Least Squares Problem & Optimization

魏家博 (Chia-Po Wei)

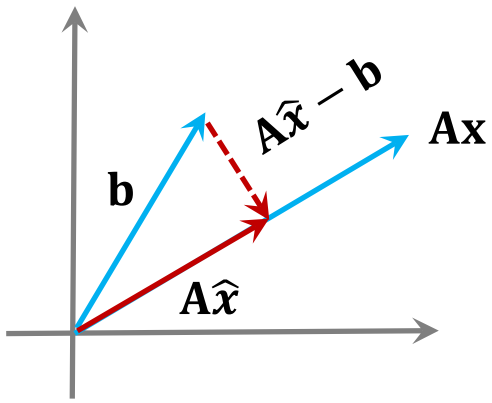
Department of Electrical Engineering

National Sun Yat-sen University

Least Squares Problem

- If a linear system $\mathbf{Ax} = \mathbf{b}$ is inconsistent, we can look for a vector $\hat{\mathbf{x}}$ for which is closest to $\mathbf{b}$. Such solution $\hat{\mathbf{x}}$ is called a least squares solution to the linear system.
- Least squares problems refer to the following optimization problem

$$\min_{\mathbf{x}} \ \|\mathbf{Ax} - \mathbf{b}\|_2^2 = \min_{\mathbf{x}} \ (\mathbf{Ax} - \mathbf{b})^T \ (\mathbf{Ax} - \mathbf{b})$$



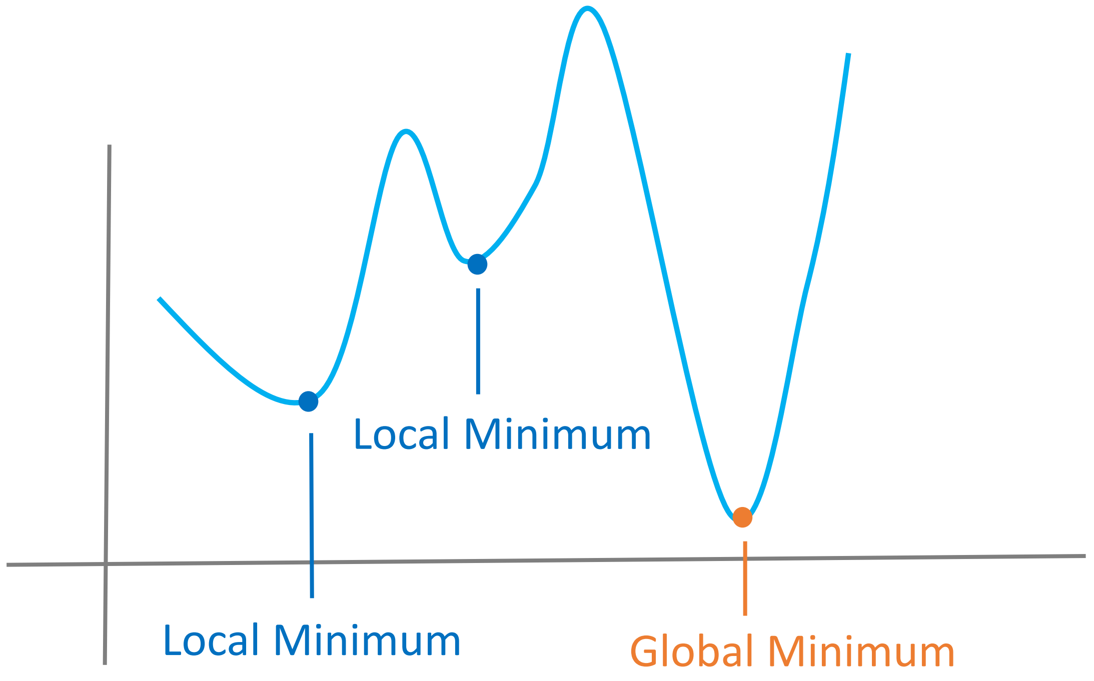
- Notice that the general form of an unconstrained optimization problem is

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

where the function $f: \mathbb{R}^n \to \mathbb{R}$ is called an objective function or a loss function.

Global and Local Minima

- We say f has a global minimum at $\mathbf{x}^*$ if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.
- We say f has a local minimum at $\mathbf{x}^*$ if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$ belonging to a neighborhood of $\mathbf{x}^*$.



Global and Local Minimizer

- We say that $\mathbf{x}^*$ is a global minimizer of the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

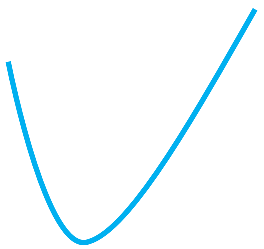
  if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. The following notation is used to denote $\mathbf{x}^*$.

