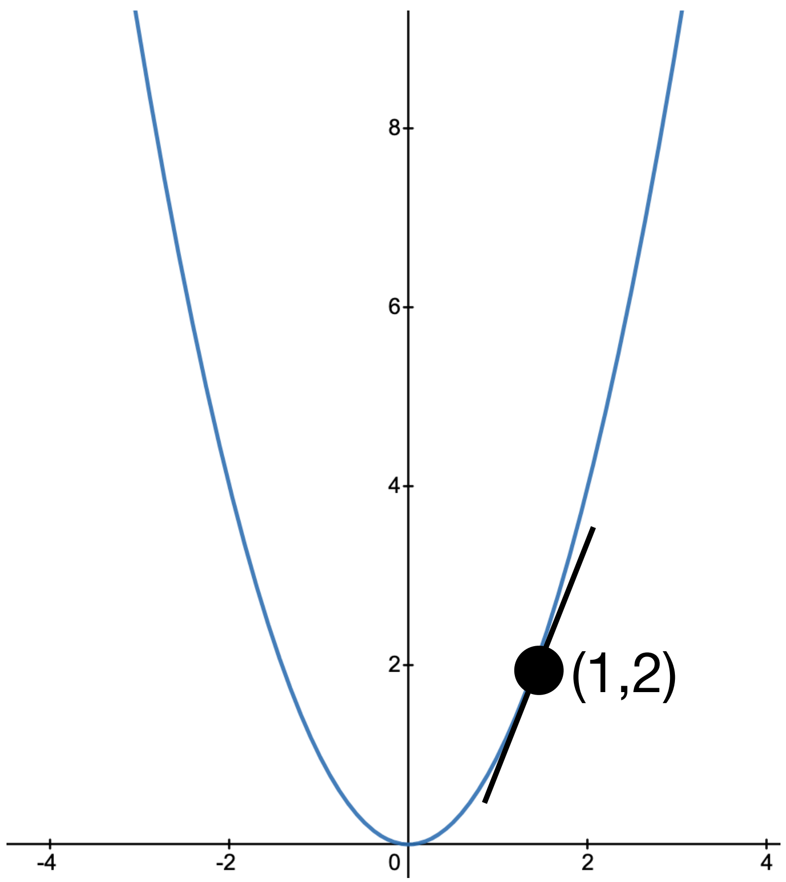
Gradient Descent

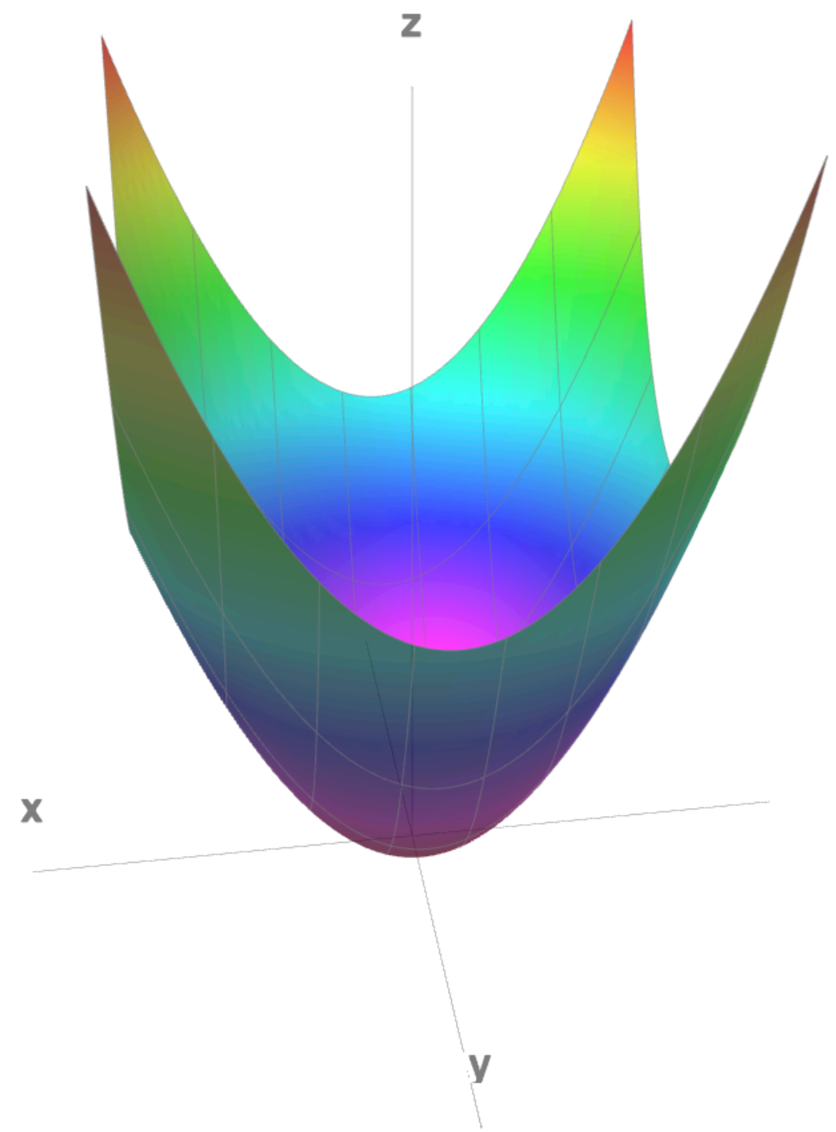
Erisa Terolli
CS 556

Functions of two variables



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

Tangent Line



$$f(x, y) = x^2 + y^2$$

Tangent Plane

Partial Derivatives

The partial derivative of a function of several variables is its derivative with respect to one of those variables, with the others held constant. The partial derivative of a function $f(x, y)$ with respect to variable x is denoted as $\dfrac{\partial f}{\partial x}$.

