1.3 NOTATION AND DEFINITIONS

CSC 462 [THPMLB-2]

DATA STRUCTURES -1

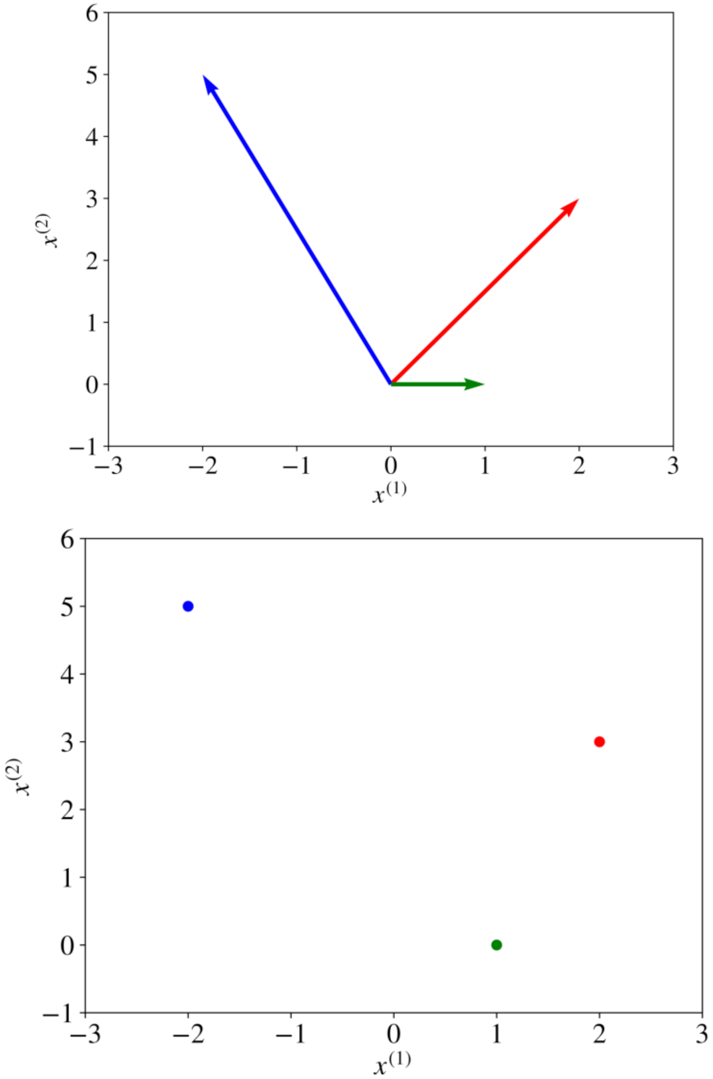
- A scalar is a simple numerical value, like 15 or −3.25.

- Variables or constants that take scalar values are denoted by an italic letter, like x or a.

- A **vector** is an ordered list of scalar values, called attributes. We denote a vector as a bold character, for example, **x** or **w**.

- We denote an attribute of a vector as an italic value with an index, like this: $w^{(j)}$ or $x^{(j)}$. The index j denotes a specific **dimension** of the vector, the position of an attribute in the list. For instance, in the vector **a** shown in red in Figure 1, $a^{(1)} = 2$ and $a^{(2)} = 3$.

- A variable can have two or more indices, like this:$x_j^{(i)}$ or like this $x_{i,j}^{(k)}$



DATA STRUCTURES -2

- A **matrix** is a rectangular array of numbers arranged in rows and columns.
  - Example $\begin{bmatrix} 2 & 4 & -3 \\ 21 & -6 & -1 \end{bmatrix}$
  - Matrices are denoted with bold capital letters, such as **A** or **W**.
- A **set** is an unordered collection of unique elements. We denote a set as a calligraphic capital character, for example, S.
  - for example, $\{1, 3, 18, 23, 235\}$ or $\{x_1, x_2, \ldots, x_n\}$
  - If a set includes all values between *a* and *b*, including *a* and *b*, it is denoted using brackets as [*a*, *b*].
  - If the set doesn't include the values *a* and *b*, such a set is denoted using parentheses like this: (*a*, *b*).