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

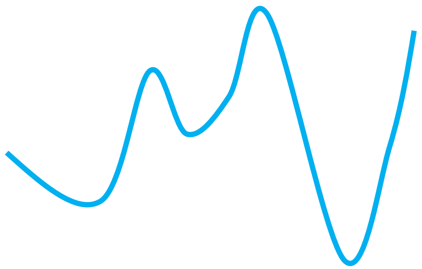
- We say that $\mathbf{x}^*$ is a local minimizer if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x}$ belonging to a neighborhood of $\mathbf{x}^*$.

Convex Objective Functions

- **Theorem**. Any local minimum of a convex objective function is also a global minimum.
- Because the objective function of the least square problem is convex, any local minimum is also a global minimum.
- The objection function of a neural network (**deep learning**) is non-convex.
  - There often exist many local minima that are not global minima.



(a) Convex function



(b) Non-convex function

Single-Variable Calculus

- Least Squares Problem

$$\min_{\mathbf{x}} \ ||\mathbf{Ax} - \mathbf{b}||_2^2 = \min_{\mathbf{x}} \ (\mathbf{Ax} - \mathbf{b})^T \, (\mathbf{Ax} - \mathbf{b})$$

- If $\mathbf{x}$ is a scalar, and let $\mathbf{A} = 2$ and $\mathbf{b} = 3$, then the above becomes

$$\min_{x} \ (2x - 3)^T \, (2x - 3) = \min_{x} \ (2x - 3)^2 = \min_{x} f(x)$$

- The global/local minimum occurs at x satisfying $f'(x) = 0$.

$$f'(x) = 2(2x - 3) \cdot 2 = 0 \quad \Longrightarrow \quad 2x - 3 = 0 \quad \Longrightarrow \quad x = 3/2.$$

Multivariable Calculus

- Least Squares Problem

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 = \min_{\mathbf{x}} (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = \min_{\mathbf{x}} f(\mathbf{x})$$

- We have

$$f(\mathbf{x}) = (\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b}$$

- The global/local minimum occurs at $\mathbf{x}$ satisfying $\nabla f(\mathbf{x}) = 0$.
- $\nabla f(\mathbf{x})$ is called the gradient of the function f.
- How can we calculate $\nabla f(\mathbf{x})$?

Matrix Derivatives

- There are six common types

	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$		

Derivatives by Vector

$$\mathbf{x} = [x_1 \quad \cdots \quad x_n]^T$$
$$\mathbf{y} = [y_1 \quad \cdots \quad y_m]^T$$

Numerator Layout Notation

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1\times n}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

Denominator Layout Notation (we use this one)

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}^T \in \mathbb{R}^{n}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n\times m}$$

- $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ is called the Jacobian matrix of $\mathbf{y}$.

Derivatives of Scalars by Vector

$$\mathbf{x} = [x_1 \quad \cdots \quad x_n]^T, \qquad \mathbf{u} = [u_1 \quad \cdots \quad u_n]^T, \qquad \mathbf{u}^T\mathbf{x} = u_1x_1 + \cdots + u_nx_n$$

(C1) If $\mathbf{u}$ is not a function of $\mathbf{x}$:

$$\frac{\partial(\mathbf{u}^T\mathbf{x})}{\partial\mathbf{x}} = \left[\frac{\partial(\mathbf{u}^T\mathbf{x})}{\partial x_1} \quad \cdots \quad \frac{\partial(\mathbf{u}^T\mathbf{x})}{\partial x_n}\right]^T = [u_1 \quad \cdots \quad u_n]^T = \mathbf{u}$$

(C2) If $\mathbf{u}$ is a function of $\mathbf{x}$:

$$\frac{\partial(\mathbf{u}^T\mathbf{x})}{\partial\mathbf{x}} = \frac{\partial\mathbf{u}}{\partial\mathbf{x}}\mathbf{x} + \frac{\partial\mathbf{x}}{\partial\mathbf{x}}\mathbf{u}$$

Computing $\nabla f(\mathbf{x})$

- $f(\mathbf{x}) = (\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{b}$
- The gradient of $f(\mathbf{x})$ is

$$\begin{aligned}\nabla f(\mathbf{x}) &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{Ax})}{\partial \mathbf{x}} - \frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{b})}{\partial \mathbf{x}} - \frac{\partial(\mathbf{b}^T\mathbf{Ax})}{\partial \mathbf{x}} \\ &= \frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{Ax})}{\partial \mathbf{x}} - 2\frac{\partial(\mathbf{b}^T\mathbf{Ax})}{\partial \mathbf{x}}\end{aligned}$$

