Optimization

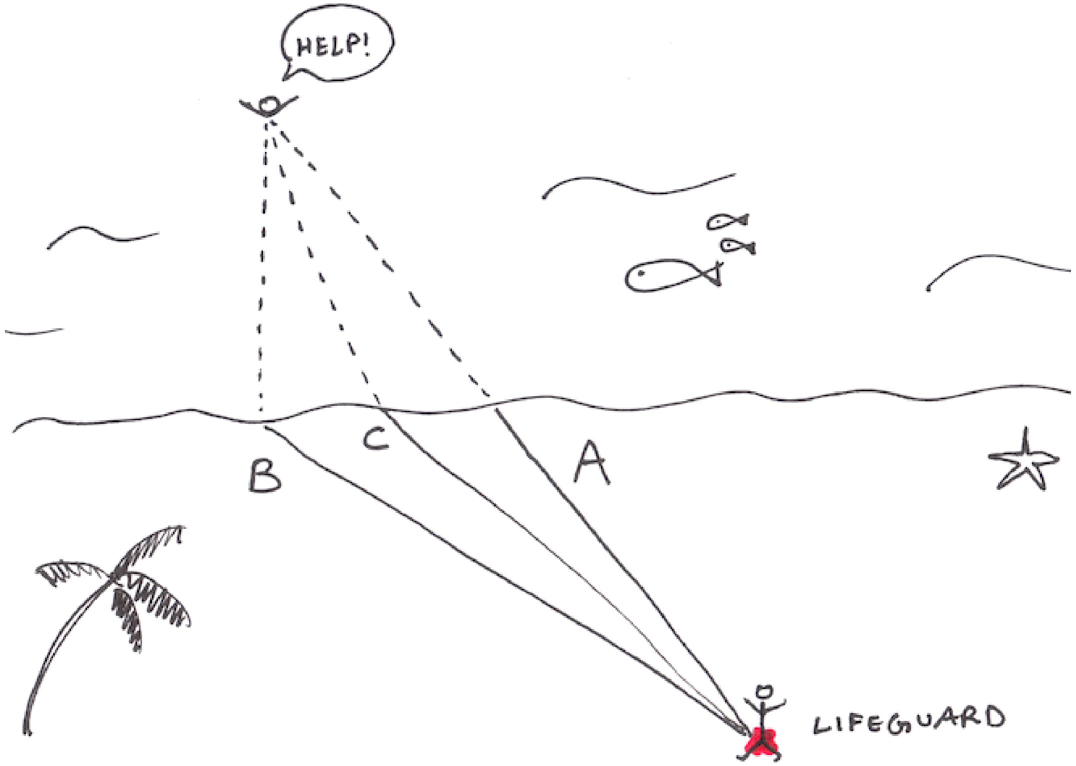
Prof. Seungchul Lee
Industrial AI Lab.

Optimization

- An important tool in
  1) Engineering problem solving and
  2) Decision science

Optimization

- Optimization



(source: http://nautil.us/blog/to-save-drowning-people-ask-yourself-what-would-light-do)

Optimization

- 3 key components
  1) Objective function
  2) Decision variable or unknown
  3) Constraints

- Procedures
  1) The process of identifying objective, variables, and constraints for a given problem (known as "modeling")
  2) Once the model has been formulated, optimization algorithm can be used to find its solutions

Optimization: Mathematical Model

- In mathematical expression

$$
\begin{array}{ll}
\min\limits_{x} & f(x) \\
\text{subject to} & g_i(x) \le 0, \qquad i = 1, \cdots, m
\end{array}
$$

  - $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$ is the decision variable
  - $f: \mathbb{R}^n \to \mathbb{R}$ is objective function
  - Feasible region: $C = \{x: g_i(x) \le 0, \ \ i = 1, \cdots, m\}$
  - $x^* \in \mathbb{R}^n$ is an optimal solution if $x^* \in C$ and $f(x^*) \le f(x), \forall x \in C$

Optimization: Mathematical Model

- In mathematical expression

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0, \qquad\qquad i = 1, \cdots, m \end{aligned}$$

- Remarks: equivalent

$$\begin{aligned} \min_{x} f(x) \quad & \leftrightarrow \quad \max_{x} -f(x) \\ g_i(x) \le 0 \quad & \leftrightarrow \quad -g_i(x) \ge 0 \\ h(x) = 0 \quad & \leftrightarrow \quad \begin{cases} h(x) \le 0 & \text{and} \\ h(x) \ge 0 \end{cases} \end{aligned}$$

Unconstrained vs. Constrained

$$\underset{x}{\text{minimize}} \quad f(x)$$

vs.