CAPITAL SIGMA AND PI

- $\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$
- $\sum_{j=1}^{m} x^{(j)} = x^{(1)} + \cdots + x^{(m)}$
- $\prod_{i=1}^{n} x_i = x_1 \cdot \ \ldots \cdot x_n$
- $a \cdot b$ means a multiplied by b and in short denoted by ab

OPERATIONS ON SETS AND VECTORS -1

- Operations on sets example: $S' = \{x^2 | x \in S, x > 3\}$
- Operations on vectors:
  - Sum of two vectors $\boldsymbol{x} + \boldsymbol{z}$ $is\ defined\ as\ the\ vector$ $[x^{(1)} + z^{(1)}, x^{(2)} + z^{(2)}, \dots, x^{(m)} + z^{(m)}]$
  - Difference of two vectors $\boldsymbol{x} - \boldsymbol{z}$ $is\ defined\ as\ the\ vector$ $[x^{(1)} - z^{(1)}, x^{(2)} - z^{(2)}, \dots, x^{(m)} - z^{(m)}]$
  - A vector multiplied by a scalar is a vector: $\boldsymbol{x}c = [cx^{(1)}, cx^{(2)}, \dots, cx^{(m)}]$
  - A **dot-product** of two vectors is a scalar. Example, $\boldsymbol{w} \cdot \boldsymbol{x} = \boldsymbol{wx} = \sum_{i=1}^{m} w^{(i)} x^{(i)}$

OPERATIONS ON SETS AND VECTORS -2

- The multiplication of a matrix **W** by a vector **x** results in another vector.

$$\mathbf{W}\mathbf{x} = \begin{bmatrix} w^{(1,1)} & w^{(1,2)} & w^{(1,3)} \\ w^{(2,1)} & w^{(2,2)} & w^{(2,3)} \end{bmatrix} \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \end{bmatrix}$$

$$\stackrel{\text{def}}{=} \begin{bmatrix} w^{(1,1)}x^{(1)} + w^{(1,2)}x^{(2)} + w^{(1,3)}x^{(3)} \\ w^{(2,1)}x^{(1)} + w^{(2,2)}x^{(2)} + w^{(2,3)}x^{(3)} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{w}^{(1)}\mathbf{x} \\ \mathbf{w}^{(2)}\mathbf{x} \end{bmatrix}$$

- When the vector is on the left side of the matrix in the multiplication, then it has to be **transposed** before we multiply it by the matrix.
  - If $\boldsymbol{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix}$, then $\boldsymbol{x}^T = [x^{(1)}, x^{(2)}]$

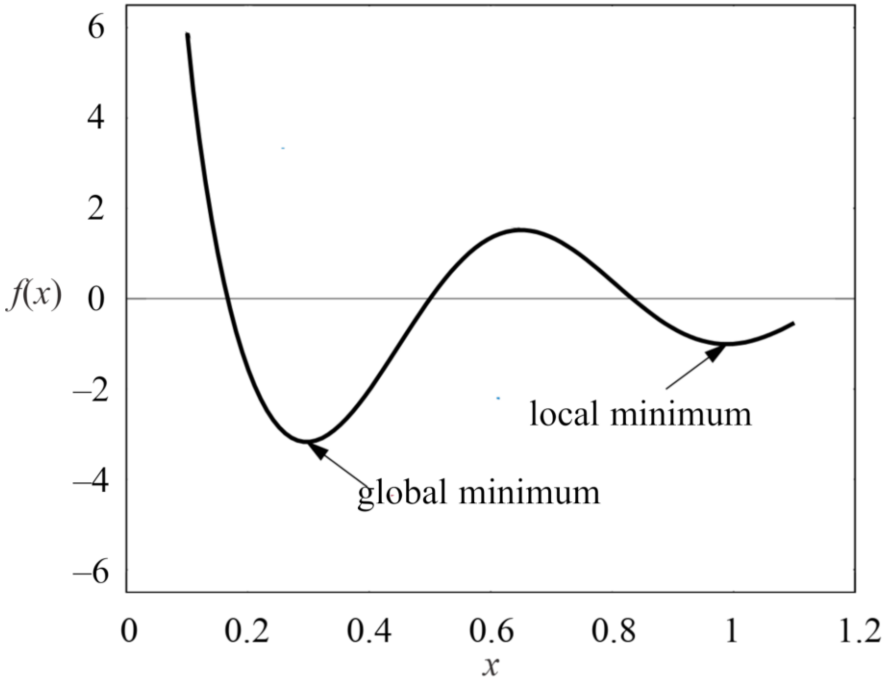
- The multiplication of the vector $\boldsymbol{x}$ by the matrix $\boldsymbol{W}$ is given by $\boldsymbol{x}^T\boldsymbol{W}$

$$\mathbf{x}^\top\mathbf{W} = \begin{bmatrix} x^{(1)} & x^{(2)} \end{bmatrix} \begin{bmatrix} w^{(1,1)} & w^{(1,2)} & w^{(1,3)} \\ w^{(2,1)} & w^{(2,2)} & w^{(2,3)} \end{bmatrix}$$

$$\stackrel{\text{def}}{=} \left[w^{(1,1)}x^{(1)} + w^{(2,1)}x^{(2)}, w^{(1,2)}x^{(1)} + w^{(2,2)}x^{(2)}, w^{(1,3)}x^{(1)} + w^{(2,3)}x^{(2)} \right]$$

FUNCTIONS

- A **function** is a relation that associates each element x of a set X, the **domain** of the function, to a single element y of another set Y, the **codomain** of the function.
- If the function is called f, this relation is denoted $y = f(x)$.
- $f(x)$ has a **local minimum** at $x = c$ if $f(x) \geq f(c)$ for every x in some open interval around $x = c$.
- The minimal value among all the local minima is called the **global minimum**.
- A vector function, denoted as $\boldsymbol{y} = \boldsymbol{f}(x)$ is a function that returns a vector $\boldsymbol{y}$. It can have a vector or a scalar argument.



MAX AND ARG MAX

Given a set of values $A = \{a_1, a_2, \ldots, a_n\}$,

- The operator $\max_{\{a \in A\}} f(a)$ returns the highest value $f(a)$ for all elements in the set A.
- The operator $\arg\max_{\{a \in A\}} f(a)$ returns the element of the set A that maximizes $f(a)$.
- Sometimes, when the set is implicit or infinite, we can write $\max f(a)$ or $\arg\max f(a)$.
- Operators $\min$ and $\arg\min$ operate in a similar manner.

ASSIGNMENT OPERATOR

- The expression $a \leftarrow f(x)$ means that the variable a gets the new value: the result of $f(x)$.
- Similarly, $\boldsymbol{a} \leftarrow [a_1, a_2]$ means that the vector variable $\boldsymbol{a}$ gets the two-dimensional vector value $[a_1, a_2]$.

Definition of the Derivative

The **derivative** of a function $f(x)$ at a point $x = a$ is defined as:

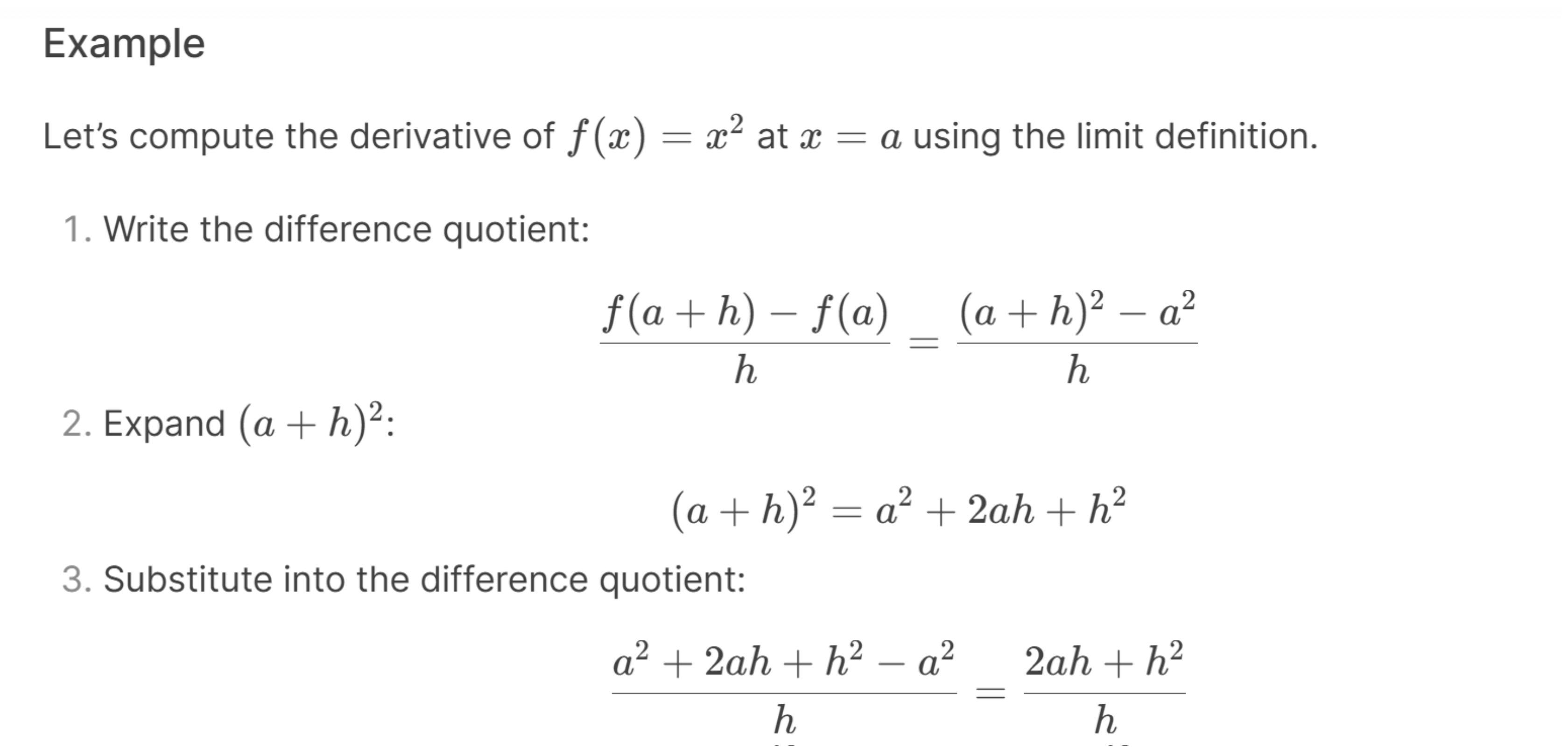
$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

Explanation of the Definition

1. $f(a+h) - f(a)$:
   - This represents the change in the function's value when the input changes from a to $a+h$.
   - It is often written as Δf, where Δ denotes "change in."
2. h:
   - This is the change in the input (Δx), which is approaching 0.
3. $\frac{f(a+h)-f(a)}{h}$:
   - This is the **average rate of change** of $f(x)$ over the interval $[a, a+h]$.
   - It is also called the **difference quotient**.
4. $\lim_{h\to 0}$:
   - The limit as h approaches 0 ensures that we are considering the **instantaneous rate of change** of $f(x)$ at the point $x = a$.

Geometric Interpretation

- The derivative $f'(a)$ represents the **slope of the tangent line** to the graph of $f(x)$ at the point $x = a$.
- As h gets smaller, the secant line (connecting $(a, f(a))$ and $(a + h, f(a + h))$) becomes closer to the tangent line at $x = a$.



Example

Let’s compute the derivative of $f(x) = x^2$ at $x = a$ using the limit definition.

1. Write the difference quotient:

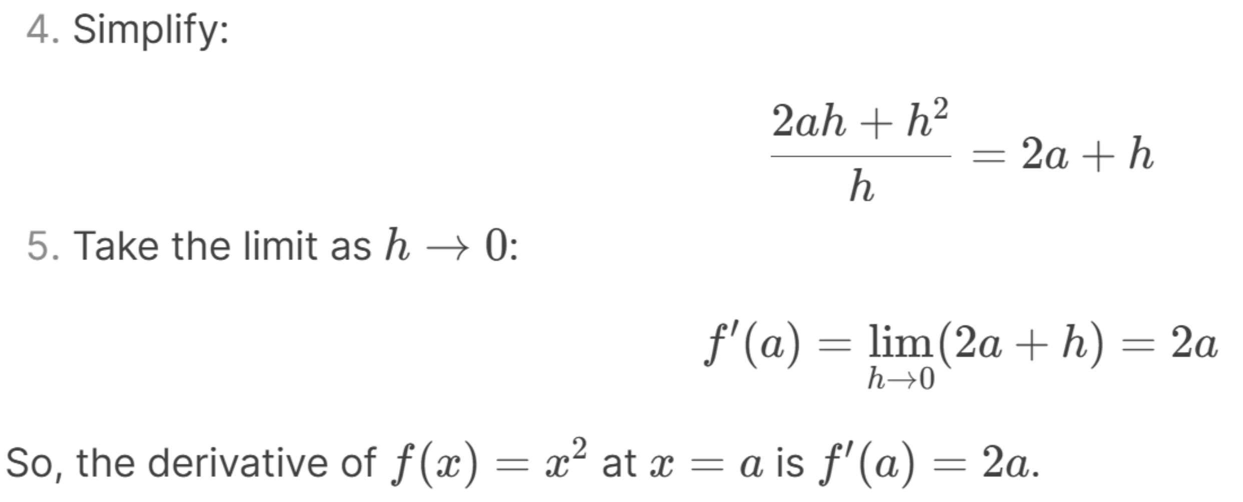
$$\frac{f(a+h) - f(a)}{h} = \frac{(a+h)^2 - a^2}{h}$$

2. Expand $(a+h)^2$:

$$(a+h)^2 = a^2 + 2ah + h^2$$

3. Substitute into the difference quotient:

$$\frac{a^2 + 2ah + h^2 - a^2}{h} = \frac{2ah + h^2}{h}$$



4. Simplify:

$$\frac{2ah + h^2}{h} = 2a + h$$

5. Take the limit as $h \to 0$:

$$f'(a) = \lim_{h \to 0} (2a + h) = 2a$$

So, the derivative of $f(x) = x^2$ at $x = a$ is $f'(a) = 2a$.

General Derivative

The derivative of $f(x)$ as a function (not just at a specific point) is:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

This gives the slope of the tangent line to $f(x)$ at any point x.

Summary

- The derivative $f'(a)$ is the **instantaneous rate of change** of $f(x)$ at $x = a$.
- It is defined as the limit of the **difference quotient** as $h \to 0$.
- Geometrically, it represents the **slope of the tangent line** to the curve at $x = a$.

DERIVATIVE AND GRADIENT-1

- A **derivative** f' of a function f is a function or a value that describes how fast f grows (or decreases).
  - If the derivative is a constant value, like 5 or -3, then the function grows (or decreases) constantly at any point x of its domain.
  - If the derivative f' is a function, then the function f can grow at a different pace in different regions of its domain.
  - The derivative of zero at x means that the function's slope at x is horizontal.
- The process of finding a derivative is called **differentiation.**

DERIVATIVE AND GRADIENT-2

- Derivatives for basic functions are known. For example,
  - $f(x) = x^2$, then $f'(x) = 2x$
  - $f(x) = 2x$, then$f'(x) = 2$
  - $f(x) = 2$, then $f'(x) = 0$
  - If the function we want to differentiate is not basic, we can find its derivative using the **chain rule.**
    - For instance if $F(x) = f(g(x))$, where f and g are some functions, then $F'(x) = f'(g(x))g'(x)$.
    - For example if $F(x) = (5x + 1)^2$ then $g(x) = 5x + 1$ and $f\left(g(x)\right) = \left(g(x)\right)^2$.
    - By applying the chain rule, we find $F'(x) = 2(5x + 1)g'(x) = 2(5x + 1)5 = 50x + 10$.

DERIVATIVE AND GRADIENT-3

- **Gradient** is the generalization of derivative for functions that take several inputs (or one input in the form of a vector or some other complex structure).

- A gradient of a function is a vector of **partial derivatives.**

- You can look at finding a partial derivative of a function as the process of finding the derivative by focusing on one of the function's inputs and by considering all other inputs as constant values.
  - *Let* $f([x^{(1)}, x^{(2)}]) \; = \; ax^{(1)} + \; bx^{(2)} + \; c$, then
  - $\frac{\partial f}{\partial x^{(1)}} = a + 0 + 0 = a, \frac{\partial f}{\partial x^{(2)}} = \, ?$

- The gradient of function f, denoted as ∇f is given by the vector $[\frac{\partial f}{\partial x^{(1)}}, \frac{\partial f}{\partial x^{(2)}}]$

RANDOM VARIABLE

- A **random variable**, usually written as an italic capital letter, like X, is a variable whose possible values are numerical outcomes of a random phenomenon.
  - Examples include a toss of a coin (0 for heads and 1 for tails), a roll of a dice, or the height of the first stranger you meet outside.
- There are two types of random variables:
  - **Discrete:** takes on only a countable number of distinct values such as red, yellow, blue or 1, 2, 3, …
  - **Continuous** (**CRV**): takes an infinite number of possible values in some interval like height, weight, and time.

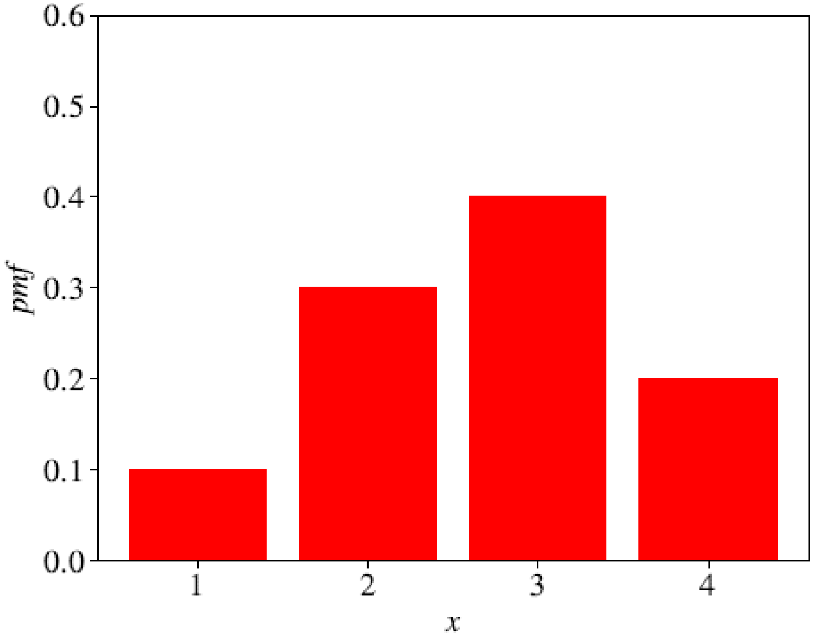
DISCRETE RANDOM VARIABLES

Let a discrete random variable X have k possible values $\{x_i\}_{i=1}^{k}$.

- The **probability distribution** of a discrete random variable is described by a list of probabilities associated with each of its possible values.

- This list of probabilities is called a **probability mass function** (pmf).
  - For example: $\Pr(X=1)=0.1, \Pr(X=2)=0.3, \Pr(X=3)=0.4, \Pr(X=4)=0.2.$
  - $\Pr(X=x_i)\in[0,1].$
  - $\sum_{i=1}^{k}\Pr(X=x_i)=1.$

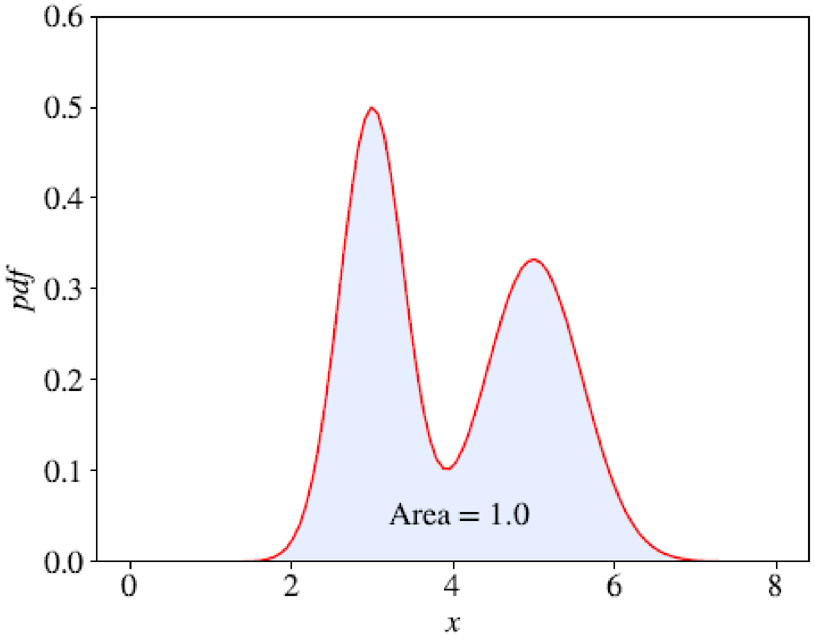
- The **expectation** of X denoted as $E[X]=\sum_{i=1}^{k}(x_i\cdot\Pr(X=x_i))$
  - **AKA mean, average** or **expected value,** frequently denoted by μ.



CONTINUOUS RANDOM VARIABLE

- Because the number of values of a continuous random variable X is infinite, the probability $\Pr(X = c)$ for any value c is 0.
- The probability distribution of a CRV (a continuous probability distribution) is described by a **probability density function** (pdf).
- The pdf is a function whose codomain is nonnegative and the area under the curve is equal to 1.
- $E[x] = \int_{\mathbb{R}} x \cdot fx(x) \quad d(x)$, where fx is the pdf of the variable X and $\int_{\mathbb{R}}$ is the *integral* of function xfx.

  In simpler terms, the expected value *E*[*X*] is calculated by multiplying each possible value *x* by its probability density *fX*(*x*) and then summing (integrating) over all possible values.
- $\int_{\mathbb{R}} fx(x) \quad d(x) = 1$: the area under the pdf curve is 1.



PARAMETERS VS. HYPERPARAMETERS

- A **hyperparameter** is a property of a learning algorithm, usually (but not always) having a numerical value.
- That value influences the way the algorithm works.
- Hyperparameters aren’t learned by the algorithm itself from data.

- **Parameters** are variables that define the model learned by the learning algorithm.
- Parameters are directly modified by the learning algorithm based on the training data.
- The goal of learning is to find such values of parameters that make the model optimal in a certain sense.

MODEL-BASED VS. INSTANCE-BASED LEARNING

- Model-based learning algorithms use the training data to create a **model** that has **parameters** learned from the training data.
- Most learning algorithms are model-based e.g. linear regression, SVM, ANN, DTs.
- After the model was built, the training data can be discarded.

Instance-based learning algorithms use the whole dataset as the model.

One instance-based algorithm frequently used in practice is **k-Nearest Neighbors** (kNN).

In classification, to predict a label for an input example the kNN algorithm looks at the close neighborhood of the input example in the space of feature vectors and outputs the label that it saw the most often in this close neighborhood.