- From (C1), $\frac{\partial(\mathbf{b}^T\mathbf{Ax})}{\partial \mathbf{x}} = \frac{\partial\left(\mathbf{A}^T\mathbf{b}\right)^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^T\mathbf{b}$
- From (C2), $\frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{Ax})}{\partial \mathbf{x}} = \frac{\partial\left(\mathbf{A}^T\mathbf{Ax}\right)^T\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial(\mathbf{A}^T\mathbf{Ax})}{\partial \mathbf{x}}\mathbf{x} + \frac{\partial \mathbf{x}}{\partial \mathbf{x}}(\mathbf{A}^T\mathbf{Ax}) = (\mathbf{A}^T\mathbf{A})^T\mathbf{x} + \mathbf{A}^T\mathbf{Ax} = 2\mathbf{A}^T\mathbf{Ax}$
- Hence, $\nabla f(x) = 2\mathbf{A}^T\mathbf{Ax} - 2\mathbf{A}^T\mathbf{b}$

Analytical Solution to Least Squares Problems

- The objective function can be written as

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{b} = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{b}$$

- Setting the gradient of f to zero, we obtain

$$\nabla f(x) = 2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b} = 0 \quad \Longrightarrow \quad \mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b} \quad \text{(Normal Equation)}$$

- If $\mathbf{A}$ has full column rank, then $\mathbf{A}^T\mathbf{A}$ is nonsingular, and we have

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

- Notice that
    - The normal equation has at least one solution, because the range space of $\mathbf{A}^T\mathbf{A}$ and the range space of $\mathbf{A}^T$ are the same, i.e., $\mathrm{R}(\mathbf{A}^T\mathbf{A}) = \mathrm{R}(\mathbf{A}^T)$.
    - If the column vectors of $\mathbf{A}$ are linearly independent, then the normal equation has exact one solution.

Multivariable Example

- If $\mathbf{x} = [x_1, x_2]^T$, and let $\mathbf{A} = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, then
- $\mathbf{Ax} - \mathbf{b} = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} x_1 - 2x_2 - 1 \\ -2x_1 - x_2 + 1 \end{bmatrix}$
- $\min_{\mathbf{x}} \ (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = \min_{\mathbf{x}} (x_1 - 2x_2 - 1)^2 + (-2x_1 - x_2 + 1)^2$
- Method 1 (manually derived, only works for very simple problems):
- $\begin{cases} x_1 - 2x_2 - 1 = 0 \\ -2x_1 - x_2 + 1 = 0 \end{cases} \Longrightarrow \begin{cases} x_1 = 0.6 \\ x_2 = -0.2 \end{cases}$

Multivariable Example (cont.)

- Method 2 (normal equation, only works for least squares problems)
- $\nabla f(x) = 2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b} = 0 \quad \Longrightarrow \quad \mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$
- $\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix}^T \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$, $\mathbf{A}^T\mathbf{b} = \begin{bmatrix} 1 & -2 \\ -2 & -1 \end{bmatrix}^T \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$
- $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 0.6 \\ -0.2 \end{bmatrix}$

Multivariable Example (cont.)

- Method 3 (gradient descent, works for general optimization problems)
- Consider the following sequence

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha\, \nabla f(\mathbf{x}_n), n \geq 0$$

- Since $-\nabla f(\mathbf{x}_n)$ is the negative gradient at $\mathbf{x}_n$, we have

$$f(\mathbf{x}_0) \geq f(\mathbf{x}_1) \geq f(\mathbf{x}_2) \geq \cdots$$

- In particular,

$$\begin{aligned}\mathbf{x}_{n+1} &= \mathbf{x}_n - \alpha\, \nabla f(x) \\ &= \mathbf{x}_n - \alpha(2\mathbf{A}^T\mathbf{A}\mathbf{x}_n - 2\mathbf{A}^T\mathbf{b}) \\ &= (\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})\mathbf{x}_n + 2\alpha\mathbf{A}^T\mathbf{b}\end{aligned}$$

- The scalar α is called the learning rate.

Gradient Descent

Pseudo code

for k in range(max_iteration):

