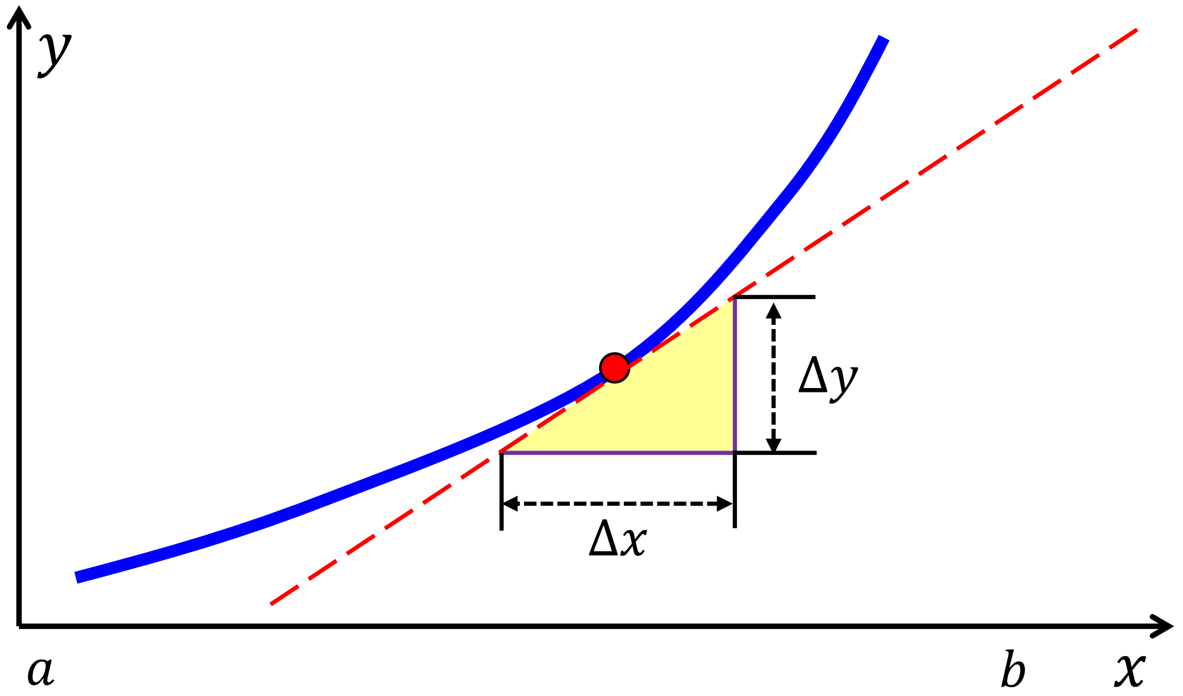
数学基础

李 波

导数与微分



$$f'(x) = \frac{df(x)}{dx} \approx \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

导数与微分

$$f'(x) = \frac{df(x)}{dx} \approx \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

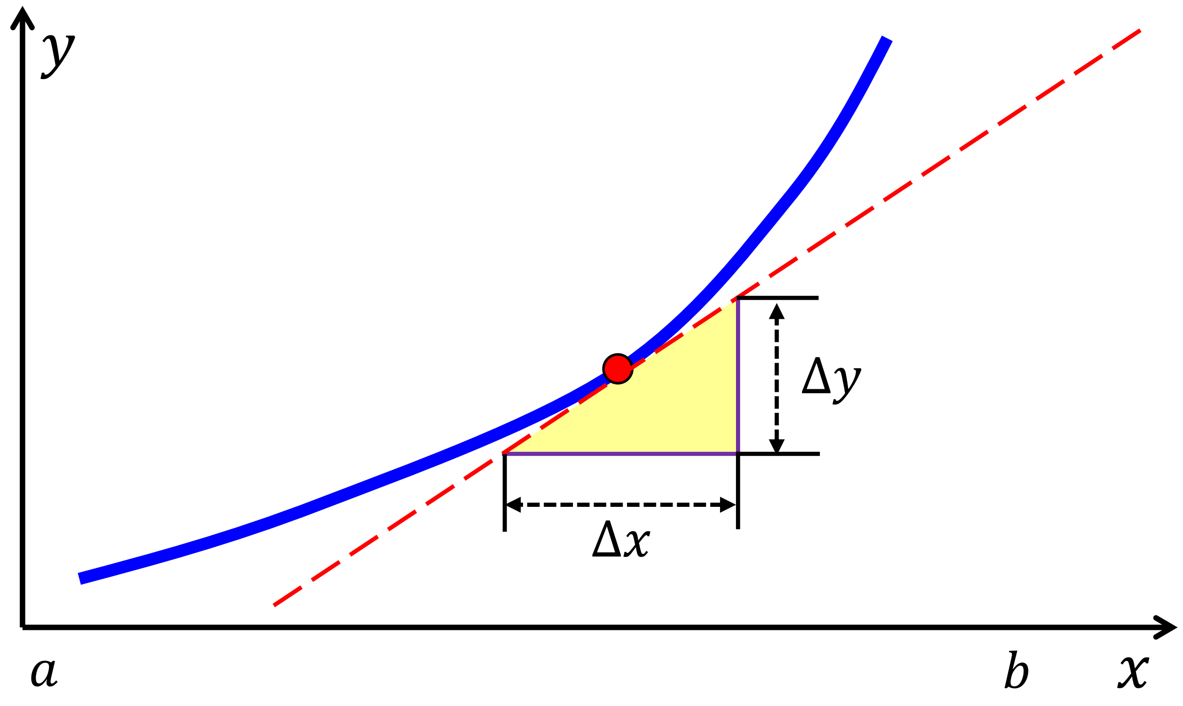
计算导数 $f(x) = x^2$

$$\lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{(x+\Delta x)^2 - x^2}{\Delta x}$$

$$= \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + \Delta x^2 - x^2}{\Delta x}$$

$$= \lim_{\Delta x \to 0} 2x + \Delta x$$

$$= 2x$$



导数与微分

$$f'(x) = \frac{df(x)}{dx} \approx \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

方程	导数
$1' = 0$	0
x^n	nx^{n-1}
e^x	e^x
$\ln(x)$	$1/x$
$\cos(x)$	$-\sin(x)$
$\sin(x)$	$\cos(x)$