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

Convex Optimization

Convex Optimization

- An extremely powerful subset of all optimization problems

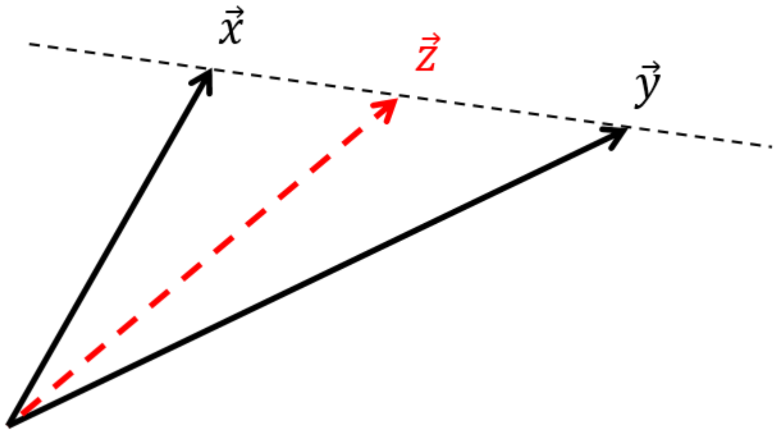
$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & x \in \mathcal{C} \end{aligned}$$

- $f: \mathbb{R}^n \to \mathbb{R}$ is a convex function and
- Feasible region C is a convex set

- Key property of convex optimization:
  - all local solutions are global solutions

Linear Interpolation between Two Points

- $\vec{z} = \theta\vec{x} + (1-\theta)\vec{y}$ and $\theta \in [0,1]$

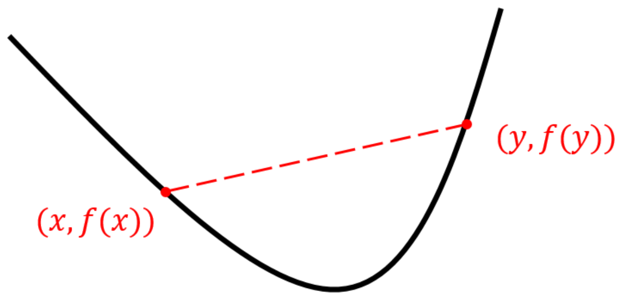


$$\vec{z} = \vec{y} + \theta(\vec{x} - \vec{y}) = \theta\vec{x} + (1-\theta)\vec{y}, \qquad 0 \leq \theta \leq 1$$

$$\text{or} \quad \vec{z} = \alpha\vec{x} + \beta\vec{y}, \qquad \alpha + \beta = 1 \text{ and } 0 \leq \alpha, \beta$$

Convex Function and Convex Set

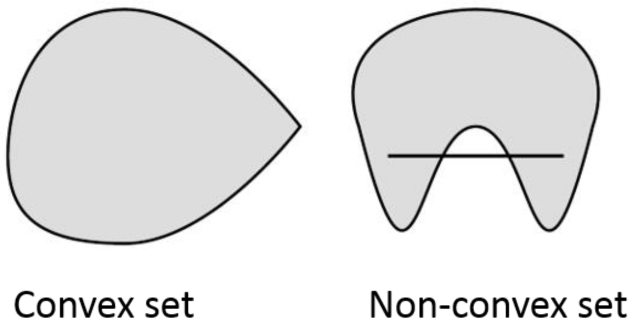
convex function



for any $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

convex set



for a $x, y \in \mathcal{C}$ and $\theta \in [0, 1]$,

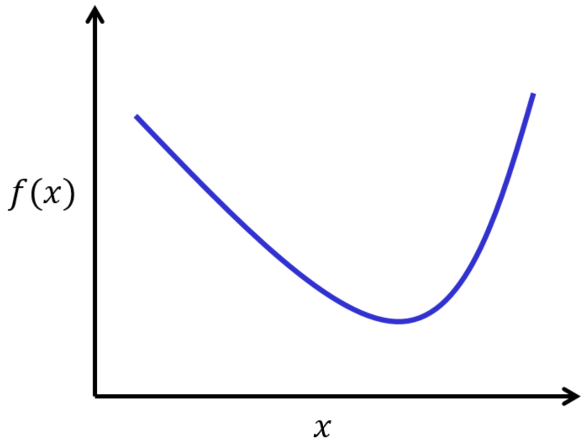
$$\theta x + (1-\theta)y \in \mathcal{C}$$

Images from http://web.stanford.edu/class/ee364a/lectures.html

Solving Optimization Problems

Solving Optimization Problems

- Starting with the unconstrained, one dimensional case



  - To find minimum point x^*, we can look at the derivative of the function $f'(x)$
  - Any location where $f'(x) = 0$ will be a "flat" point in the function

- For convex problems, this is guaranteed to be a global minimum

Solving Optimization Problems

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- Generalization for multivariate function $f: \mathbb{R}^n \rightarrow \mathbb{R}$
  - the gradient of f must be zero

$$\nabla_x f(x) = 0$$

- For defined as above, *gradient* is a *n*-dimensional vector containing partial derivatives with respect to each dimension

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

- For continuously differentiable f and unconstrained optimization, optimal point must have

$$\nabla_x f(x^*) = 0$$

How do we Find $\nabla_x f(x) = 0$

- Direct solution
  - In some cases, it is possible to analytically compute x^* such that $\nabla_x f(x^*) = 0$

$$f(x) = 2x_1^2 + x_2^2 + x_1 x_2 - 6x_1 - 5x_2$$

$$\Longrightarrow \nabla_x f(x) = \begin{bmatrix} 4x_1 + x_2 - 6 \\ 2x_2 + x_1 - 5 \end{bmatrix}$$

$$\Longrightarrow x^\star = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$q(x_1, \cdots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} x_i x_j = x^T H x$$

Gradients

- Matrix derivatives

y	$\frac{\partial y}{\partial x}$
Ax	A^T
$x^T A$	A
$x^T x$	$2x$
$x^T Ax$	$Ax + A^T x$

How to Find $\nabla_x f(x) = 0$

y	$\frac{\partial y}{\partial x}$
Ax	A^T
$x^T A$	A
$x^T x$	$2x$
$x^T A x$	$Ax + A^T x$

- Direct solution
  - In some cases, it is possible to analytically compute x^* such that $\nabla_x f(x^*) = 0$

$$f(x) = 2x_1^2 + x_2^2 + x_1 x_2 - 6x_1 - 5x_2$$

$$\Longrightarrow \nabla_x f(x) = \begin{bmatrix} 4x_1 + x_2 - 6 \\ 2x_2 + x_1 - 5 \end{bmatrix}$$

$$\Longrightarrow x^\star = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Examples

y	$\frac{\partial y}{\partial x}$
Ax	A^T
$x^T A$	A
$x^T x$	$2x$
$x^T A x$	$Ax + A^T x$

- affine function $g(x) = a^T x + b$

$$\nabla g(x) = a, \qquad \nabla^2 g(x) = 0$$

- quadratic function $g(x) = x^T P x + q^T x + r$, $\qquad P = P^T$

$$\nabla g(x) = 2Px + q, \qquad \nabla^2 g(x) = 2P$$

- $g(x) = \|Ax - b\|^2 = x^T A^T A x - 2b^T A x + b^T b$

$$\nabla g(x) = 2A^T A x - 2A^T b, \qquad \nabla^2 g(x) = 2A^T A$$

Revisit: Least-Square Solution

- Scalar Objective: $J = \|Ax - y\|^2$

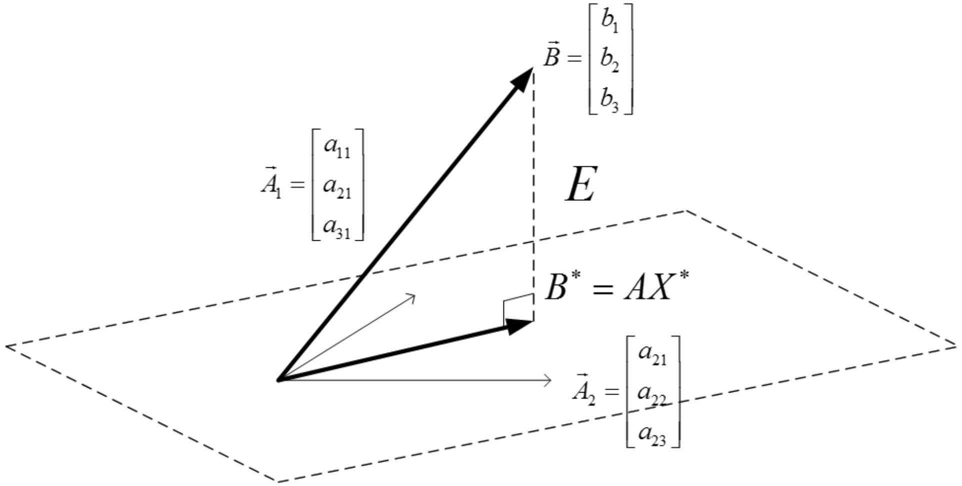
y	$\frac{\partial y}{\partial x}$
Ax	A^T
$x^T A$	A
$x^T x$	$2x$
$x^T A x$	$Ax + A^T x$

$$
\begin{aligned}
J(x) &= (Ax - y)^T (Ax - y) \\
&= \left(x^T A^T - y^T\right)(Ax - y) \\
&= x^T A^T A x - x^T A^T y - y^T A x + y^T y
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial J}{\partial x} &= A^T A x + \left(A^T A\right)^T x - A^T y - \left(y^T A\right)^T \\
&= 2A^T A x - 2A^T y = 0
\end{aligned}
$$

$$\Longrightarrow \quad \left(A^T A\right) x = A^T y$$

$$\therefore \quad x^* = \left(A^T A\right)^{-1} A^T y$$



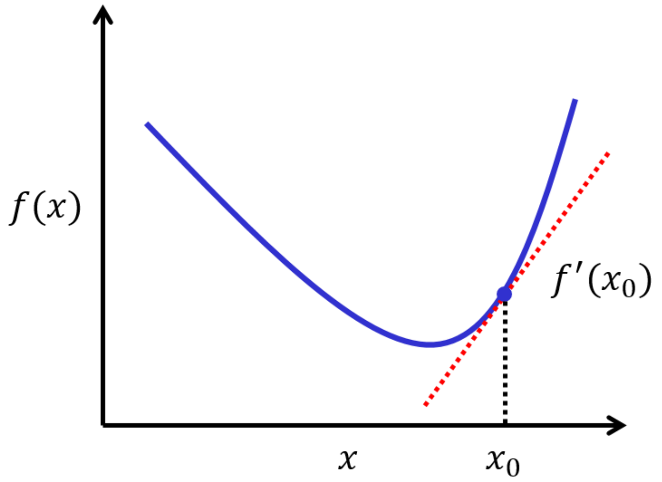
$$\min_X \|E\|^2 = \min_X \|AX - B\|^2$$

$$X^* = \left(A^T A\right)^{-1} A^T B$$

$$B^* = AX^* = A\left(A^T A\right)^{-1} A^T B$$

How do we Find $\nabla_x f(x) = 0$

- Iterative methods
  - More commonly the condition that the gradient equal zero will not have an analytical solution, require iterative methods

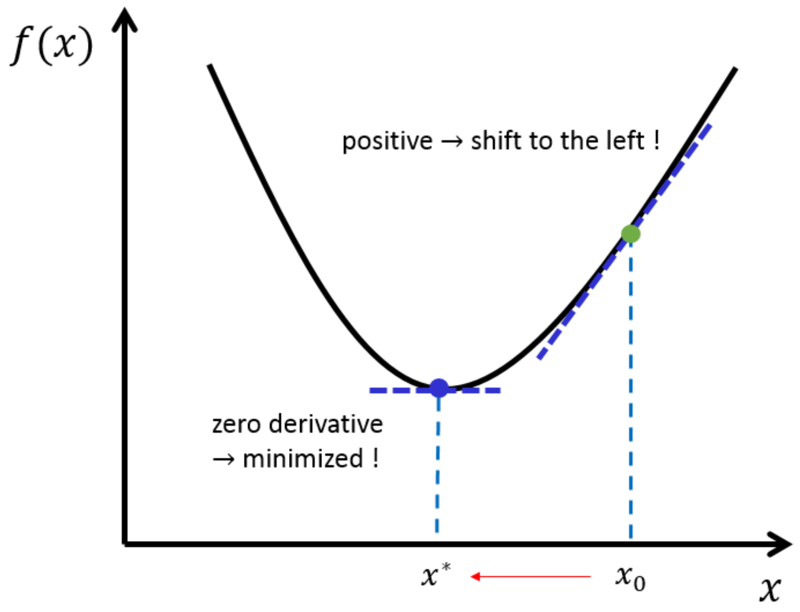


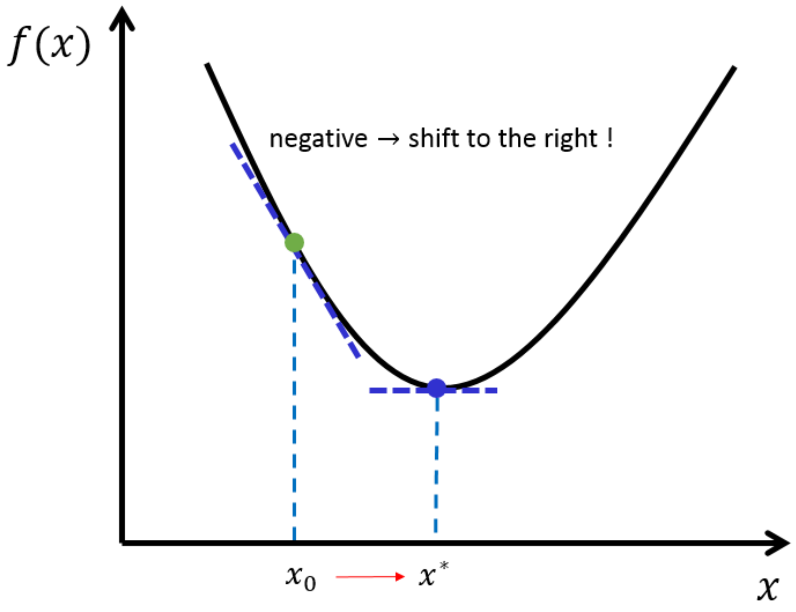
  - The gradient points in the direction of "steepest ascent" for function f

Descent Direction (1D)

- It motivates the *gradient descent* algorithm, which repeatedly takes steps in the direction of the negative gradient

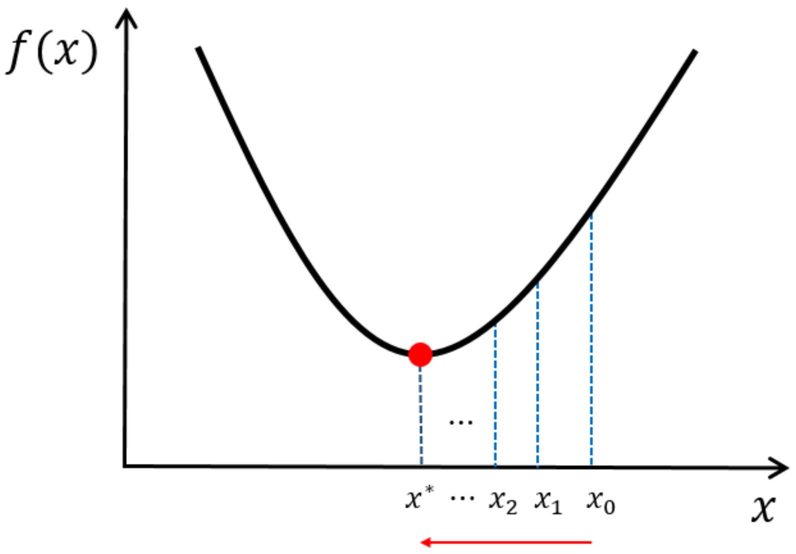
$$x \leftarrow x - \alpha \nabla_x f(x) \qquad \text{for some } \textit{step size } \alpha > 0$$





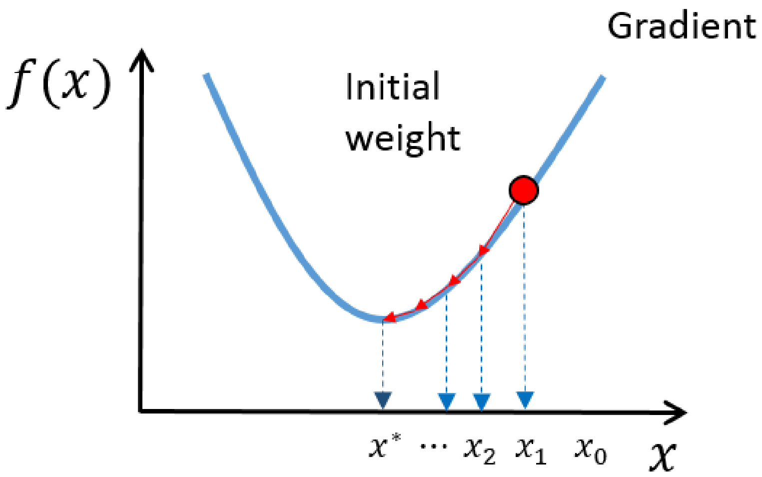
Gradient Descent

Repeat: $x \leftarrow x - \alpha \nabla_x f(x)$ for some *step size* $\alpha > 0$

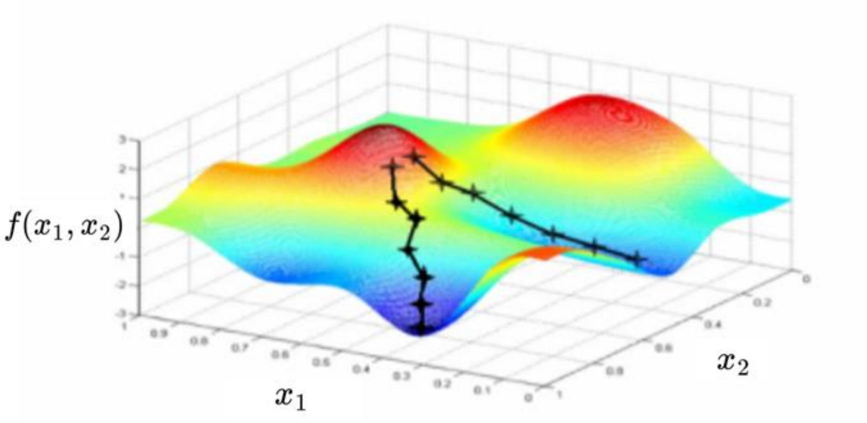


Gradient Descent in High Dimension

Repeat: $x \leftarrow x - \alpha \nabla_x f(x)$ for some *step size* $\alpha > 0$

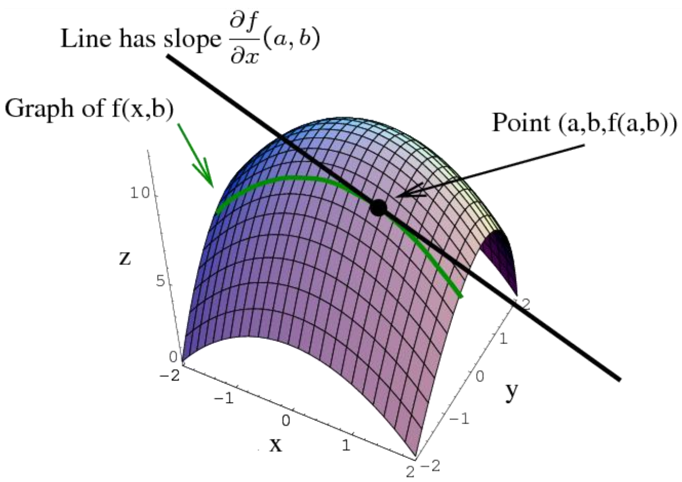


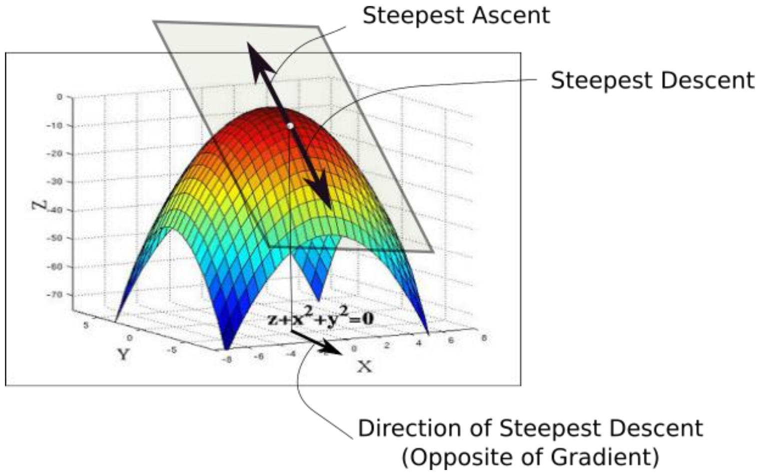
Global cost minimum $J_{\min}(\omega)$

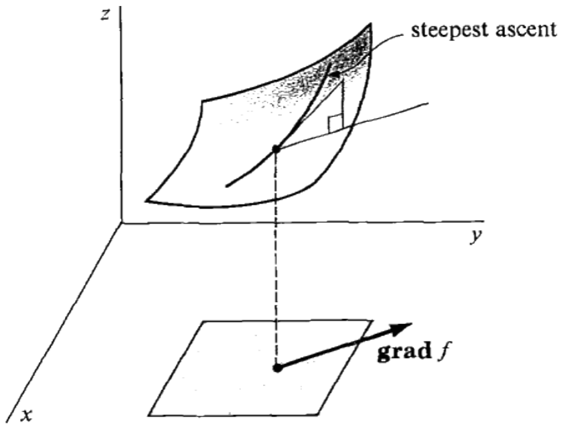


Gradient Descent in High Dimension

Repeat: $x \leftarrow x - \alpha \nabla_x f(x)$ for some *step size* $\alpha > 0$