$x^{k+1} = x^k - \alpha \nabla f(x^k)$

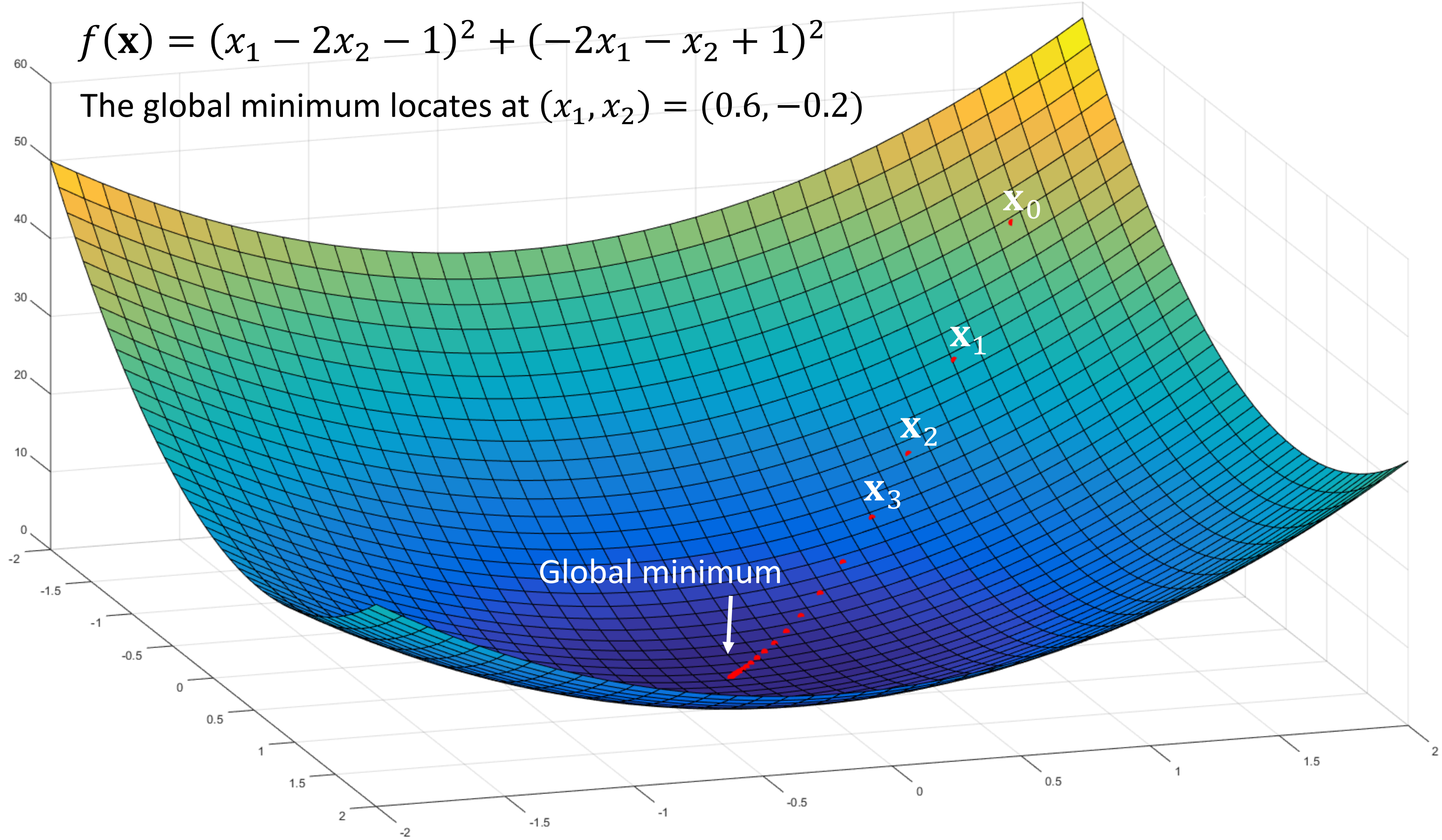
if $\left\| x^{k+1} - x^k \right\|_2 < 10^{-8}$: #stopping criterion

break

Multivariable Example (cont.)

$$f(\mathbf{x}) = (x_1 - 2x_2 - 1)^2 + (-2x_1 - x_2 + 1)^2$$

The global minimum locates at $(x_1, x_2) = (0.6, -0.2)$



Multivariable Example (cont.)

$\alpha = 0.02$

k	(x_1, x_2)	$f(\mathbf{x})$
0	(-2.00, 2.00)	58.00
1	(-1.48, 1.56)	37.12
2	(-1.06, 1.21)	23.76
3	(-0.73, 0.93)	15.20
4	(-0.46, 0.70)	9.73
5	(-0.25, 0.52)	6.23
⋮	⋮	⋮
29	(0.60,-0.20)	0

$\alpha = 0.01$

k	(x_1, x_2)	$f(\mathbf{x})$
0	(-2.00, 2.00)	58.00
1	(-1.74, 1.78)	46.98
2	(-1.51, 1.58)	38.05
3	(-1.30, 1.40)	30.82
4	(-1.11, 1.24)	24.97
5	(-0.94, 1.10)	20.22
⋮	⋮	⋮
60	(0.60,-0.20)	0

- If the learning rate is too small, then the convergence speed might be slow.
- If the learning rate is too large, then the update sequence cannot converge.

References

[1] Steven J. Leon, *Linear Algebra with Applications*, Pearson, 2015.

[2] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press (http://web.stanford.edu/~boyd/cvxbook/)

[3] Kaare B. Petersen and Michael S. Pedersen, *The Matrix Cookbook* (https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf)