Tutorial 3

Xiaojian Xu

02/15/2019

# ESE415 OPTIMIZATION

# Guideline

HW2 Recap

HW3 Recitation

Partial Derivative, Derivative, Gradient, Directional Derivative and Hessian

# 1. HW2 Recap

Lipschitz Constant vs. Largest Eigenvalue of Hessian Matrix

**Problem 5**

We consider an optimization problem that often arises in machine learning. We would like to assign a label $y \in \{-1, +1\}$ to a given input vector $\boldsymbol{x} \in \mathbb{R}^n$. For example, $\boldsymbol{x}$ might denote a vectorized image with $n$ pixels and $y$ a class it corresponds to (*e.g.*, $+1$ for *bird* and $-1$ for *not bird*). We seek to build a linear model

$$\varphi(\boldsymbol{x}; \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_n x_n = \boldsymbol{x}^\mathsf{T} \boldsymbol{w}_1 + w_0 = \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}^\mathsf{T} \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{w} = (\boldsymbol{w}_1, w_0) \in \mathbb{R}^{n+1}$, with $\boldsymbol{w}_1 = (w_1, \ldots, w_n) \in \mathbb{R}^n$, represent the weights. We train the model on a dataset of $m$ data points $\{\boldsymbol{x}_i, y_i\}_{i=1,\ldots,m}$, stored as a label vector $\boldsymbol{y} \in \{-1, +1\}^m$ and a data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{n \times m}$. The columns of $\boldsymbol{X}$ correspond to the input data-vectors $\boldsymbol{x}_i$. The training is performed by minimizing the following cost function

$$f(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \varphi(\boldsymbol{x}_i; \boldsymbol{w}))^2 = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{w}\|^2, \tag{2}$$

where $\boldsymbol{A} = \left[\boldsymbol{X}^\mathsf{T} \mid \mathbf{1}\right] \in \mathbb{R}^{m \times (n+1)}$ is transposed and augmented version of the data-matrix, where every row is made of an input data-vector followed by a constant 1 as in eq. (1). Given a minimizer $\boldsymbol{w}^* \in \mathbb{R}^{n+1}$ of $f$, we can predict the label for any vector $\boldsymbol{x}$ as

$$\widehat{y} = \mathsf{sgn}(\boldsymbol{x}^\mathsf{T} \boldsymbol{w}_1 + w_0) \quad \text{where} \quad \mathsf{sgn}(x) \triangleq \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ +1 & \text{if } x > 0 \end{cases}.$$

In this exercise, we will work with the dataset stored in `dataset.mat`. This file along with the initial script are stored in the archive `assignment2.zip`.

(a) Download, run, and study `ClassificationGM.m` for minimizing eq. (2) via the *gradient method (GM)*. Run the code with parameters $\gamma =$ `stepSize` $= 10^{-4}$, `maxIter` $= 100$, and `tol` $= 10^{-6}$, and the random initialization $\boldsymbol{w}^0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ (*i.e.*, the initialization consists of independent and identically distributed Normal random variables). What is the final value of $f$? Include the generated plot with your solutions.

(b) Provide the expression for the gradient and the Hessian of $f$. Comment on the convexity of $f$.

(c) Analytically find the minimizer of $f$ and comment on the uniqueness of the solution.

(d) Modify the script to implement *GM with backtracking line-search* (as described on Page 4 of Lecture 7). Run the algorithm using the step reduction parameter $\beta = 0.5$, initial $\gamma =$ `stepSize` $= 1$, and fix all the other parameters as in (a). Comment on the convergence of the algorithm. What is the final value of $f$? What is the final value of $\gamma$? Include the generated plot with your solutions. Additionally, print and submit your code with your solutions.

(e) Compute the Lipschitz constant $L$ of $\nabla f(\boldsymbol{w})$. What is the range of possible step-sizes for this problem?

(f) **(Bonus)** Modify the script in order to implement the *accelerated gradient method (AGM)*, with the following update rule for the acceleration parameter

$$\theta_0 = 1 \quad \text{and} \quad \theta_t = \frac{1 + \sqrt{1 + 4\theta_{t-1}^2}}{2}.$$

Run AGM using a *fixed* step size $\gamma =$ `stepSize` $= 10^{-3}$ and fix all the other parameters as in (a). Comment on the convergence of the method. What is the final value of $f$? Include the generated plot with your solutions. Additionally, print and submit your code with your solutions.