导数与微分

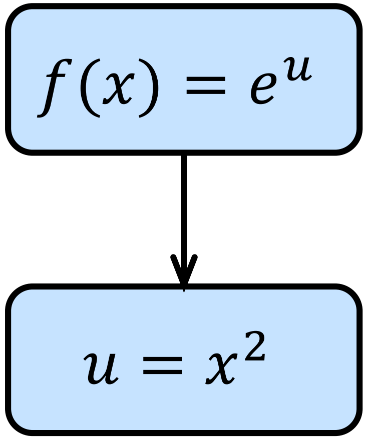
对于复合函数 $f(g(x))$，可以用**链式法则**计算导数

$$\frac{df(g(x))}{dx} = \frac{df(g(x))}{dg(x)}\frac{dg(x)}{dx}$$

例题：

$$f(x) = e^{x^2}$$

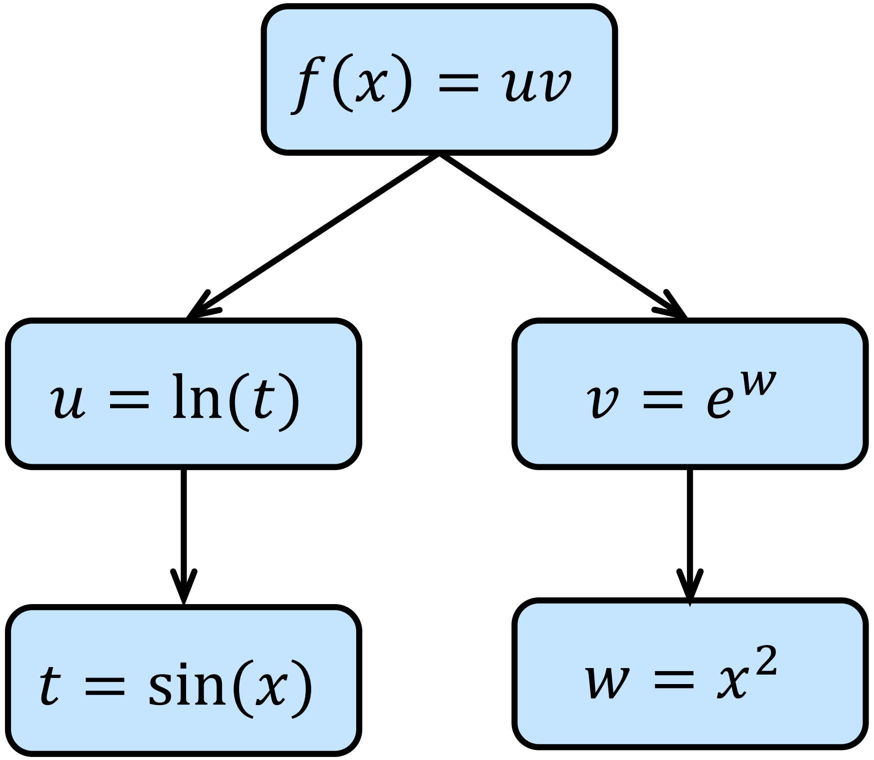
$$\frac{df(x)}{dx} = \frac{df(x)}{du}\frac{du}{dx} = e^{u}2x = 2xe^{x^2}$$



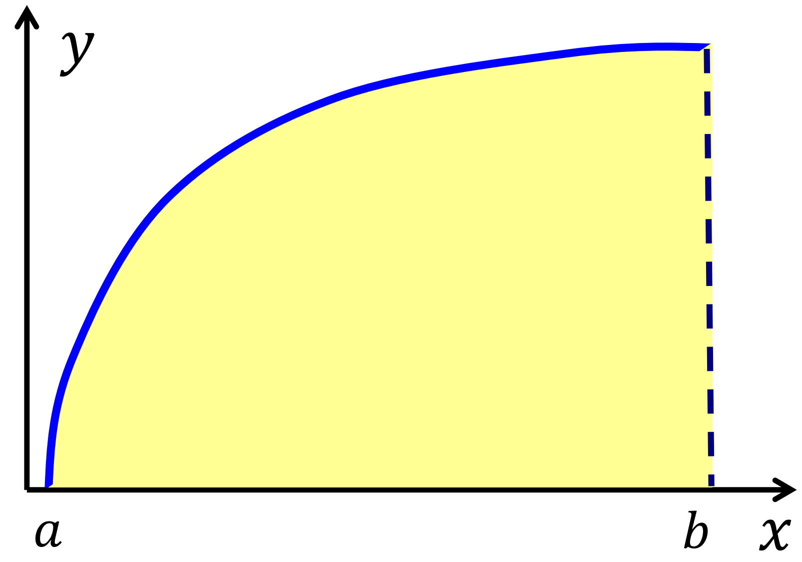
导数与微分

例题：计算导数 $f(x) = (\ln(\sin(x)))\, e^{x^2}$

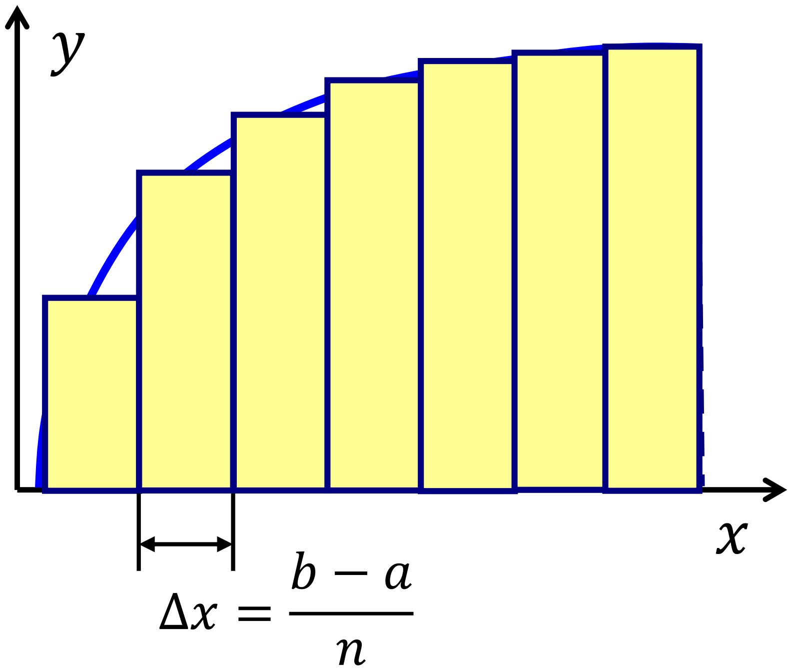
$$
\begin{aligned}
\frac{df(x)}{dx} &= \frac{df(x)}{du}\frac{du}{dx} + \frac{df(x)}{dv}\frac{dv}{dx} \\
&= v\frac{du}{dx} + u\frac{dv}{dx} \\
&= v\frac{du}{dt}\frac{dt}{dx} + u\frac{dv}{dw}\frac{dw}{dx} \\
&= v\frac{1}{t}\frac{dt}{dx} + ue^{w}\frac{dw}{dx} \\
&= v\frac{1}{t}\cos(x) + ue^{w}2x \\
&= e^{x^2}\frac{1}{\sin(x)}\cos(x) + (\ln(\sin(x))e^{x^2}2x
\end{aligned}
$$



积分



$\approx$



$$\int_a^b f(x)dx \approx \lim_{n\to\infty}\sum_{i=1}^{n}\frac{b-a}{n}f(x_i)$$

积分

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} \Delta x f(x_i)$$

微积分基本定理

1. $f(x)$ 为在 $[a,b]$区间上的连续函数。$g(x)=\int_a^x f(x)dx$ 在区间 $[a,b]$ 也连续而且在区间(a,b)上可微分，其导数为 $g'(x)=f(x)$或

$$\frac{d\int_a^x f(x)dx}{dx}=f(x)$$

2. $f(x)$ 为在 $[a,b]$区间上的连续函数，那么

$$\int_a^b f(x)dx=F(b)-F(a)$$

其中$F(x)$ 为$f(x)$的原函数，即$F'(x)=f(x)$。

积分

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} \Delta x f(x_i)$$

$$\int_a^b 1dx = x\Big|_a^b = b - a$$

$$\int_a^b e^x dx = e^x\Big|_a^b = e^b - e^a$$

$$\int_a^b x^n dx = \frac{x^{n+1}}{n+1}\Bigg|_a^b = \frac{b^{n+1}}{n+1} - \frac{a^{n+1}}{n+1}$$

$$\int_a^b \frac{1}{x} dx = \ln(x)\Big|_a^b = \ln(b) - \ln(a)$$

$$\int_a^b \cos(x)dx = \sin(\mathrm{x})\Big|_a^b = \sin(b) - \sin(a)$$

$$\int_a^b \sin(x)dx = -\cos(\mathrm{x})\Big|_a^b = -\cos(b) + cos(a)$$

问题

x is a standard Normally distributed random variable with pdf $f(x)$

$$\int_{-1}^{1} f(x)dx = ?$$

问题

$$f(x) = \int_0^x g(y)dy$$

$$\frac{df(x)}{dx} = ?$$

问题

$$f(x) = \int_0^x g(x, y)dy$$

$$\frac{df(x)}{dx} = ?$$

概率

- 令 A 为一个事件，比如
  - ✓ 明天下雨
  - ✓ 早晨起床晚了
  - ✓ 考试挂科
  - ✓ ……

- 事件A发生的概率表示为 $p(A)$。 $p(A)$ 有如下性质:

$$0 \le p(A) \le 1$$

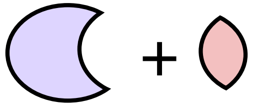
  - ✓ $p(\mathrm{A}) = 1$ 表示事件必定发生.
  - ✓ $p(\mathrm{A}) = 0$ 表示事件基本不会发生.
  - ✓ 实践中，往往统计事件的频率表示其概率.

