# Preliminary to machine learning

## — 有关范数、凸、矩阵微分、和梯度计算

Jingbo Xia

Huazhong Agricultural University

*xiajingbo.math@gmail.com*

2023-12-05

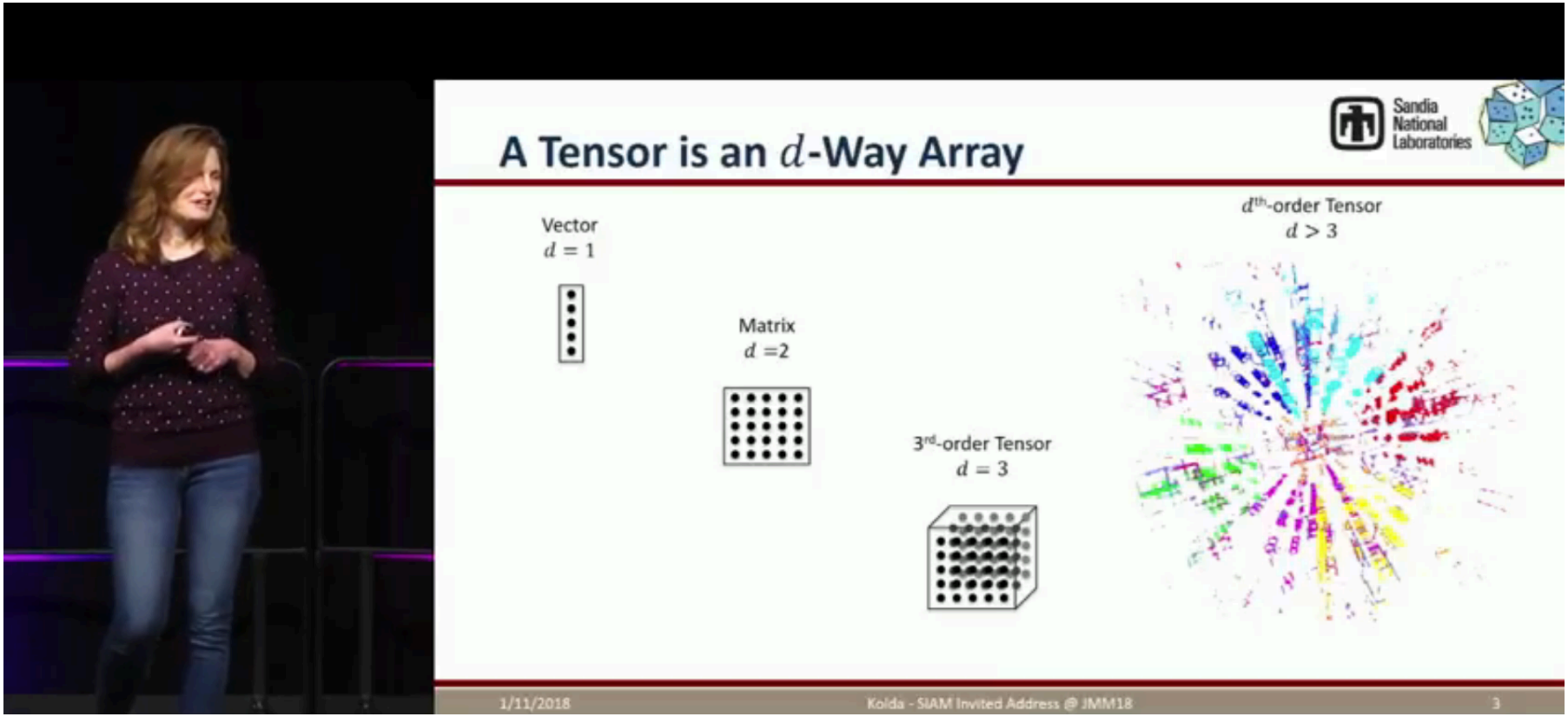
# Table of contents I

# Outline

# 从几何空间到任意维的 Rn 空间

**Why arbitrary n-dimensional space.**

We call the $n$-dimensional space $R^n$.




1

[1] Picture from Tarama Kolda's talk in SIAM, Jan 2018.

# 从几何空间到任意维的 Rn 空间

## Point, or vector?

How do you treat a "vector" in a $X$-$Y$ plane? A dot, an arrow with point and line segment, or a vector $(x, y)$?


2

[2] 代数脑, or 几何脑？

# 从几何空间到任意维的 Rn 空间

**Isomorphism of geometrical thinking and algebraic thinking**

**术语的挑选**：A 2-dimensional plane. "$Oxy$-plane" or $R^2$.
In $Oxy$-plane, the vector referred is an arrow with both length and direction, and the addition follows the rules of "Parallelogram rule" / Law of addition of vectors.
In $R^2$, the vector refers to a two-dimensional array $(x, y)$, and the addition follows $(x_1, y_1) + (x_2, y_2) - (x_1 + x_2, y_1 + y_2)$.

The two sets with additive computation are isomorphic same.
From the group theory, the set $G$ with a multiplication $\cdot$ is denoted as a group (群) [3], if it satisfies:
1. For each $x, y \in G$, $x \cdot y \in G$;
2. Exist an element $e$, s.t., $xe = ex = x$;
3. For each $x \in G$, there is a $z$, s.t., $xz = zx = e$, and $z$ is actually the inverse of $x$, and so $z := x^{-1}$.

Apparently, the $Oxy$ plane with addition and $R^2$ with addition are both groups.

[3]参考：群论, group theory.

# 从几何空间到任意维的 Rn 空间

## Isomorphism of geometrical thinking and algebraic thinking

[Definition]Isomorphism (同构). The groups $S_1(+)$ and $S_2(\cdot)$ are isomorphic if there is a bijection $f$: $S_1 \longrightarrow S_2$, which satisfies:

$$\text{For each } x, y \in S_1, f(x+y) = f(x) \cdot f(y).$$

Therefore, one will not differentiate the geometrical understanding and algebraic understanding.

因此，上述准备帮助我们理解任意维的 $R^n$中的点 (dot), 线段 (line segment) 和超平面 (hyper plane).

# 从几何空间到任意维的 Rn 空间

## Dot, line segment and hyper plane in Rn space

What is dot?
What is line segment?
What is hyper plane?

# 从几何空间到任意维的 Rn 空间

## Dot, line segment and hyper plane in Rn space

What is dot?

$X \in \mathbb{R}^n$ is a dot, or a vector. In another word, a dot with a "higher" geometrical taste.

What is line segment?

A line segment is the set of all dots which connect $X_1 \in \mathbb{R}^n$ and $X_2 \in \mathbb{R}^n$ in a "straight" manner. Think about it...

What is hyper plane?

I guess you are thinking about $Ax + By + cz = 0$, notation used in high school textbook.

# 从几何空间到任意维的 Rn 空间

## Dot, line segment and hyper plane in Rn space

We have given definition of dot in $R^n$, before go ahead to more definitions, e.g., convex sets, we need to define a "line segment" in $\mathbb{R}^n$. But, why[4]?
First, let's consider the existence of parameter $\theta$ in $\mathbb{R}^1$ space.

**Existence of** $\theta$: Let $x_1, x_2$ be two scalar value in $\mathbb{R}$, then for each value $x_0 \in (x_1, x_2)$, there exists a parameter $\theta$ such that

$$x_0 = \theta x_1 + (1 - \theta) x_2.$$

Proof: Proof of this theorem is straightforward, if one consider the ratio of $\frac{x_2 - x_0}{x_2 - x_1}$.
I will leave this as an assignment.

[4] Thank graduate student, Ms. YingLiu, who suggested to add this in the slides when she was in my DM class, Fall, 2018.

# 从几何空间到任意维的 Rn 空间

## Dot, line segment and hyper plane in Rn space

Then, what is "line segment" in $\mathbb{R}^2$ or $\mathbb{R}^n$?

**Line segment in the $\mathbb{R}^2$ space**: for two points $X_1 = (x_{11}, x_{12}), X_2 = (x_{21}, x_{22}) \in \mathbb{R}^2$, the line segment connected two points consists of points $X_0$ which satisfies:

$$X_0 = \theta X_1 + (1 - \theta) X_2,$$

where $\theta \in [0, 1]$.

**Line segment in the $\mathbb{R}^n$ space** is a natural extension of the above scenario to $\mathbb{R}^n$.

Then, what is "plaine" in $\mathbb{R}^2$ or $\mathbb{R}^n$?

**Plane in the $\mathbb{R}^3$ space**: $Ax + By + Cz + D = 0$

**Hyperplane in the $\mathbb{R}^n$ space** is a natural extension of the above scenario to $\mathbb{R}^n$: $< W, X > +b = 0$.

# Outline

# 范数 (Norm)

Why is norm important?

# 范数 (Norm)

Because measure is important?

# 范数 (Norm)

Because measure is important?

# 范数 (Norm)

## Definition of Norm

**Norm**: A norm $||\cdot||$ is a map from $\mathbb{R}^n$ to $\mathbb{R}$, which satisfies:

$$||X|| \geq 0, \text{ for any } X \in \mathbb{R}^n; \text{ and } ||X|| = 0, \text{ only if } X = 0,$$

$$||cX|| = |c|||X||, \text{ for any } X \in \mathbb{R}^n;$$

$$||X + Y|| \leq ||X|| + ||Y||, \text{ for any } X, Y \in \mathbb{R}^n. \text{ (Triangular inequality.)}$$

# 范数 (Norm)

## lp norms

For better details, the $lp$ norms are defined as below:

$$l_p\ norm = ||x||_p = \begin{cases} \#(i\ |x_i \neq 0), & \text{if } p = 0 \text{ (This norm doesn't follow} \\ \sum_i |x_i|, & \text{if } p = 1 \\ \sqrt{\sum_i x_i^2}, & \text{if } p = 2 \\ (\sum_i x_i^p)^{\frac{1}{p}}, & \text{for arbitrary } p \in \mathbb{N} \\ (\sum_i x_i^\infty)^{\frac{1}{\infty}} = \max(|x_i|), & \text{if } p = \infty \end{cases} \tag{1}$$

# 范数 (Norm)

## lp norms

**思考**

It is your turn to think about the shape of the "norm-ball" now.

---

$\{X \in \mathbb{R}^n \mid \ X$ satisfies: $||X|| \leq C$, where $C$ is a constant.$\}$

# 范数 (Norm)

## lp norms

### 思考

It is your turn to think about the shape of the "norm-ball" now.

$\{X \in \mathbb{R}^n \mid X$ satisfies: $||X|| \leq C$, where $C$ is a constant.$\}$

### 思考

《矩阵论》中还定义了诸多不同的向量范数和矩阵范数。

他们的用意在哪里？

# Outline

# 凸性 (Convexity) I

## 凸集 (Convex set)

A set $C$ is convex if the line segment between any two points in $C$ lies in $C$, i.e., if for any $x_1, x_2 \in C$ and any $\theta$ with $0 \leq \theta \leq 1$, we have

$$\theta x_1 + (1 - \theta) x_2 \in C.$$

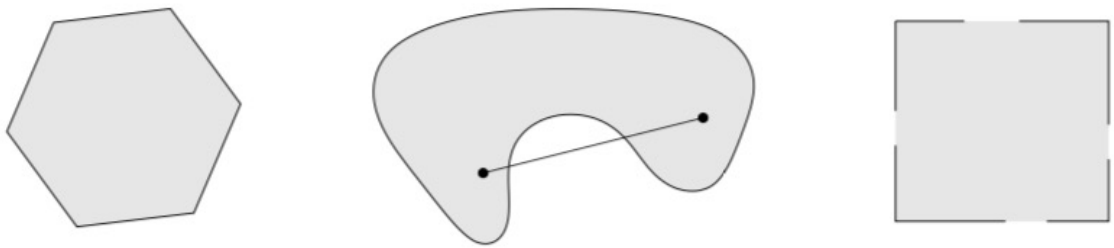


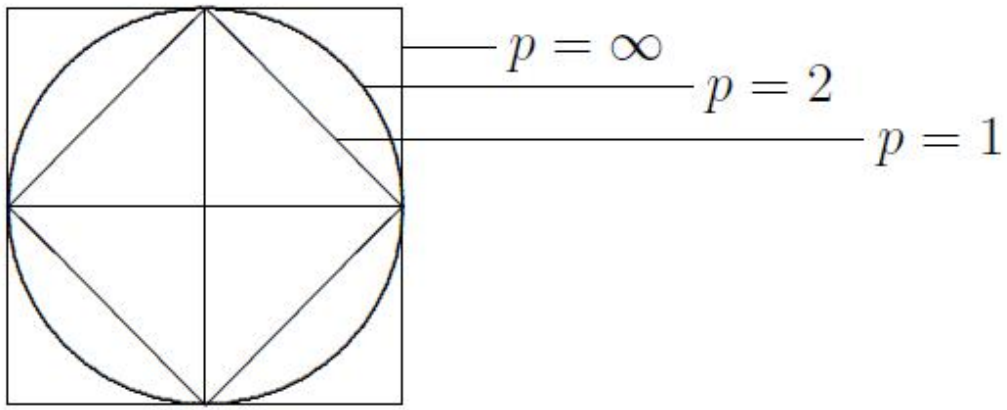
图 1: Some simple convex and nonconvex sets. *Left*. The hexagon, which includes its boundary (shown darker), is convex. *Middle*. The kidney shaped set is not convex, since the line segment between the two points in the set shown as dots is not contained in the set. *Right*. The square contains some boundary points but not others, and is not convex.

# 凸性 (Convexity) II

## 凸集 (Convex set)

**Some examples for convex sets in $\mathbb{R}^n$:**

1. Hyperplane: $< W, X > +b = 0$. This is an example we already discussed in SVM.
2. Half divided space by hyper plane: $< W, X > +b \geq 0$.
3. Norm-ball: $||x||_2 \leq C$. Here we denote $|| \cdot ||_2$ without discussion, but you can just think it in a Euclidean taste. We will discuss the norm later.



# 凸性 (Convexity) I

## 凸优化 (Convex optimization)

A (scalar) convex optimization problem is one of the form

$$\begin{cases} \text{minimize } f_0(x) \\ \text{subject to } f_i(x) \leq b_i, i = 1, \cdots, m \end{cases} \tag{2}$$

*a* *b*

[a] 凸优化问题。
[b] 注：凸优化要求函数和可行区间均具有凸性。

where the functions $f_0, \cdots, f_m : \mathbb{R} \to \mathbb{R}$ are convex (**凸函数**), i.e., satisfy

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y))$$

for all $x, y \in \mathbb{R}$ and all $\alpha,\ \beta \in \mathbb{R}$ with $\alpha + \beta = 1$, $\alpha \geq 0, \beta \geq 0$.

# 凸性 (Convexity) II

## 凸优化 (Convex optimization)

A (general) convex optimization problem is the form

$$\begin{cases} \text{minimize } f_0(X) \\ \text{subject to } f_i(X) \leq 0, i = 1, \cdots, p, \\ \qquad\qquad f_i(X) = 0, i = p+1, \cdots, m \end{cases} \tag{3}$$

[a]

[a]更为一般的凸优化问题

where the functions $f_0, \cdots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex, i.e., satisfy

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y))$$

for all $x, y \in \mathbb{R}^n$ and all $\alpha,\ \beta \in \mathbb{R}$ with $\alpha + \beta = 1$, $\alpha \geq 0, \beta \geq 0$. [5]

[5]凸优化问题中，局部最优解即为全局最优解，这使得寻优步骤极大增速。这使得凸性是优化问题中较受欢迎的性质。

# Outline

# 梯度 (Gradient) 和 Hessian 矩阵 I

## Gradient

In mathematics, the gradient is a multi-variable generalization of the derivative. While a derivative can be defined on functions of a single variable, for functions of several variables, the gradient takes its place. The gradient is a vector-valued function, as opposed to a derivative, which is scalar-valued.

Consider a room in which the temperature is given by a scalar field, $T$, so at each point $(x, y, z)$ the temperature is $T(x, y, z)$. (Assume that the temperature does not change over time.) At each point in the room, the gradient of $T$ at that point will show the direction in which the temperature rises most quickly. The magnitude of the gradient will determine how fast the temperature rises in that direction.

# 梯度 (Gradient) 和 Hessian 矩阵 II

## Gradient

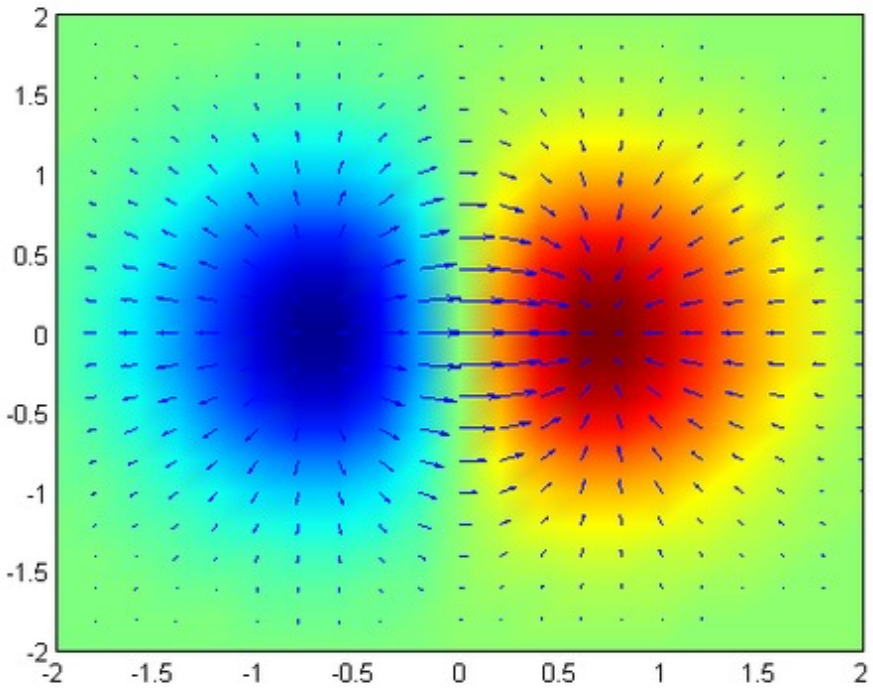



图 2: Gradient of the 2D function $f(x, y) = xe^{(x2+y2)}$ is plotted as blue arrows over the pseudocolor plot of the function.

# 梯度 (Gradient) 和 Hessian 矩阵 III

## Gradient

Consider a surface whose height above sea level at point $(x, y)$ is $H(x, y)$. The gradient of $H$ at a point is a vector pointing in the direction of the steepest slope or grade at that point. The steepness of the slope at that point is given by the magnitude of the gradient vector.

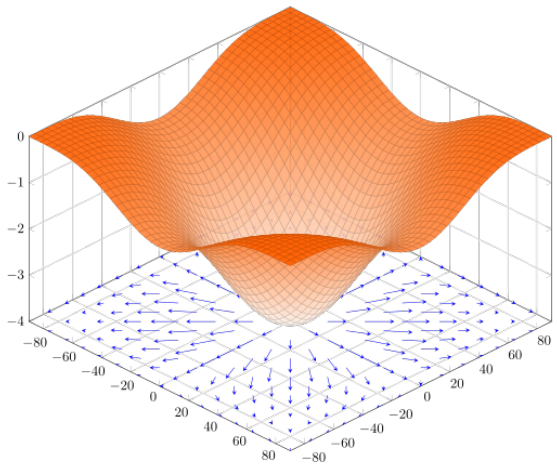



图 3: The gradient of the function $f(x, y) =?(\cos^2 x + \cos^2 y)^2$ depicted as a projected vector field on the bottom plane.

**Cartesian coordinates**

In the three-dimensional Cartesian coordinate system with a Euclidean metric, the gradient, if it exists, is given by:

$$\nabla f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k} = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix}.$$

For example, the gradient of the function $f(x, y, z) = 2x + 3y^2 - \sin(z)$ is

$$\nabla f = 2\mathbf{i} + 6y\mathbf{j} - \cos(z)\mathbf{k} = \begin{pmatrix} 2 \\ 6y \\ -\cos(z) \end{pmatrix}.$$

And generally, for a $x \in \mathbb{R}^n$ and a multi-variant function $f(x)$,

$$\nabla f(x) = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n})^T.$$

## 定理

*If* $f(x) = b^T x = x^T b$, *then* $\nabla f(x) = b$.
*If* $f(x) = x^T A x$, *then* $\nabla f(x) = A^T x + Ax$.
*If* $f(x) = g(x)^T h(x)$, *with* $g, h : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, *then*

$$\nabla f(x) = \begin{pmatrix} h(x)^T \frac{\partial g(x)}{\partial x_1} \\ \vdots \\ h(x)^T \frac{\partial g(x)}{\partial x_n} \end{pmatrix} + \begin{pmatrix} g(x)^T \frac{\partial g(x)}{\partial x_1} \\ \vdots \\ g(x)^T \frac{\partial h(x)}{\partial x_n} \end{pmatrix} \tag{4}$$

## 定理 (Basic result)

*For* $w, x \in \mathbb{R}^n$, *assume* $f(x, w) =< x, w >= x^T w$, *we have:*
$\frac{\partial f(x,w)}{\partial w} = x$.

Proof: The proof is straightforward.
Denote $x = (x_1, x_2, \cdots, x_n)^T$, and $w = (w_1, w_2, \cdots, w_n)^T$. And straightforward verification shows that $\frac{\partial f(x,w)}{\partial w} = \begin{pmatrix} \frac{\partial f(x,w)}{\partial w_1} \\ \vdots \\ \frac{\partial f(x,w)}{\partial w_n} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x$.

# 梯度 (Gradient) 和 Hessian 矩阵 I

## Hessian 矩阵

In mathematics, the Hessian matrix or Hessian is a square matrix of second-order partial derivatives of a scalar-valued function, or scalar field. It describes the local curvature of a function of many variables. The Hessian matrix was developed in the 19-th century by the German mathematician Ludwig Otto Hesse and later named after him. Hesse originally used the term "functional determinants".

图 4: Ludwig Otto Hesse (1811 - 1874)

# 梯度 (Gradient) 和 Hessian 矩阵 III

## Hessian 矩阵

Suppose $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a function taking as input a vector $x \in \mathbb{R}^n$and outputting a scalar $f(x) \in \mathbb{R}$; if all second partial derivatives of $f$ exist and are continuous over the domain of the function, then the Hessian matrix $H$ of $f$ is a square $n \times n$ matrix, usually defined and arranged as follows:

$$H(x) = \nabla^2 f(x) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \, \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \, \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \, \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \, \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

# 梯度 (Gradient) 和 Hessian 矩阵 I

## Hessian Matrix: Second derivative test

The Hessian matrix of a convex function is positive semi-definite.
Refining this property allows us to test if a critical point x is a local maximum, local minimum, or a saddle point, as follows:

If the Hessian is positive definite at $x$, then $f$ attains an isolated local minimum at $x$. If the Hessian is negative definite at $x$, then $f$ attains an isolated local maximum at $x$. If the Hessian has both positive and negative eigenvalues then $x$ is a saddle point for $f$. Otherwise the test is inconclusive. This implies that, at a local minimum (respectively, a local maximum), the Hessian is positive-semi-definite (respectively, negative semi-definite).

# 梯度 (Gradient) 和 Hessian 矩阵 I

## Multivariant Newton method

Hessian matrices are used in large-scale optimization problems within Newton-type methods because they are the coefficient of the quadratic term of a local Taylor expansion of a function. That is, for $x \in \mathbb{R}^n$, and $x_0$ a vector near $x$,

$$y = f(x) \approx f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T H(x_0)(x - x_0), \tag{5}$$

where $\nabla f$ is the gradient.

[A note here][6].

# 梯度 (Gradient) 和 Hessian 矩阵 II

## Multivariant Newton method

Take the second-order Taylor expansion around the point $x_o$ in equation (5), we would like to find a point at which the quadratic form on the right hand side is minimized. Supposing that $H(x_0)$ is positive definite, the unique minimum is obtained at

$$x = x_0 - H(x_0)^{-1}\nabla f(x_0). \tag{6}$$

So, Newton's method can be summed up as:

$$x^{(t+1)} = x^{(t)} - \alpha_t H(x^{(t)})^{-1}\nabla f(x^{(t)}). \tag{7}$$

[Assignments][7]:

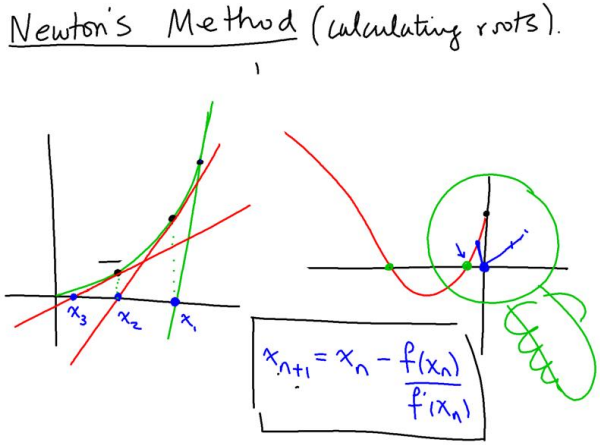
Answersheet to (1). From equation (5), we have

$$0 = \nabla f(x) = \nabla f(x_0) + H(x_0)(x - x_0),$$

and we have $x = x_0 - H(x_0)^{-1}\nabla f(x_0)$.

Note: Newton's method converges quadratically near the root, and it requires inverting a Hessian matrix, $H(x_0)$, which is $O(n^3)$ with standard technique. As a result, each iteration can be quite expensive for large $n$.
Quasi-Newton methods overcome many of these limitations.

Answersheet to (2). The multivariate Newton's method for minimization is very similar to Newton's method for root finding.




Here,

$$f'(x_t) = k = \frac{f(x_t)}{x_t - x_{t+1}}, \text{ and } x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}.$$

In the case of the minimization, is to replace $f$ to $f'$ and find the root of $f'$, so we have the Newton iteration in terms of $f$:

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)},$$

What is the uni-variate form of the Newton's method?

[6]Please be noted that the Hessian appears in the second-order Taylor expansion of a scalar field (if you need recall something in Calculus):

$$y = f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}\nabla^2 f(x_0)(x - x_0)^2 + O((x - x_0)^3),$$

where $x \in \mathbb{R}$, and $\nabla$ denotes derivative symbol.

[7](1). Prove formula (6);
(2). Recall and list the uni-variant form of this formula.

Gradient descent (aka., steepest descent) is for minimizing multidimensional smooth convex objective functions of the form $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$.

## 定理 (Gradient Descent)

*1:* ***Input:*** *Initial point* $w_0$*, gradient norm tolerance* $\varepsilon$
*2:* *Set t=0*
*3:* ***while*** $||\nabla \mathcal{J}(w_t)|| \geq \varepsilon$ ***do***
*4:* $\quad w_{t+1} = w_t - \eta_t \nabla_t J(w_t)$
*5:* $\quad t = t + 1$
*6:* ***end while***
*7:* ***Return:*** $w_t$

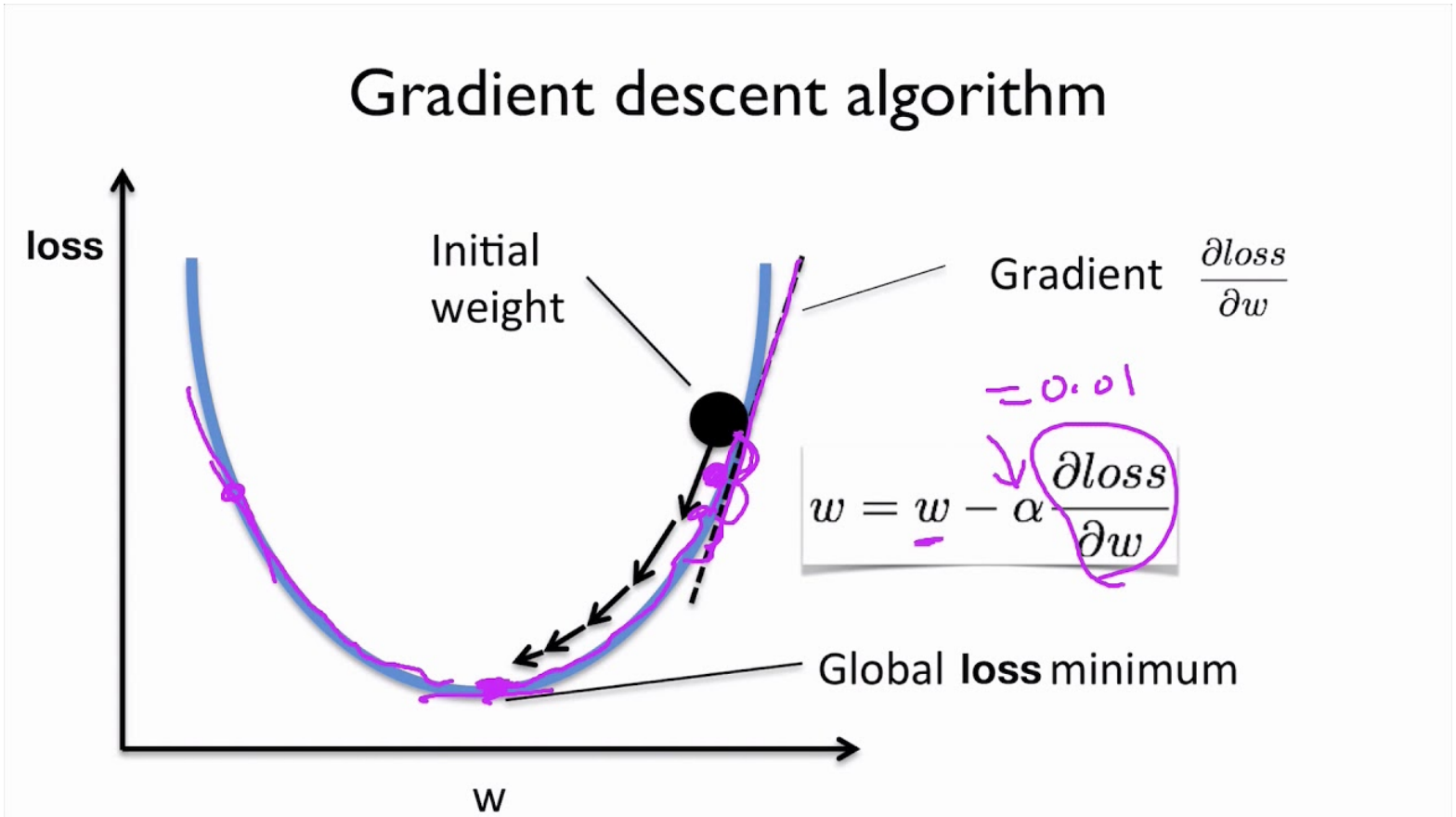
Here,

$$w_{t+1} = w_t - \eta_t \nabla_t J(w_t) \tag{8}$$

is the core iteration.

# 梯度 (Gradient) 和 Hessian 矩阵 II

## 梯度下降 (Gradient Descent)

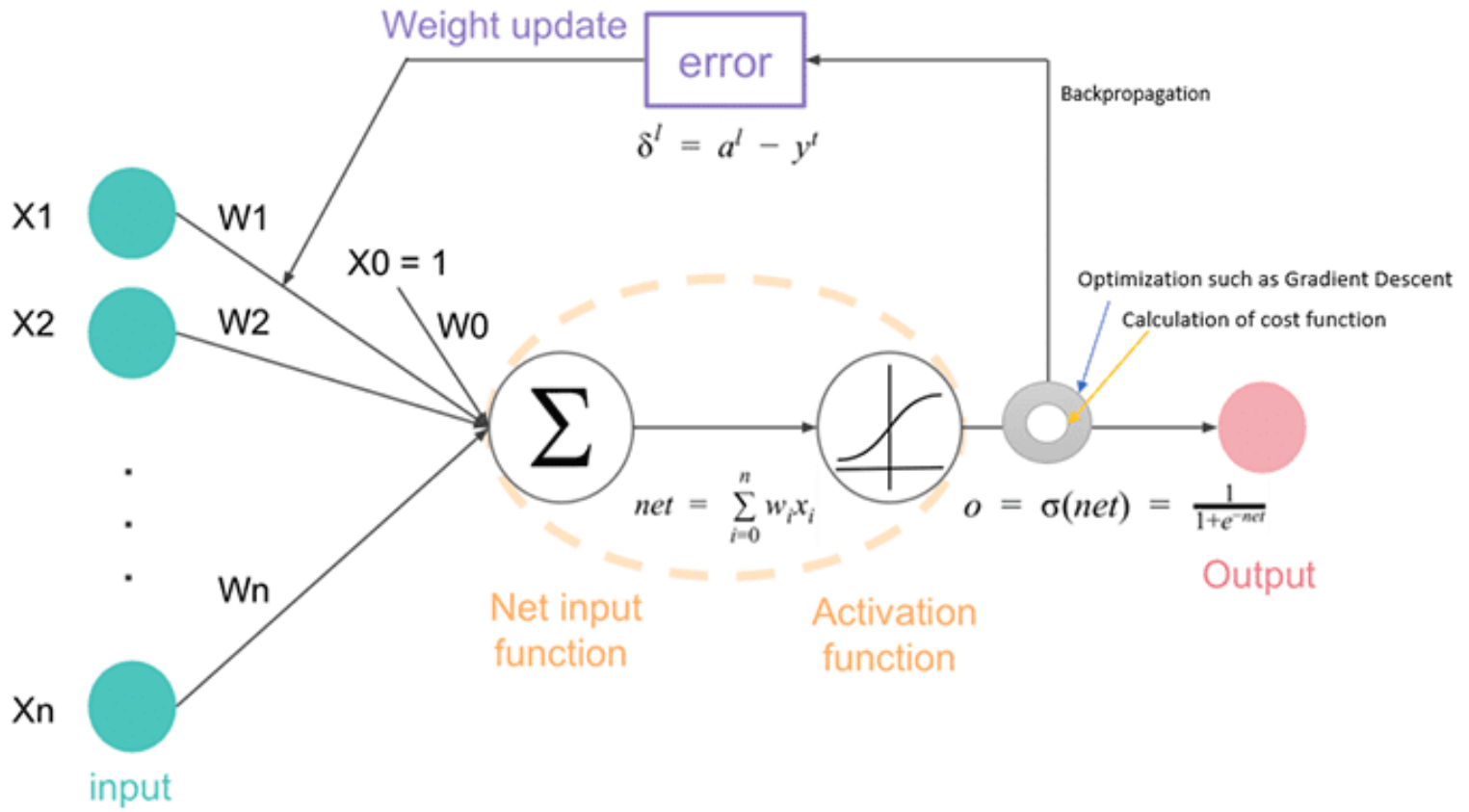



[a]

[a]针对 Loss 函数的参数寻优过程

# 梯度 (Gradient) 和 Hessian 矩阵 III

## 梯度下降 (Gradient Descent)




[a]

[a] How does gradient descent power neural networks and deep learning? https://pyimagesearch.com/2021/05/05/gradient-descent-algorithms-and-variations/

### 代码实现

Some Pytorch codes: [a]

```
...
MLPClassifier(max_iter=50, alpha=1e-4, solver='sgd', ....)
...
```

[a] https://www.cambridgespark.com/info/neural-networks-in-python

# 梯度 (Gradient) 和 Hessian 矩阵 I

## Gradient Descent or Newton's Method. Difference?

A vivid interpretation come from Quora[8].

**What's the difference between Gradient Descent and Newton's Method?**

Suppose you were on a hill and wanted to climb down to the lowest point in the valley below. There are two possible ways you could achieve this.
(1) The gradient descent way: You look around your feet and no farther than a few meters from your feet. You find the direction that slopes down the most and then walk a few meters in that direction. Then you stop and repeat the process until you can repeat no more. This will eventually lead you to the valley!
(2) The Newton way: You look far away. Specifically you look around in a way such that your line of sight is tangential to the mountain surface where you are. You find the point in your line of sight that is the lowest and using your awesome spiderman powers ...you jump to that point! Then you repeat the process until you can repeat no more!

# 梯度 (Gradient) 和 Hessian 矩阵 II

## Gradient Descent or Newton's Method. Difference?

So which one do you think will get you to the bottom of the valley faster?

Reference:

[8] https://www.quora.com/What-are-the-relationship-between-Gradient-Descent-and-Newtons-Method

# Outline

# 一般的运算技巧

我们已经关注到 Loss 函数的求解、梯度计算的列式，都是离不开矩阵层面的微分计算的。

一般而言，对矩阵函数 $f(A)$求微分计算 $\frac{\partial f(A)}{\partial A}$ 只需要掌握两个技巧。

一是记忆一些常见结论，并与传统意义下 $f(x)$的微分计算做结果联想。
二是按 $A$的分量进行相应计算，从头推理。

# 跟迹 (Trace) 有关的一些运算技巧

## 定理 (Frobenius norm and Trace)

*For each* $A \in \mathbb{R}^{m\times n}$, $||A||_F^2 = Tr(A^TA)$, *where* $Tr(\cdot)$ *refers to trace of the matrix.*

Proof: The proof is straightforward.

# 跟迹 (Trace) 有关的一些运算技巧

## 定理 (Trace(AB))

*For each* $A \in \mathbb{R}^{m\times n}$ *and* $B \in \mathbb{R}^{n\times m}$, $Tr(AB) = Tr(BA)$.

Proof: The proof is straightforward.

## 定理 (Derivative of Trace(AB))

*For each* $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times m}$, $\frac{\partial Tr(AB)}{\partial A} = B^T$.

Proof: The proof is straightforward.

## 定理 (Derivativeof Trace(AB))

*For each* $A, B \in \mathbb{R}^{m\times n}$, $\frac{\partial Tr(A^TB)}{\partial A} = B$.

Proof: From the above the theorem, we know that $\frac{\partial Tr(AB)}{\partial A^T} = B$. Substitute $A$ with $A^T$, the result follows.

# 跟迹 (Trace) 有关的一些运算技巧

**Exercise:**

[Assignments:] Assume $A, B \in \mathbb{R}^{n\times n}$, please show that

$$\frac{\partial Tr(ABA^T)}{\partial A} = AB^T + AB.$$

**Hints:**

Here, you'd solve the derivative of $ABA^T$, and the solution requires some knowledge of multiplicative differential. It is kind of easy, as well. Please just recall how do you solve the similar differential problem in $\frac{d}{dx}f(x)g(x)$ with a multiplicative form.

Won't be hard.

# 跟迹 (Trace) 有关的一些运算技巧

**Answer sheet:**

[Solve:] Assume $A, B \in \mathbb{R}^{n\times n}$, we have

$$\frac{\partial Tr(ABA^T)}{\partial A} = \frac{\partial Tr(ABA_c^T)}{\partial A} + \frac{\partial Tr(A_cBA^T)}{\partial A}$$

$$= (BA_c^T)^T + \frac{\partial Tr(A^TA_cB)}{\partial A}$$

$$= (BA^T)^T + AB = AB^T + AB.$$

Thus the result follows.

# Acknowledgements


Thank you!