequisites.

Caveat: we will introduce these concepts as needed for this course, but not always in the most mathematically rigorous way.

More Standard Description of ML



We will use the following terminology throughout the course:

Sets

A **set** is a collection of **distinct, unordered** objects.

- Every object in a set is called an **element**.
- The order in which elements are written does not matter.
- Elements may be anything (e.g., numbers, words, or other sets).

Examples

$$\mathbb{N} = \{0, 1, 2, \ldots\}, \qquad \mathbb{R}, \qquad \emptyset = \{\}, \qquad \{\{1, 2\}, \{3\}\}.$$

Elements, Cardinality, and Subsets

For a set $\mathcal{A}$:

- $x \in \mathcal{A}$ means that x is an **element** of $\mathcal{A}$.
- $|\mathcal{A}|$ is the **cardinality** of $\mathcal{A}$: its number of elements.
- $\mathcal{B} \subseteq \mathcal{A}$ means that $\mathcal{B}$ is a **subset** of $\mathcal{A}$: every element of $\mathcal{B}$ is also in $\mathcal{A}$.

Example

Let $\mathcal{A} = \{2, 4, 6\}$. Then

$$2 \in \mathcal{A}, \qquad |\mathcal{A}| = 3, \qquad \{2, 6\} \subseteq \mathcal{A}.$$

Element versus subset: $2 \in \mathcal{A}$, whereas $\{2\} \subseteq \mathcal{A}$.

Operations on Sets

Let $\mathcal{A} = \{1, 2, 3\}$ and $\mathcal{B} = \{3, 4\}$.

Operation	**Meaning**	**Result**
$\mathcal{A} \cup \mathcal{B}$	In $\mathcal{A}$ or $\mathcal{B}$	$\{1, 2, 3, 4\}$
$\mathcal{A} \cap \mathcal{B}$	In both sets	$\{3\}$
$\mathcal{A} \setminus \mathcal{B}$	In $\mathcal{A}$ but not $\mathcal{B}$	$\{1, 2\}$

Power Sets

The **power set** $\mathcal{P}(\mathcal{A})$ is the set of all subsets of $\mathcal{A}$.

If $\mathcal{A} = \{a, b\}$, then

$$\mathcal{P}(\mathcal{A}) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}.$$

> If $|\mathcal{A}| = k$, then $|\mathcal{P}(\mathcal{A})| = 2^k$.

Intervals (Sets of Real Numbers)

Intervals describe continuous ranges of possible real values.

Interval	**Endpoint rule**	**Set-builder form**
$[1, 4]$	Include both endpoints	$\{x \in \mathbb{R} \mid 1 \leq x \leq 4\}$
$(1, 4)$	Exclude both endpoints	$\{x \in \mathbb{R} \mid 1 < x < 4\}$
$[1, 4)$	Include 1; exclude 4	$\{x \in \mathbb{R} \mid 1 \leq x < 4\}$
$(1, 4]$	Exclude 1; include 4	$\{x \in \mathbb{R} \mid 1 < x \leq 4\}$

Set-Builder Notation

$$\{x \in \mathbb{N} \mid x \text{ is even}\}$$

Candidate	Source Set	Condition
x	$x \in \mathbb{N}$	x is even
An arbitrary object	Where candidates come from	Which candidates are retained

The vertical bar $\mid$ means **such that**.

$$\{x \in \mathbb{N} \mid x \text{ is even}\} = \{0, 2, 4, 6, \ldots\}$$

Similarly, $\{x \in \mathbb{R} \mid -1 < x \leq 2\} = (-1, 2]$.

Quick Check: Set-Builder Notation

Which expression represents the positive real numbers?

- **A** $\{x \in \mathbb{R} \mid x > 0\}$
- **B** $\{x \in \mathbb{N} \mid x > 0\}$
- **C** $\{x > 0 \mid x \in \mathbb{R}\}$
- **D** $\{0, \infty\}$

A. Candidates come from $\mathbb{R}$, and the condition retains those greater than zero.

Tuples

Set

$$\{200, 350, 500\}$$

- Unordered
- Values are distinct
- Uses braces

Tuple

$$(350, 200, 350)$$

- Ordered
- Repetition is allowed
- Uses parentheses

⚠ $(1, 2) \neq (2, 1)$, but $\{1, 2\} = \{2, 1\}$.

Cartesian products

Suppose we want all ordered pairs formed by choosing one element from $\mathcal{X}$ and one from $\mathcal{Y}$.

The Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$ is

$$\mathcal{X} \times \mathcal{Y} = \{(x, y) \mid x \in \mathcal{X},\ y \in \mathcal{Y}\}.$$

For $\mathcal{X} = \{1, 2\}$ and $\mathcal{Y} = \{200, 300\}$,

$\mathcal{X} \times \mathcal{Y} =$

$$\{(1, 200), (1, 300), (2, 200), (2, 300)\}.$$

> Cartesian products combines set space.

> For finite sets, $|\mathcal{X} \times \mathcal{Y}| = |\mathcal{X}|\,|\mathcal{Y}|$.

Products with More Components

$$\mathcal{X}_1 \times \cdots \times \mathcal{X}_d = \{(x_1, \ldots, x_d) \mid x_j \in \mathcal{X}_j\}.$$

A house described by size, age, and number of rooms can be represented by a three-component tuple.

Vectors

Tuple

$$(120, 10, \text{brick})$$

Components may have different types. Arithmetic may not make sense.

Vector

$$\mathbf{x} = (120, 10, 3)^\top \in \mathbb{R}^3$$

Numerical components support addition, scaling, and other operations.

> Every vector can be viewed as a tuple, but not every tuple is a vector.

Vectors in $\mathbb{R}^d$

For the house example,

$$\mathbf{x} = \begin{pmatrix} 120 \\ 10 \\ 3 \end{pmatrix} \in \mathbb{R}^3.$$

For vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$ and scalar $c \in \mathbb{R}$,

$$\mathbf{x} + \mathbf{z} \in \mathbb{R}^d, \qquad c\mathbf{x} \in \mathbb{R}^d.$$

Transpose and Dot Product

The transpose turns a column vector into a row vector:

$$\mathbf{x}^\top = (x_1, \dots, x_d).$$

The dot product produces one real number:

$$\mathbf{w}^\top \mathbf{x} = \sum_{j=1}^{d} w_j x_j.$$

Example: Dot Product

$$\mathbf{x} = \begin{pmatrix} 120 \\ 10 \\ 3 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} 2 \\ -1 \\ 20 \end{pmatrix}.$$

The dot product combines the three componentwise products:

$$\mathbf{w}^\top \mathbf{x} =$$

$$\mathbf{w}^\top \mathbf{x} = 2(120) - 1(10) + 20(3) = 290.$$

Matrices Stack Row Vectors

Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$. A matrix can be formed by stacking their transposes:

$$\mathbf{M} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,d} \\ x_{2,1} & \cdots & x_{2,d} \end{bmatrix} \in \mathbb{R}^{2\times d}.$$

Each row is one transposed vector. Stacking house feature vectors produces a feature matrix.

Example: From a Table to a Matrix

Observed Houses:

Size (m^2)	**Age**	**Rooms**	**Price ($)**
90	25	3	310,000
140	8	4	510,000
115	40	3	355,000

Feature Matrix:

$$\mathbf{M} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \mathbf{x}_3^\top \end{bmatrix} = \begin{bmatrix} 90 & 25 & 3 \\ 140 & 8 & 4 \\ 115 & 40 & 3 \end{bmatrix}.$$

Matrix-Vector Multiplication

For $\mathbf{w} \in \mathbb{R}^d$, multiplying by $\mathbf{M}$ computes one dot product per row:

$$\mathbf{M}\mathbf{w} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \end{bmatrix} \mathbf{w} = \begin{pmatrix} \mathbf{x}_1^\top \mathbf{w} \\ \mathbf{x}_2^\top \mathbf{w} \end{pmatrix} \in \mathbb{R}^2.$$

For example,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} =$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1(1) + 2(2) \\ 3(1) + 4(2) \end{pmatrix} = \begin{pmatrix} 5 \\ 11 \end{pmatrix}.$$

Representing Dataset

A **dataset** is a collection of $(\text{feature}, \text{label})$ tuples, where each feature is a vector.