# HW2 Recap: problem 5 (e)

- In this case, Lipschitz constant L can be set as the largest eigenvalue of the hessian matrix *Hf(w)*. Here we show the proofs. The proofs contain three parts using following hints .

Hints(a)

**Theorem 4 (Spectral decomposition).** Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Then

1. All of the eigenvalues of $\boldsymbol{A}$ are real.
2. There are $n$ mutually orthogonal, unit-norm eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ corresponding to $n$ eigenvalues $\lambda_1, \ldots, \lambda_n$ (with repetitions included). Therefore, the matrix $\boldsymbol{A}$ can be decomposed in the form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\boldsymbol{U} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_n]$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.
3. We can represent $\boldsymbol{A}$ as

$$\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top.$$

Hints(b)

**Definition 3.** A function $f$ has a *Lipschitz continuous gradient* with constant $L > 0$, when

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|,$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. We will denote this class of functions by $\mathcal{C}_L^1$.

a). First we show that for a symmetric matrix A, we $\|Ax\| \le |\lambda_{max}| \|x\|$ where $\lambda_{max}$ is maximum eigenvalue

*Note that a real symmetric matrix is diagonalizable by an orthogonal matrix.

⇒ So we could have $A = U \Lambda U^T$ (as shown in Theorem 4 (2) above) where $\Lambda = \text{diag}(\lambda_1 \cdots \lambda_n)$.

∴ For $\forall x$,

$x^T A \cdot x = x^T \cdot U \Lambda U^T \cdot x$,

$= (U^T x)^T \cdot \Lambda \cdot (U^T x)$,

Putting $y = U^T x \Rightarrow x^T A x = y^T \Lambda y$

$\|Ax\|_2 = (Ax)^T Ax = x^T A^T A x = x^T \cdot A^2 \cdot x = x^T \cdot (U \Lambda U^T)(U \Lambda U^T) \cdot x$

$= x^T \cdot U \Lambda^2 U^T x = (U^T x)^T \cdot \Lambda^2 \cdot (U^T x)$

Putting $y = U^T x$

$\Rightarrow \|Ax\|_2 = y^T \Lambda^2 y = \sum \lambda_i^2 y_i^2 \le \sum \lambda_{max}^2 y_i^2 = \lambda_{max}^2 \|y\|^2$

$\because \|y\|^2 = y^T \cdot y = (U^T x)^T \cdot (U^T x) = x^T U \cdot U^T \cdot x = x^T x = \|x\|^2$

$\Rightarrow \|Ax\|^2 \le \lambda_{max}^2 \|y\|^2 = \lambda_{max}^2 \|x\|^2$

$\therefore \|Ax\| \le |\lambda_{max}| \|x\|$

b). now we show that for function $f(w) = \frac{1}{2} \|y - Aw\|^2$, we have $\|A^T A (w_1 - w_2)\| \le L \|w_1 - w_2\|$, $w_1, w_2 \in \mathbb{R}^n$

Proof:

$f(w) = \frac{1}{2} \|y - Aw\|^2$

$\nabla f(w) = A^T (Aw - y)$ (gradient)

$Hf(w) = A^T A$ (Hessian)

As shown in lecture notes, $f(w)$ has a Lipschitz continuous gradient with constant $L > 0$, when

$\|\nabla f(w_1) - \nabla f(w_2)\| \le L \|w_1 - w_2\|$, $\forall w_1, w_2 \in \mathbb{R}^n$

plug in $\nabla f(w) = A^T (Aw - y)$, ⇒

$\|A^T (Aw_1 - y) - A^T (Aw_2 - y)\| \le L \|w_1 - w_2\|$

$\|A^T A (w_1 - w_2)\| \le L \|w_1 - w_2\|$

c). Now we know that for function $f(w) = \frac{1}{2} \|y - Aw\|^2$

setting $w_1 - w_2 \to x$, $A^T A \to A$.

from (a) ⇒ $\|A^T A \cdot (w_1 - w_2)\| \le |\lambda_{max}| \|w_1 - w_2\|$ ---- *1

$\lambda_{max}$ is the largest eigenvalue of $A^T A$.

from (b) ⇒ $\|A^T A (w_1 - w_2)\| \le L \|w_1 - w_2\|$ ---- *2

L is the Lipschitz constant of $f(w)$.

⇒ Set $L_0$ as the smallest Lipschitz constant.

then we could set $L_0 = |\lambda_{max}|$, $L \ge |\lambda_{max}|$

# Proofs

# 2. HW2 Recitation (P1)

## Problem

The goal of this problem is to give some intuition about different rates of convergence of $\{f(\boldsymbol{x}^t)\}_{t\in\mathbb{N}}$ to $f^* = f(\boldsymbol{x}^*)$. As a reminder, the rate of convergence is characterized via two values $p \in \mathbb{N}$ and $\beta \in [0, \infty)$. Specifically, the goal is to find the largest $p$ such that

$$\beta = \limsup_{t\to\infty} \frac{|f(\boldsymbol{x}^t) - f^*|}{|f(\boldsymbol{x}^{t-1}) - f^*|^p} < \infty.$$

Let $\delta \in (0, 1)$ and consider $r_t \triangleq |f(\boldsymbol{x}^t) - f^*|$. Categorize the convergence of the following sequences as *sublinear*, *linear*, *superlinear*, and *quadratic*. Justify and explain your answer.

a) $r_t = \frac{\delta}{t}$

~~b) $r_t = \delta^t$~~

~~c) $r_t = \delta^{t^2}$~~

~~d) $r_t = \delta^{2^t}$~~

## Hints

**Definition 2.** Let $\{f_k\}$ be a convergent sequence with $f_k \to f$. The *order of convergence* of $\{f_k\}$ is

$$p^* = \sup\left\{p \in \mathbb{N} : \limsup_{k\to\infty} \frac{|f_k - f|}{|f_{k-1} - f|^p} < \infty\right\}.$$

The larger $p^*$ implies faster convergence.

**Definition 3.** Let $p$ denote the order of convergence of $\{f_k\}$ to $f$ and let

$$\beta = \limsup_{k\to\infty} \frac{|f_k - f|}{|f_{k-1} - f|^p}.$$

We say that the convergence is

- *linear* or *geometric* if $p = 1$ and $\beta \in (0, 1)$.
- *superlinear* if $p = 1$ and $\beta = 0$.
- *sublinear* if $p = 1$ and $\beta = 1$.
- *quadratic* if $p = 2$ and $\beta < \infty$.

# 2. HW3 Recitation (P1)

Solution

a) The convergence is *sublinear* as for $p = 1$, we have that

$$\frac{r_t}{r_{t-1}} = \frac{\frac{\delta}{t}}{\frac{\delta}{t-1}} = \frac{t-1}{t} = 1 - \frac{1}{t} \xrightarrow[t\to\infty]{} 1 = \beta.$$

Note that for any other $p > 1$, the value given by $\limsup$ will be unbounded.

# 3. Partial derivative, derivative, gradient, directional derivative and Hessian

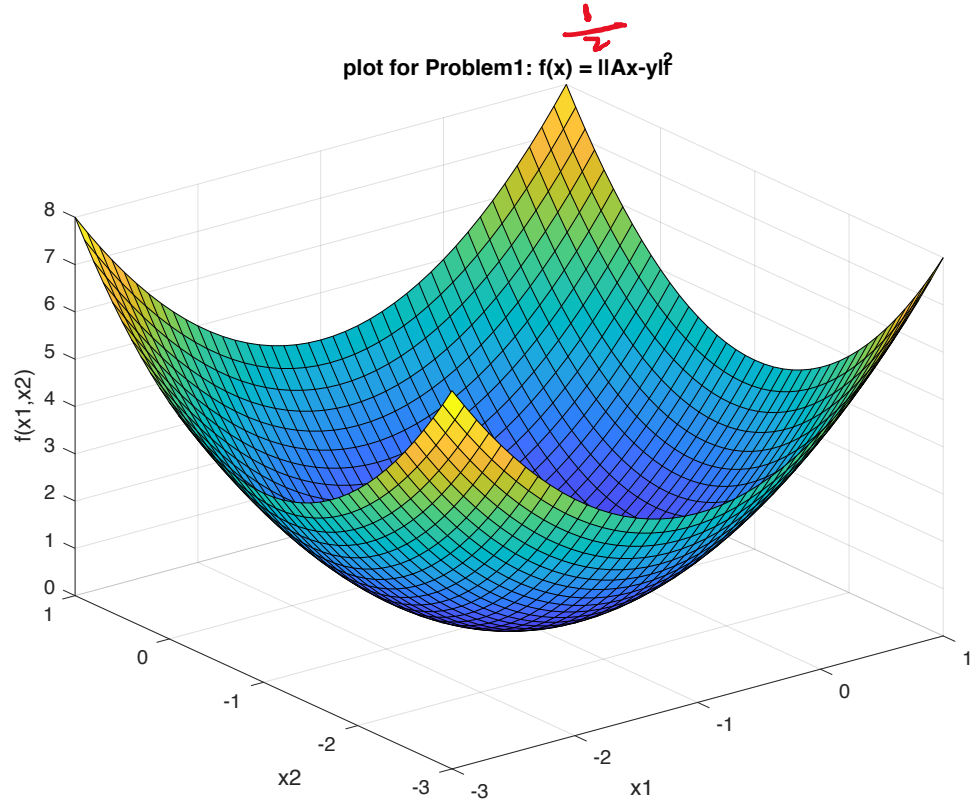
1. We will consider functions of form

$$f(\boldsymbol{x}) \quad \text{where} \quad \boldsymbol{x} \in \mathbb{R}^n$$

2. Vectors will be generally assumed to be laid out in the column form

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \quad \text{and} \quad \boldsymbol{x}^\mathsf{T} = [x_1 \cdots x_n],$$

where the second notation indicates the transpose.




$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 \quad \text{where} \quad \boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} -2 \\ 0 \end{bmatrix} \quad \text{with} \quad \boldsymbol{x} \in \mathbb{R}^2.$$

# 3.1 Partial Derivative

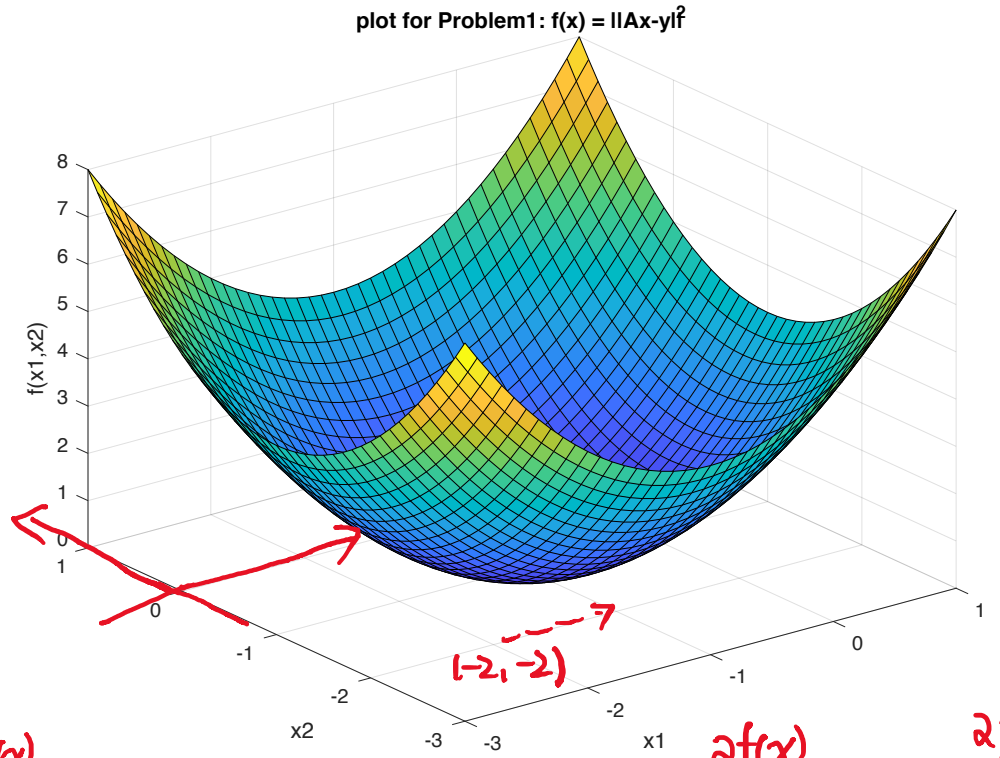
3. We denote and define the partial derivative of $f$ with respect to coordinate $x_i$ as

$$\frac{\partial f(\boldsymbol{x})}{\partial x_i} \triangleq \lim_{\delta \to 0} \frac{f(\boldsymbol{x} + \delta \boldsymbol{e}_i) - f(\boldsymbol{x})}{\delta},$$

where $\boldsymbol{e}_i \in \mathbb{R}^n$ is a vector that is 0 in all coordinates except at $i$, where it is 1.

Highlight:

- Partial derivative implies the changing rate on one coordinate direction at a certain point .
- Its sign implies the increasement or reduction of the function and its magnitude implies the changing speed (rate).




$\frac{\partial f(x)}{\partial x_1} = 2x_1 + 2$

$\frac{\partial f(x)}{\partial x_2} = 2x_2 + 2$

at point $(-2,-2) \Rightarrow \frac{\partial f(x)}{\partial x_1} = -2 = \frac{\partial f(x)}{\partial x_2}$

# 3.2 Derivative

4. We denote the derivative of a vector-valued function $\boldsymbol{y}(\boldsymbol{x}) \in \mathbb{R}^m$ with respect to $\boldsymbol{x} \in \mathbb{R}^n$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \mathrm{D}\boldsymbol{y}(\boldsymbol{x}) \triangleq \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Highlight:

- Derivative here is defined for the functions f: $R^n \rightarrow R^m$;
- Set $y_i = g_i(\mathbf{x})$, each row can be regarded as the gradient of $\nabla g_i(\mathbf{x})^T$;
- When m = 1, it becomes gradient.

Example:

$f(x) = Ax$, $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^{n \times 1}$

$\Rightarrow f(x) \in \mathbb{R}^{m \times 1}$ (vector)

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$f(x) = \begin{bmatrix} x_1 + 4x_2 , & 2x_1 + 5x_2, & 3x_1 + 6x_2 \end{bmatrix}^T$$

# 3.3 Gradient

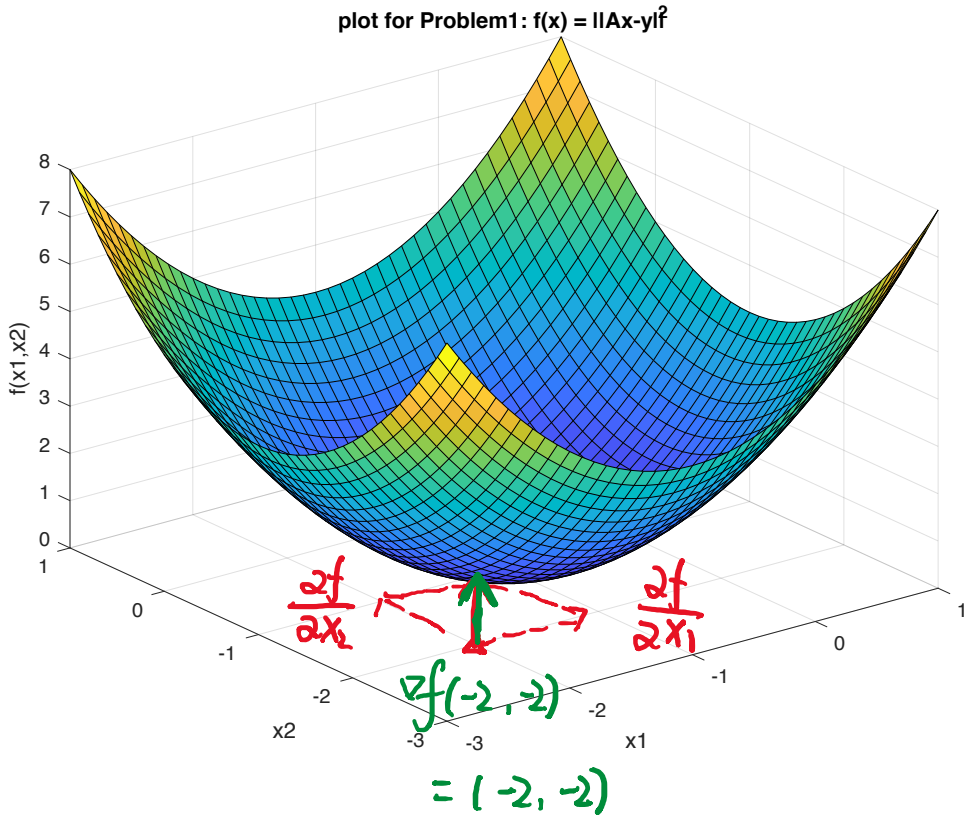
5. Gradient of $f$ at $\boldsymbol{x} \in \mathbb{R}^n$

$$\nabla f(\boldsymbol{x}) = [\mathrm{D}f(\boldsymbol{x})]^{\mathsf{T}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix},$$

where we adopted a column-layout convention for convenience.

Highlight:

- Gradient here is defined for the functions f: $R^n \rightarrow R$;
- It is a vector consists of partial derivative.
- It implies the direction that function *f* changes fastest.



# 3.4 Directional Derivative

**Definition 1.** The directional derivative of $f(\boldsymbol{x})$ along $\boldsymbol{u} \in \mathbb{R}^n$, $\|\boldsymbol{u}\| = 1$, is

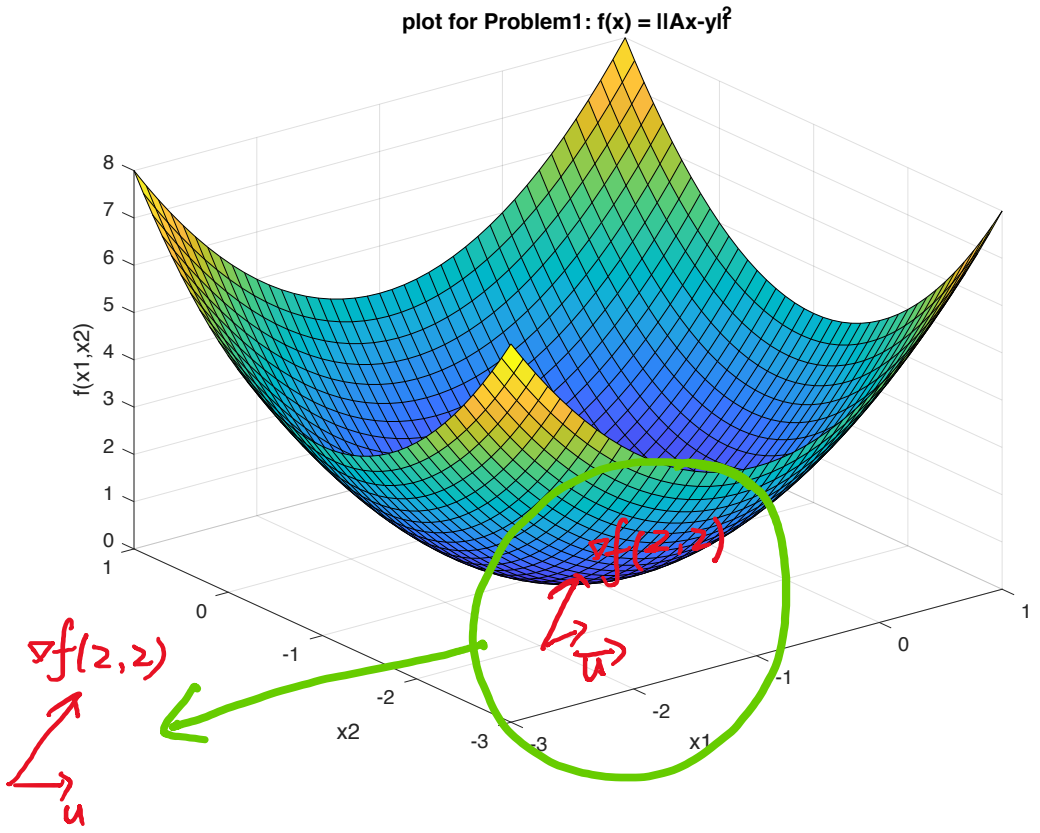
$$\nabla_{\boldsymbol{u}} f(\boldsymbol{x}) = \lim_{\gamma \to 0} \frac{f(\boldsymbol{x} + \gamma \boldsymbol{u}) - f(\boldsymbol{x})}{\gamma} = \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{u}.$$

*Proof.* We consider the Taylor's expansion $f(\boldsymbol{x} + \gamma \boldsymbol{g}) = f(\boldsymbol{x}) + \gamma \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{g} + o(\gamma)$, where $o(\gamma)/\gamma \to 0$ when $\gamma \to 0$. Then, we have that

$$\frac{f(\boldsymbol{x} + \gamma \boldsymbol{g}) - f(\boldsymbol{x})}{\gamma} = \nabla f(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{g} + \frac{o(\gamma)}{\gamma}.$$

Highlight:

- Directional derivative  is a scaler;
- It show the projection of the gradient on a certain direction **u;**
- It measures the changing speed at the direction **u.**
- u = - ∇f(**x**) /|| ∇f(**x**) ||points in the direction of the fastest local decrease of *f* (this is the how gradient method works)



# 3.5 Hessian

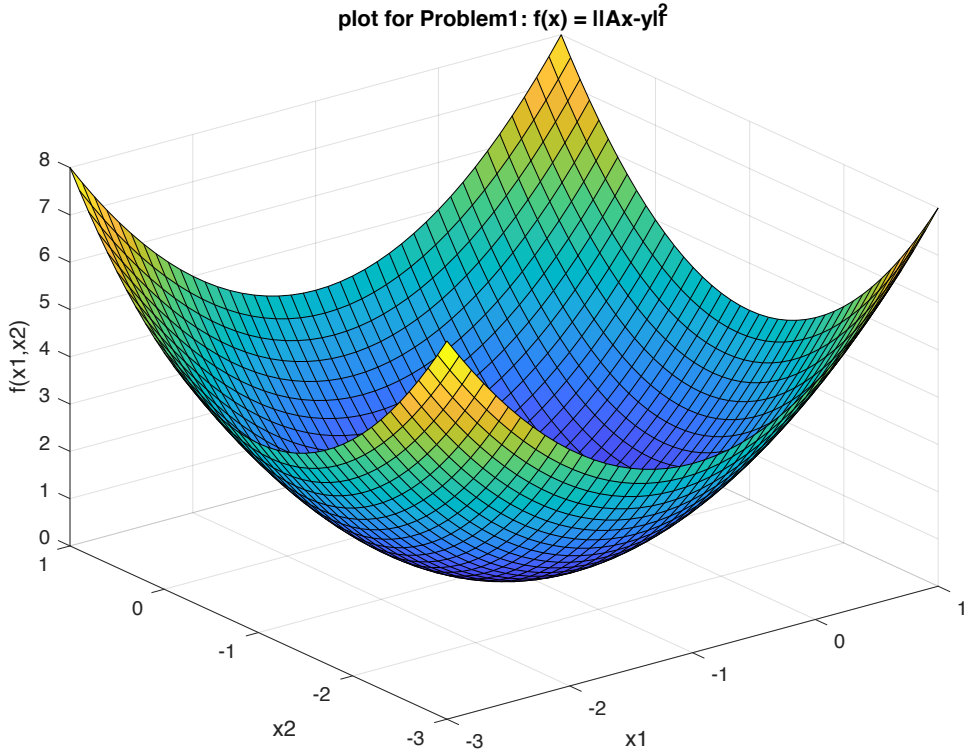
6. Hessian of $f$ at $\boldsymbol{x} \in \mathbb{R}^n$

$$\mathrm{H}f(\boldsymbol{x}) = \mathrm{D}^2 f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix},$$

where we note that Hessian is a symmetric matrix, *i.e.*, $\mathrm{H}f(\boldsymbol{x}) = [\mathrm{H}f(\boldsymbol{x})]^\mathsf{T}$.

Highlight:

- Hessian is defined here is for function f: $R^n \rightarrow R$;
- Hessian is a matrix with size n*n;



# Thanks