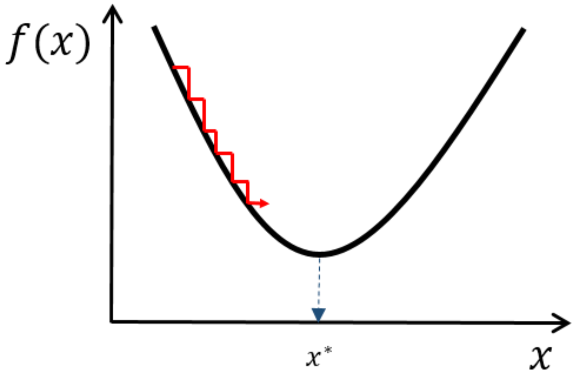




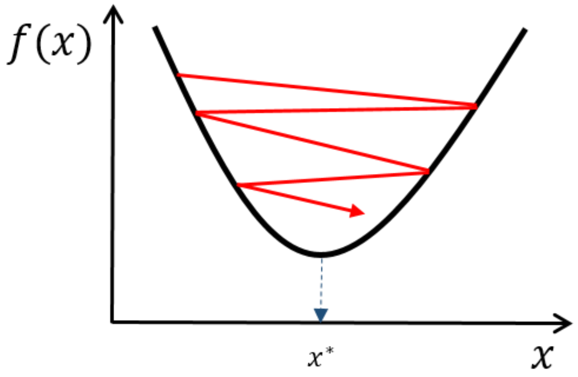


Choosing Step Size α

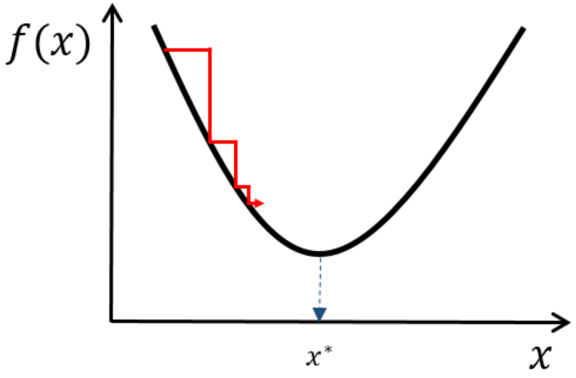
- Learning rate



Too small: converge very slowly

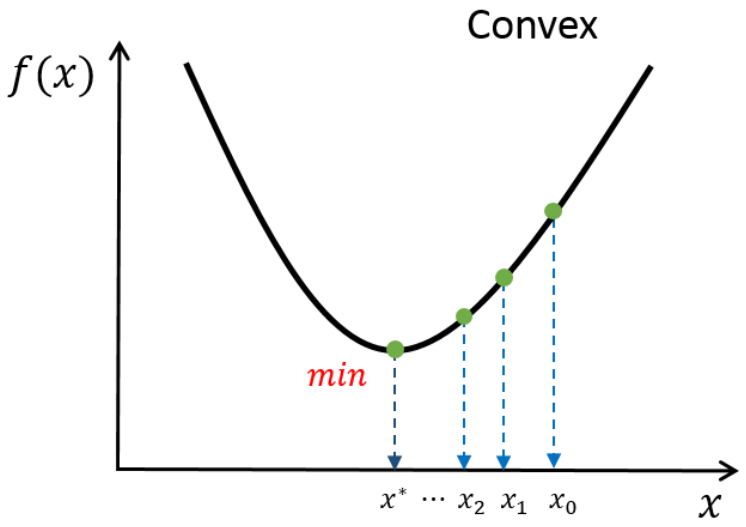


Too big: overshoot and even diverge

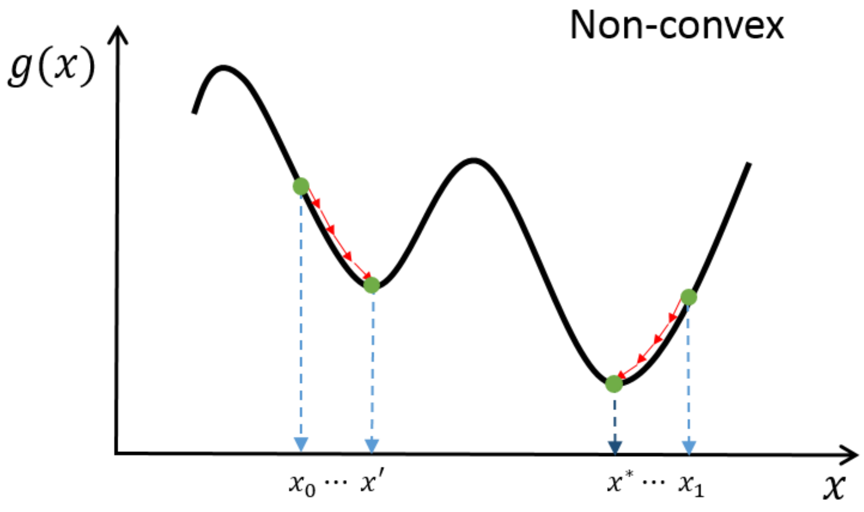


Reduce size over time

Where will We Converge?



Any local minimum is a global minimum



Multiple local minima may exist

- Random initialization
- Multiple trials

Gradient Descent

$$\min \quad (x_1 - 3)^2 + (x_2 - 3)^2$$
$$= \min \quad \frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 & 6 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 18$$

$$f = \frac{1}{2}X^T H X + g^T X$$
$$\nabla f = HX + g$$

- Update rule: $X_{i+1} = X_i - \alpha_i \nabla f(X_i)$

```
H = np.matrix([[2, 0],[0, 2]])
g = -np.matrix([[6],[6]])

x =  np.zeros((2,1))
alpha = 0.2

for i in range(25):
    df = H*x + g
    x = x - alpha*df

print(x)
```

```
[[ 2.99999147]
 [ 2.99999147]]
```

y	$\frac{\partial y}{\partial x}$
Ax	A^T
$x^T A$	A
$x^T x$	$2x$
$x^T A x$	$Ax + A^T x$

Practically Solving Optimization Problems

- The good news: for many classes of optimization problems, people have already done all the “hard work” of developing numerical algorithms
  - A wide range of tools that can take optimization problems in “natural” forms and compute a solution

- We will use CVX (or CVXPY) as an optimization solver
  - Only for convex problems
  - Download: https://www.cvxpy.org/

- Gradient descent
  - Neural networks/deep learning
  - TensorFlow