Feature Vector

$\mathbf{x}_1 = (90, 25, 3)^\top \in \mathcal{X}$

Label

$y_1 = 310 \in \mathcal{Y}$

Example

$(\mathbf{x}_1, y_1) \in \mathcal{X} \times \mathcal{Y}$

$$\mathcal{D} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n.$$

Functions

Functions

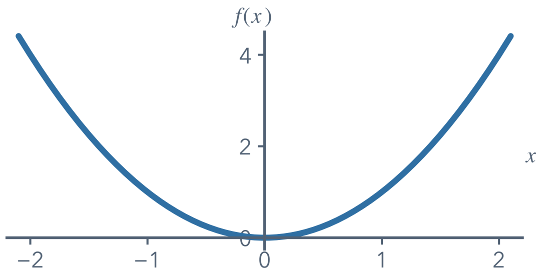
The predictor $\widehat{f}$ and the learner $\mathcal{A}$ are functions.

> A **function** f from a set $\mathcal{X}$ to a set $\mathcal{Y}$ assigns to each element of $\mathcal{X}$ exactly one element of $\mathcal{Y}$, denoted as $f : \mathcal{X} \to \mathcal{Y}$.

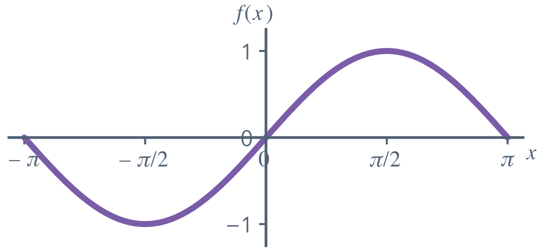
- $\mathcal{X}$ is the **Domain**; $\mathcal{Y}$ is the **Codomain**
- The **range** is the set of values that f actually produces. It may be smaller than $\mathcal{Y}$. For example, let $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. The range of f is $[0, \infty)$.

Example: Functions

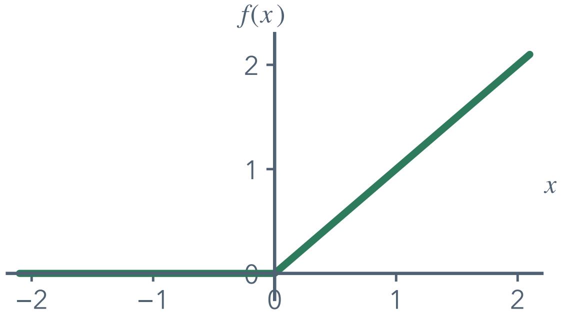
$$f(x) = x^2$$



$$f(x) = \sin x$$



$$f(x) = \begin{cases} 0, & x \leq 0, \\ x, & x > 0, \end{cases} \text{ or } f(x) = \max(0, x).$$



Representing Predictor and Learner

Predictor

$$f : \mathcal{X} \to \mathcal{Y}$$

Takes a feature vector and produces a prediction.

Learner

$$\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$$

Takes a dataset and produces a predictor.

Summations, Intergrations, Derivatives

Summations

The symbol $\sum$ denotes summation:

$$\sum_{x \in \mathcal{X}} f(x) = \sum_{i=1}^{n} f(x_i) = f(x_1) + \cdots + f(x_n)$$

This adds $f(x_i)$ for every discrete input $x_i \in \mathcal{X}$.

For example, let $f(x) = x^2$ and $\mathcal{X} = \{1, 2, 3\}$:

$$\sum_{x \in \mathcal{X}} f(x) = f(1) + f(2) + f(3) = 1 + 4 + 9 = 14.$$

> Use $\sum$ when the inputs are discrete: we can list them and add the values.

Intergrations

If x can be **any real value** in $[a, b]$, we cannot list every input. How can we accumulate the corresponding values of $f(x)$?

We divide the interval into narrow pieces of width Δx and form a weighted sum.

The **integral** of $f(x)$ with respect to x over $[a, b]$ is denoted by

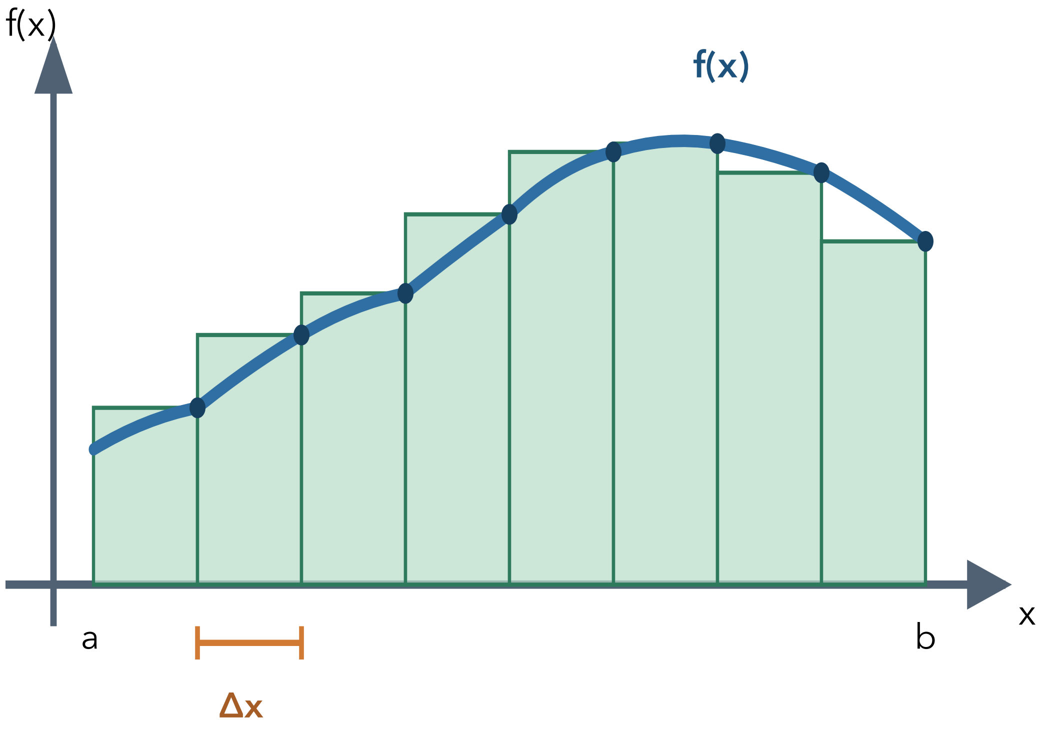
$$\int_a^b f(x)\, dx = \lim_{\Delta x \to 0} \sum_i f(x_i)\, \Delta x.$$

Intergrations: Explanations

We can interpret $\int_a^b f(x)\,dx$ as the total area under the curve of $f(x)$ between (a, b).

Partition $[a, b]$ into narrow parts of width Δx. The ith rectangle has height $f(x_i)$, so its area is

$$\underbrace{f(x_i)}_{\text{height}}\underbrace{\Delta x}_{\text{width}}.$$



The integral:

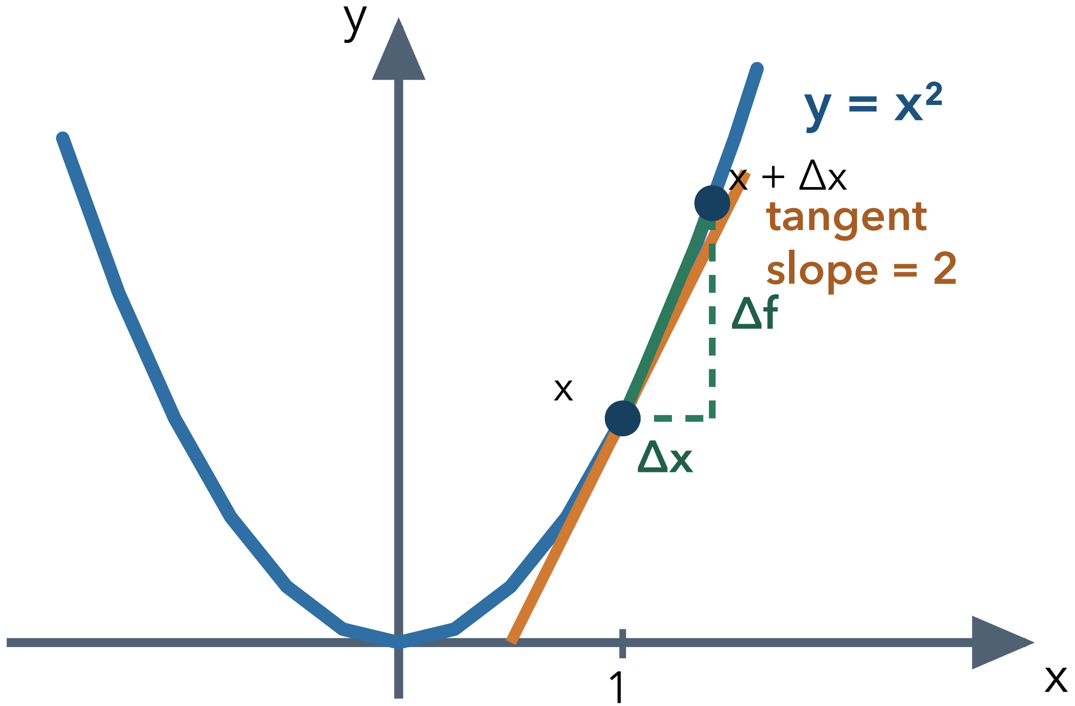
$$\int_a^b f(x)\,dx \approx \sum_i f(x_i)\,\Delta x.$$

Derivatives

The **derivative** of f at x measures the local rate at which $f(x)$ changes as x changes. It is denoted by $f'(x)$ or $\frac{df}{dx}(x)$.

For a finite change Δx,

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



> This is the slope of the **tangent line** at x.

Example: Derivatives

Function	**Derivative**
c	0
x^k	kx^{k-1}
e^x	e^x
$af(x) + bg(x)$	$af'(x) + bg'(x)$

The Chain Rule

A **composed function** has the form $f(x) = g(h(x))$. Its derivative is

$$f'(x) = g'(h(x))h'(x).$$

For example, let

$$h(x) = x^2, \qquad g(x) = \exp(x), \qquad f(x) = g(h(x)) = \exp(x^2).$$

$$\begin{aligned} f'(x) &= g'(h(x))h'(x) \\ &= \exp(x^2)(2x). \end{aligned}$$

ⓘ Therefore, $\boxed{f'(x) = 2x\exp(x^2)}$.

Partial Derivatives

A function may have several inputs. A **partial derivative** measures how its output changes when one input changes and the others are held fixed.

For $f(x_1, \ldots, x_d)$, the partial derivative with respect to x_i is denoted as

$$\frac{\partial f}{\partial x_i}(x_1, \ldots, x_d).$$

Example: Partial Derivatives

Consider

$$f(x_1, x_2) = 2x_1x_2^2.$$

Change x_1

Hold x_2 fixed

$$\frac{\partial f}{\partial x_1} = 2x_2^2$$

Change x_2

Hold x_1 fixed

$$\frac{\partial f}{\partial x_2} = 4x_1x_2$$

Gradients

The **gradient** collects every partial derivative of a scalar function into one vector.

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \dfrac{\partial f}{\partial x_d}(\mathbf{x}) \end{pmatrix} \in \mathbb{R}^d.$$

For $f(x_1, x_2) = 2x_1 x_2^2$,

$$\nabla f(x_1, x_2) = \begin{pmatrix} 2x_2^2 \\ 4x_1 x_2 \end{pmatrix}, \qquad \nabla f(1, 2) = \begin{pmatrix} 8 \\ 8 \end{pmatrix}.$$

Refer to course note (Chap 1 - 2) for more detailed review

Next Lecture

More math review: Probability

So that we can will describe uncertainty and randomness more formally.