Example

To find the partial derivative of $f(x, y) = x^2 + y^2$ with respect to x :

1. Treat all other variables as constant (y in this case).
2. Differentiate the function using the normal rules of differentiation

$$Fix\ y = 2,\ f(x,2) = x^2 + 2^2$$

$$\frac{\partial f}{\partial x} = 2x$$

Gradient Definition

The gradient is the vector of partial derivatives of a function with respect to its variables. The gradient is denoted by the symbol ∇ . The gradient of a function $f(x, y, z)$ is represented by ∇f and is defined as:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix}$$

Gradient Example

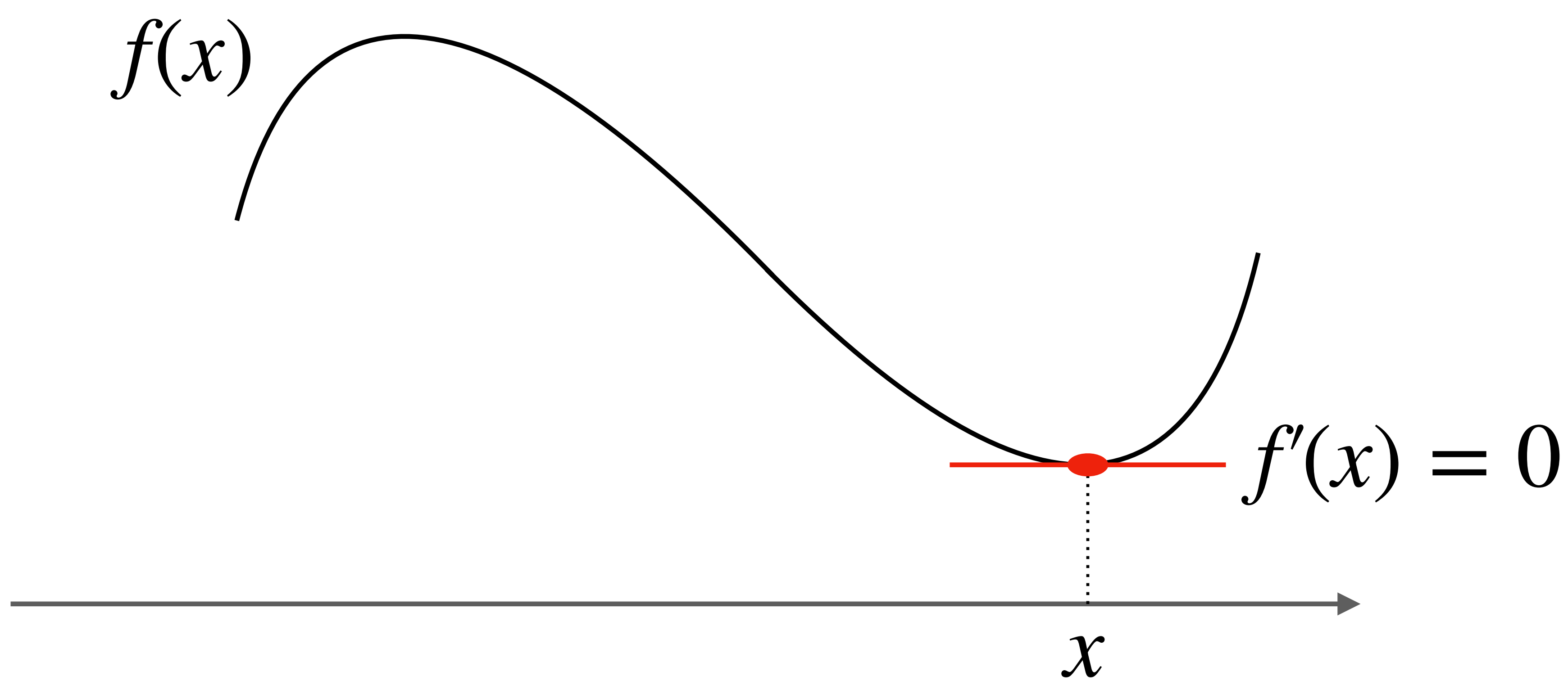
Compute the gradient of $f(x, y) = x^2 + y^2$ at point $(1,2)$.

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\nabla f_{(1,2)} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

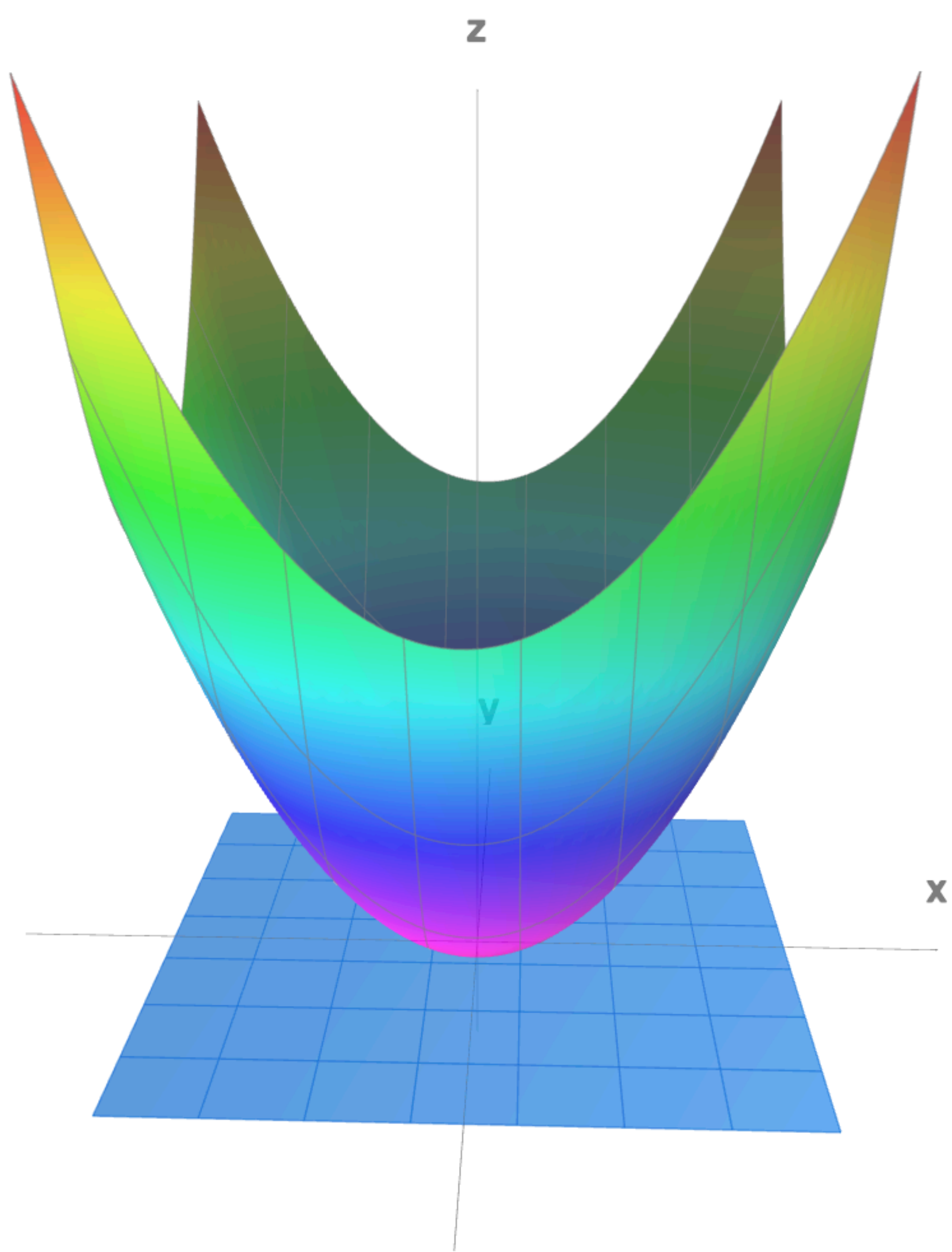
Optimizing functions

Find the value of x that minimizes the $f(x)$.



Optimizing functions

Find the value of x and y that minimizes the $f(x, y)$.



Minimum found at the point where both slopes are 0.

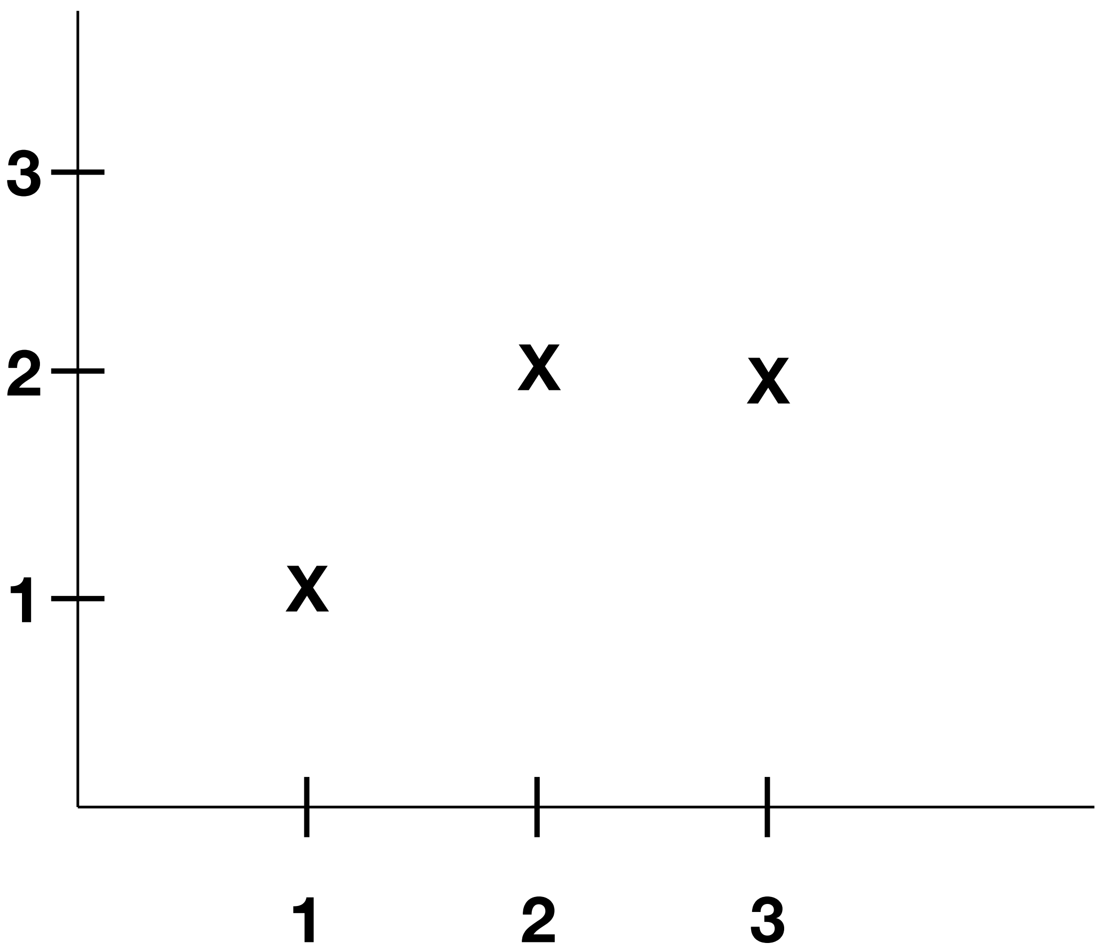
$$\frac{\partial f}{\partial x} = 2x = 0 \rightarrow x = 0$$

$$\frac{\partial f}{\partial y} = 2y = 0 \rightarrow y = 0$$

Minimum found at $(0,0)$.

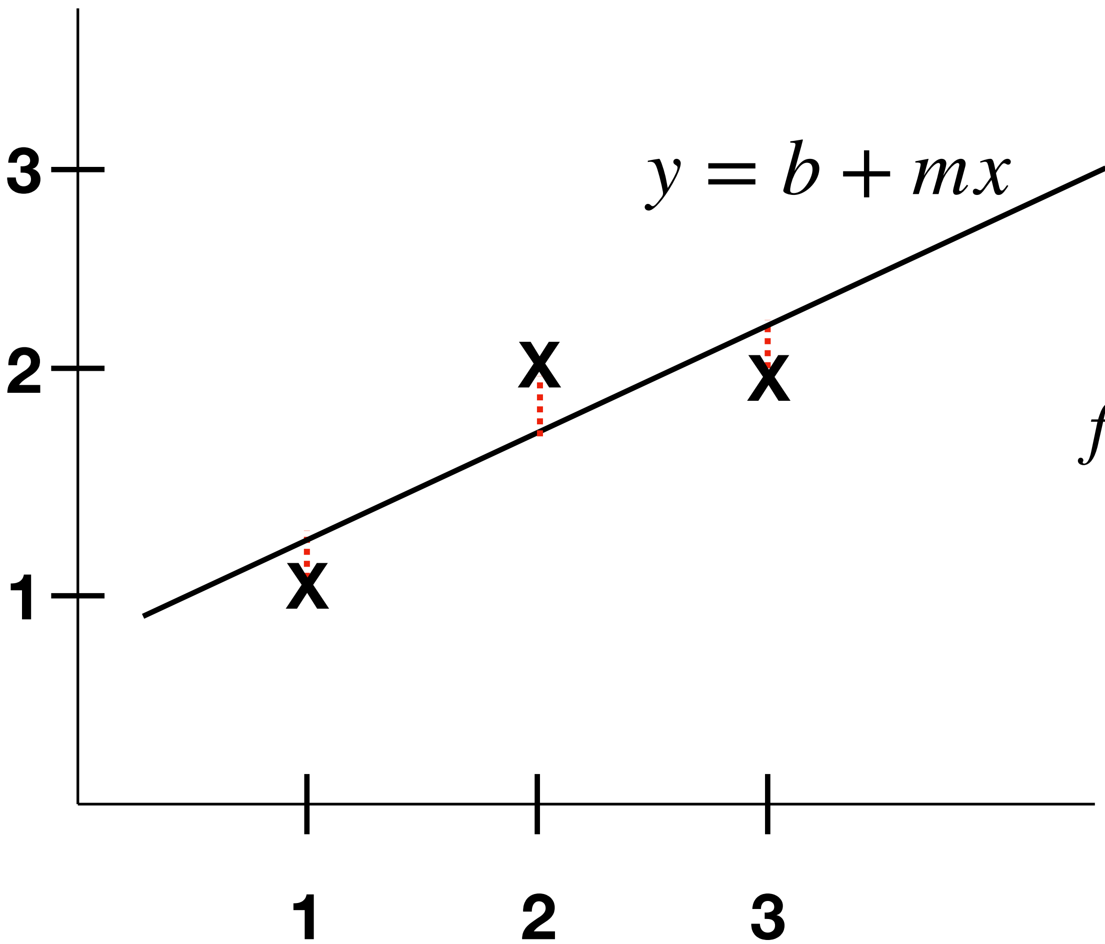
Optimizing with gradients

Find the closest line to the points (1,1), (2,2) and (3,2).



Optimizing with gradients

Find the closest line to the points (1,1), (2,2) and (3,2).



Find the optimal m and b to minimize the sum of squared distances between the points and the line.

$$f(m, b) = (m + b - 1)^2 + (2m + b - 2)^2 + (3m + b - 2)^2$$

$$\frac{\partial f}{\partial m} = 28m + 12b - 22$$

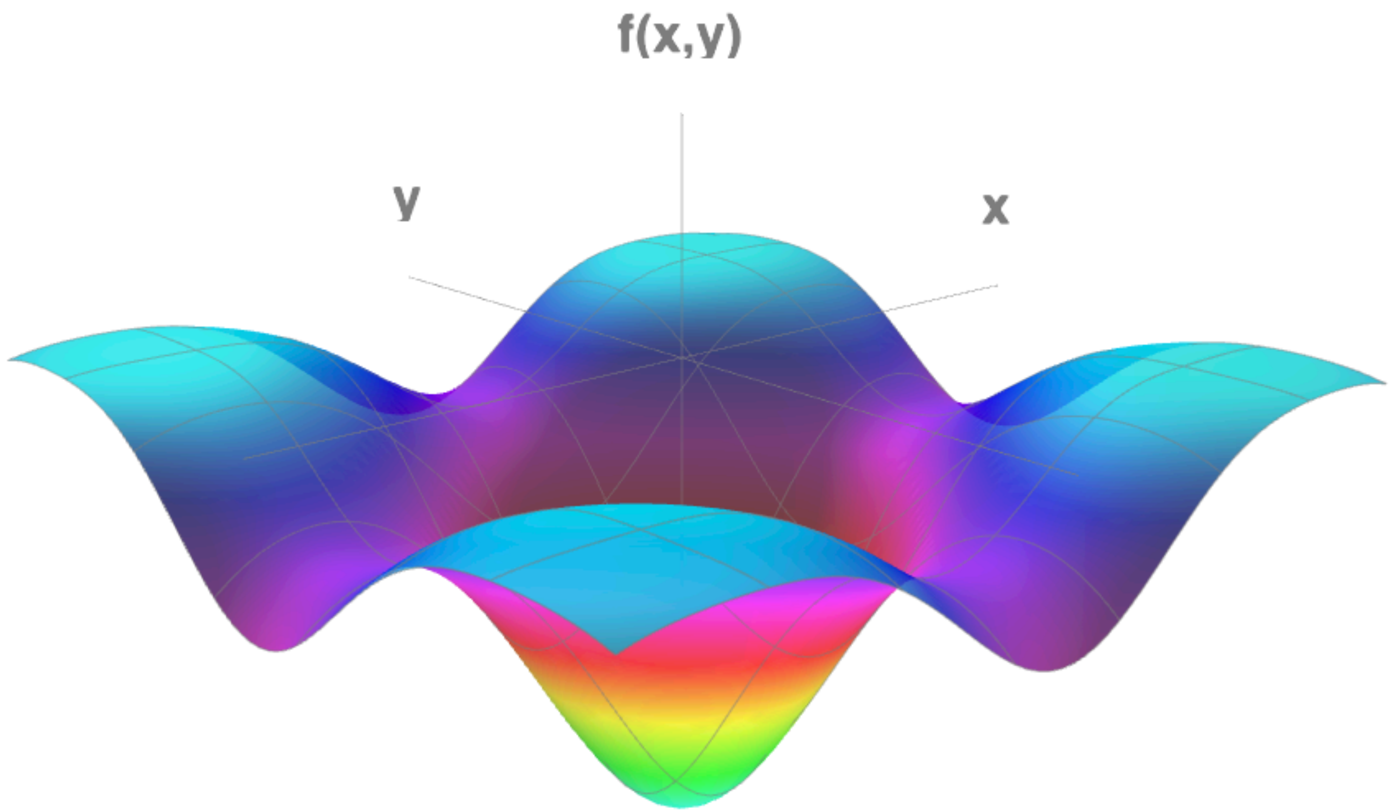
$$\frac{\partial f}{\partial b} = 12m + 6b - 10$$

$$\begin{cases} 28m + 12b = 22 \\ 12m + \ \ 6b = 10 \end{cases} \rightarrow \quad \begin{matrix} m = \dfrac{1}{2} \\ b = \dfrac{2}{3} \end{matrix}$$

Gradient Descent Intuition

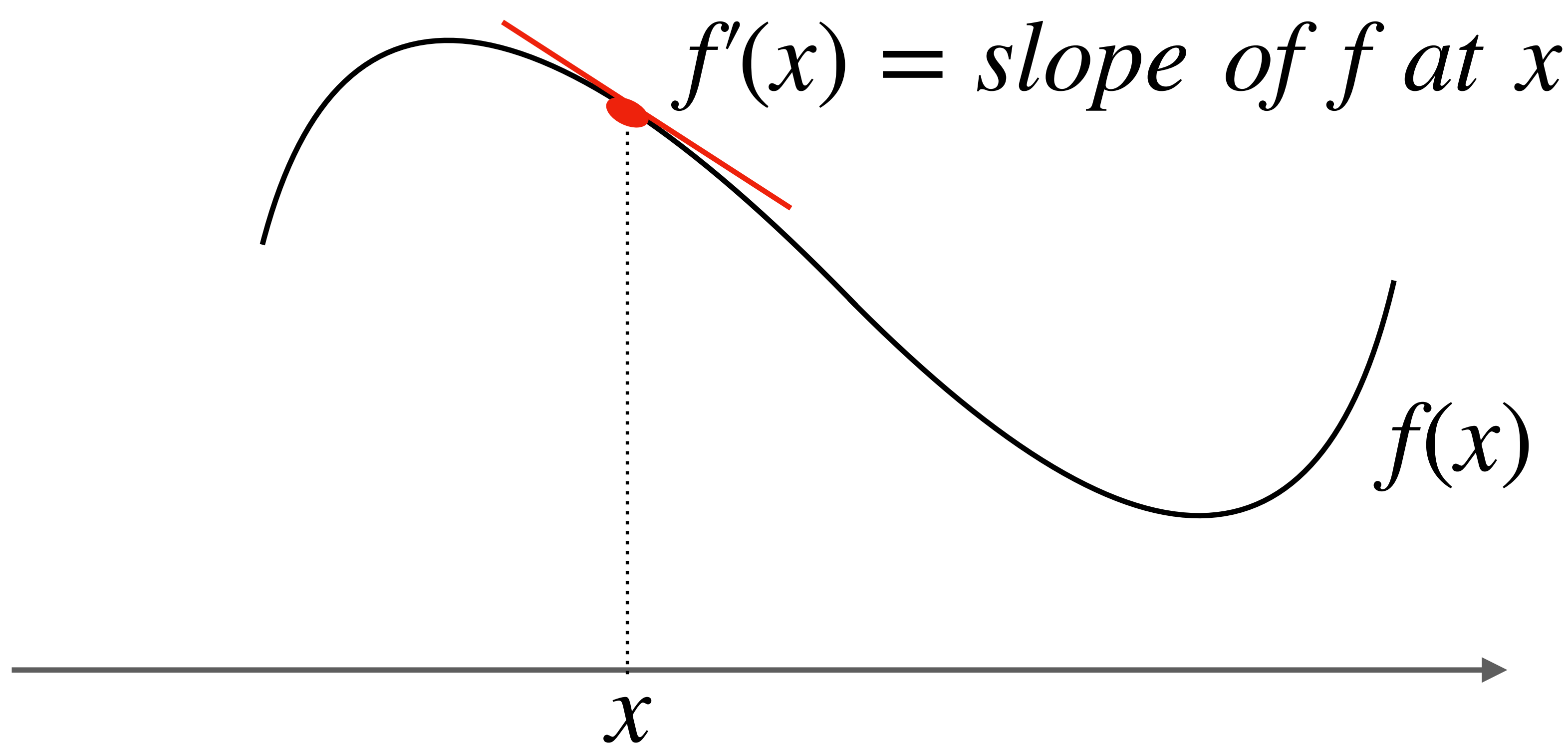
Task

Find the parameters x and y that minimize the function $f(x, y)$.



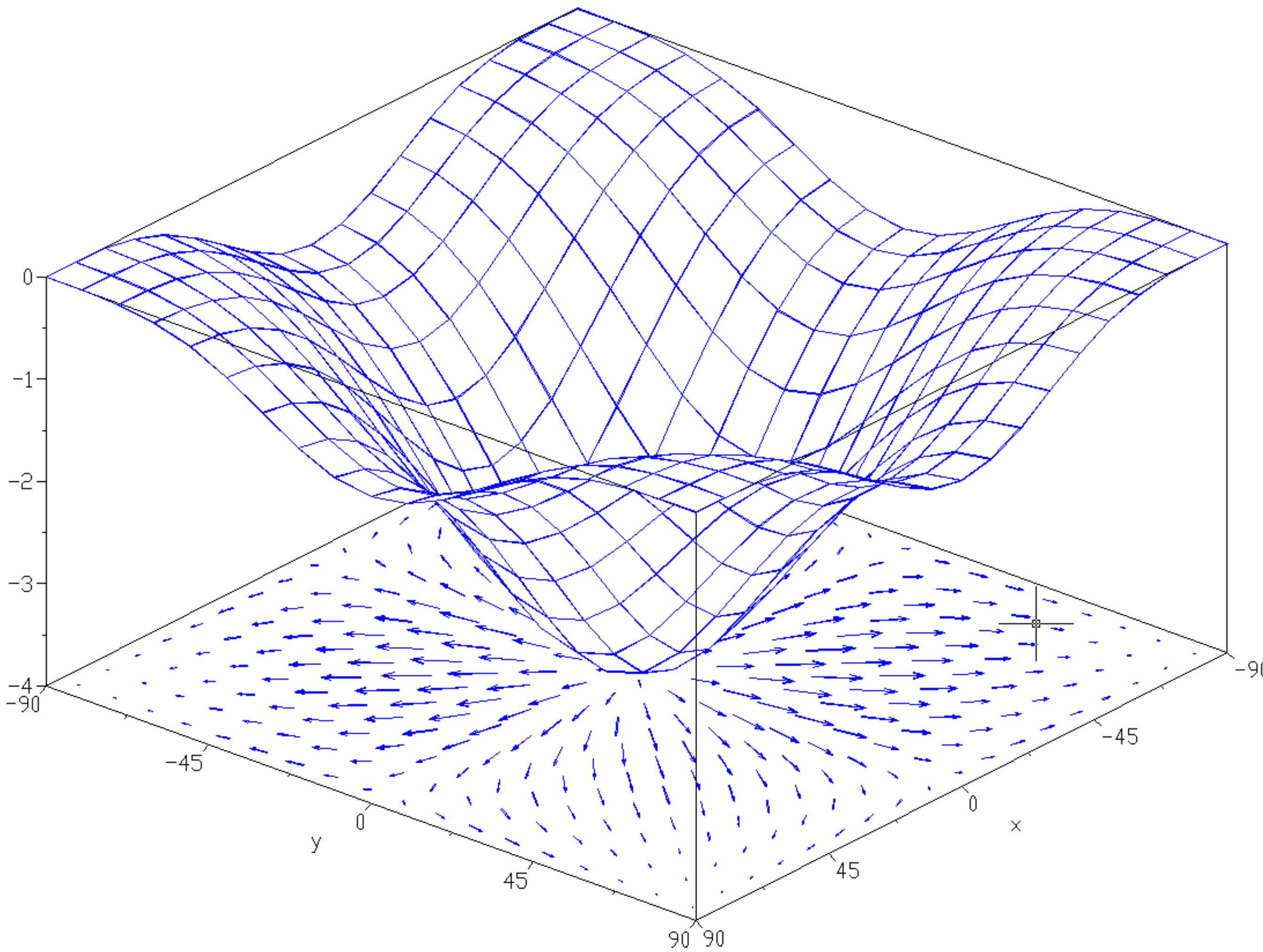
Gradient of a function in 1D

The derivative $f'(x)$ of $f(x)$ tells the direction and intensity of the increase of $f(x)$ at point x.



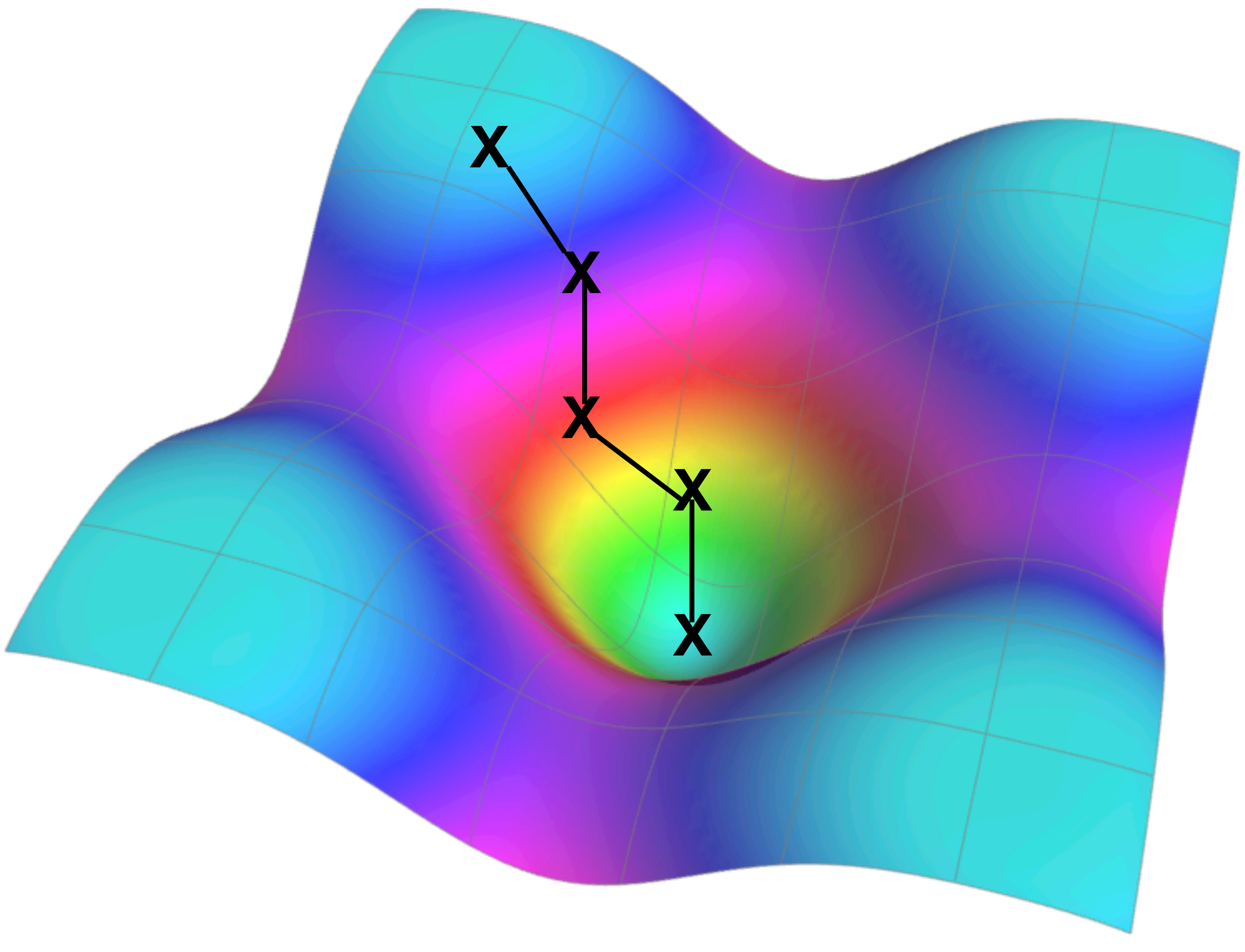
Gradient of a function

The gradient $\nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \ldots\right]$ of $f(x, y, \ldots)$ tells the direction and intensity of the maximum increase of $f()$ at $(x, y, \ldots)$.



Gradient Descent

Find the minimum of $f()$ by repeatedly following $-\nabla f$.



Gradient Descent Pseudocode

Function: $f(x, y)$, Goal: find minimum of $f(x, y)$.

1. Pick and arbitrary starting point (x_0, y_0)

2. Repeat until convergence:

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} - \alpha \nabla f(x_{k-1}, y_{k-1})$$

α is the step size, or step length, or learning rate

The Learning Rate

- If the learning rate α is too small, gradient descent can be slow.

- If α is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.

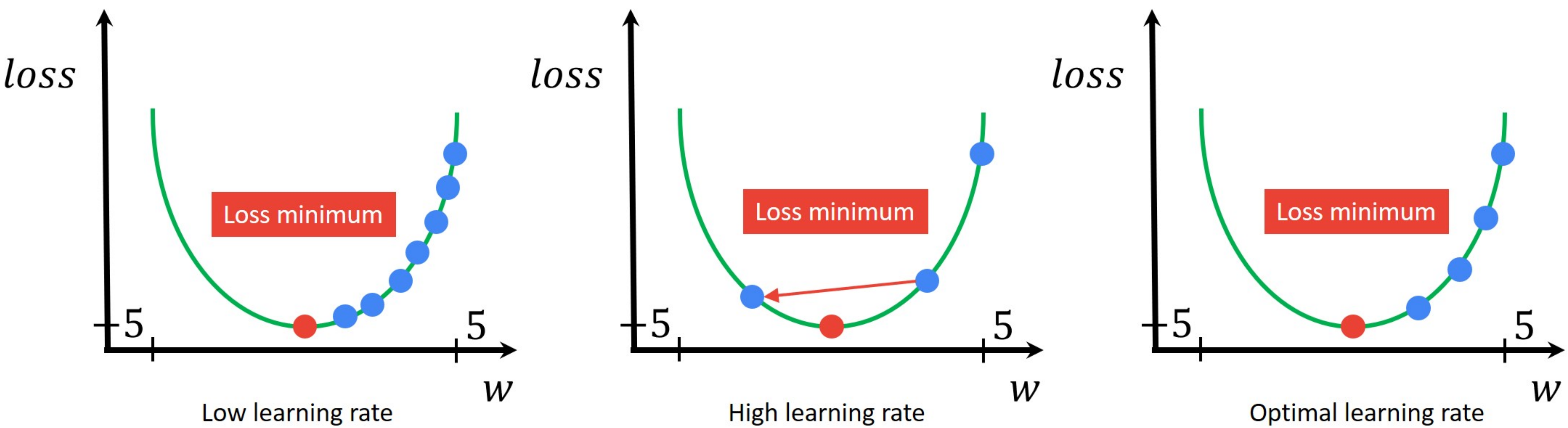


Image Credit: datahacker.rs

Gradient Descent Pros and Cons

Pros

- Can be applied to every dimensions and space
- Easy to implement

Cons

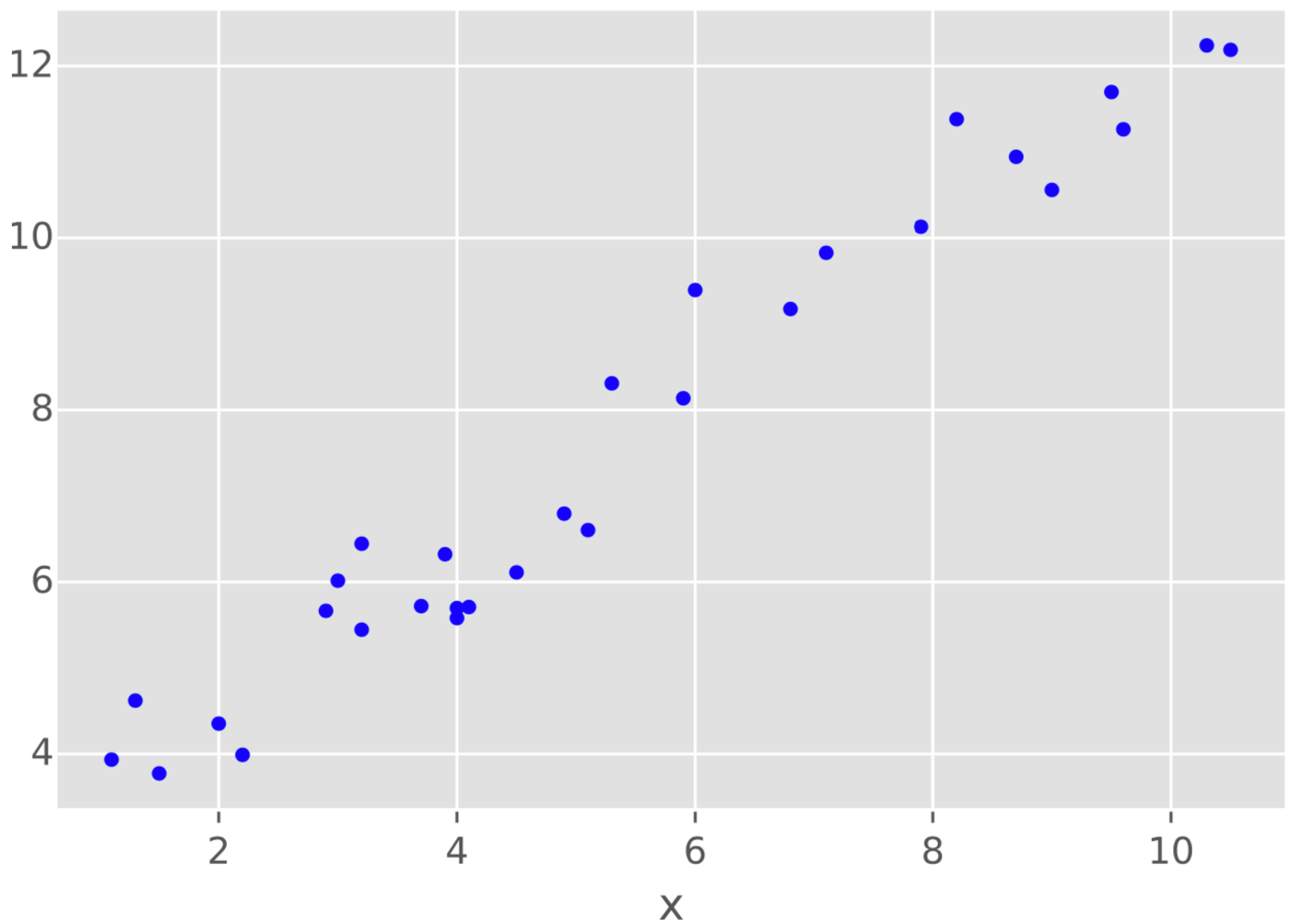
- Local minima problem
- Relatively slow close to minimum
- For non-differentiable functions, gradient methods are ill-defined

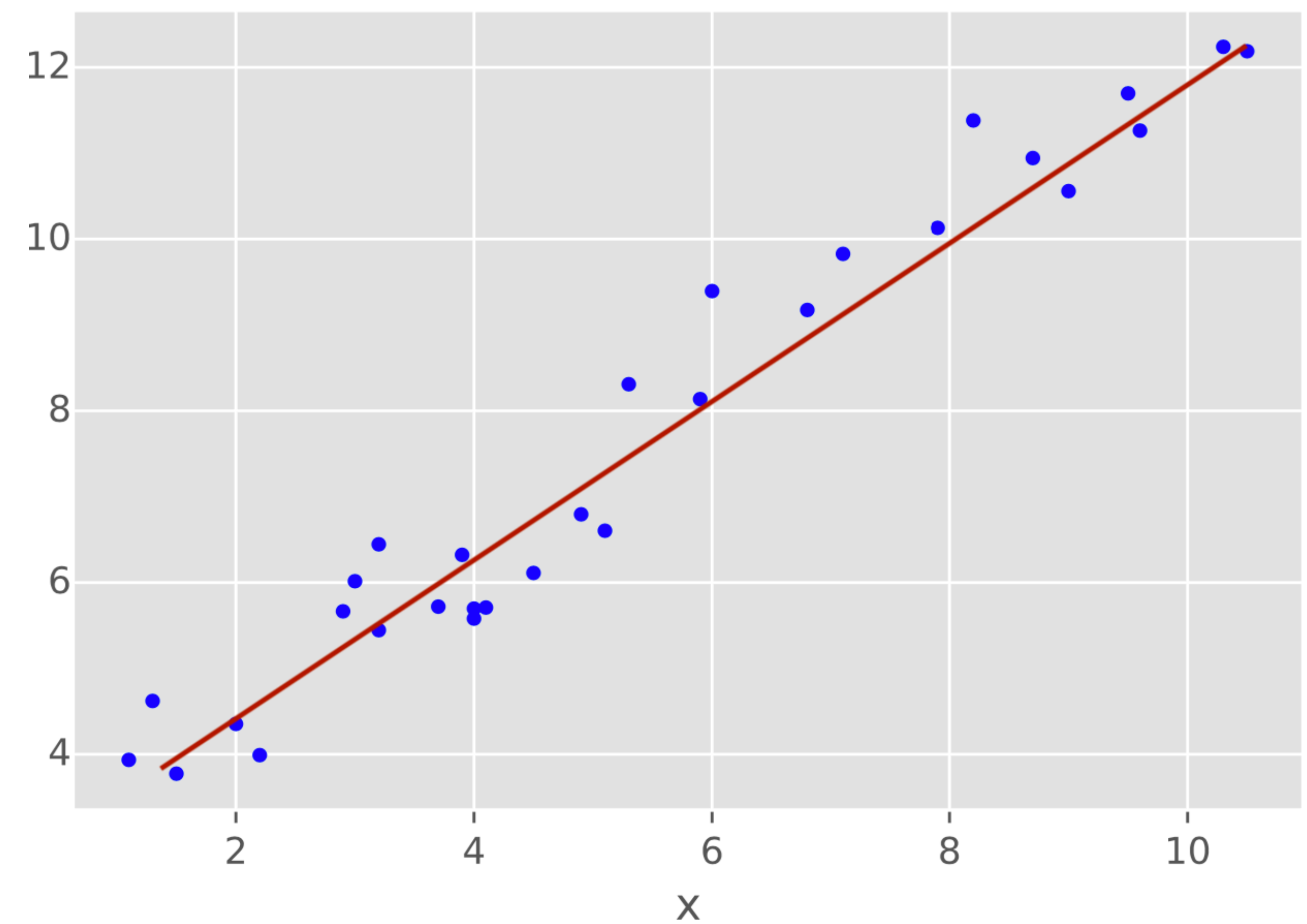
Thank you!

Application to Linear Regression

Find the hypothesis function which minimizes the loss:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2, \quad h_\theta((x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$$





Gradient Descent for Linear Regression

1. Start with some arbitrary θ.

2. Repeat until convergence:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta),\ (j = 1,2,\ldots,n)$$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)},\ (j = 1,2,\ldots,n)$$