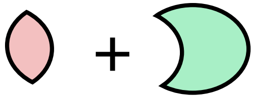
概率

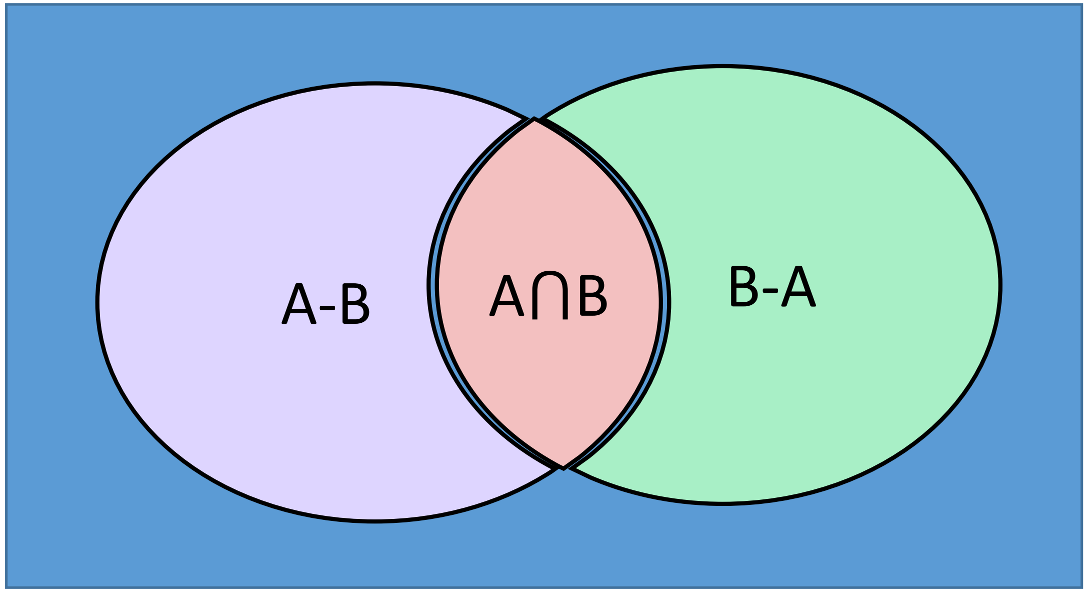
- 事件A和事件B同时发生的概率被称为**事件A和事件B的联合概率**，表示为
$$p(A\cap B)$$
- **事件A并事件B的概率**是指两个事件至少有一个发生的概率
$$p(A\cup B)$$
- **事件A与事件B互斥**，是指这两个事件不可能同时发生
$$p(A\cap B)=0 \quad \text{or} \quad p(A\cup B)=p(A)+p(B)$$
- **事件A与事件B相互独立**是指这两个事件没有任何关系。在概率上，有如下等式成立
$$p(A\cap B)=p(A)p(B)$$

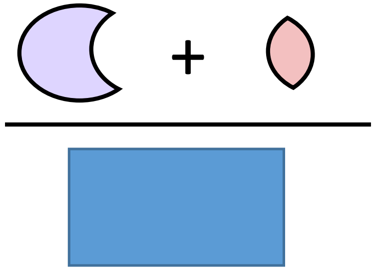
概率：文氏图

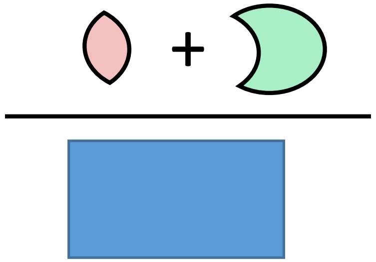
总事件 = 

事件A = 

事件B = 

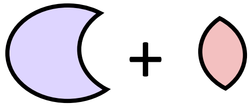


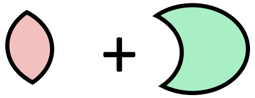
$p(A)$ = 

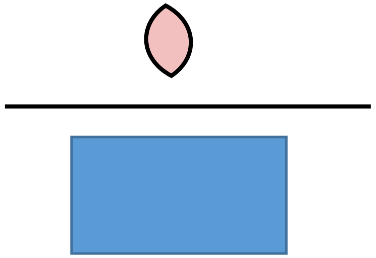
$p(B)$ = 

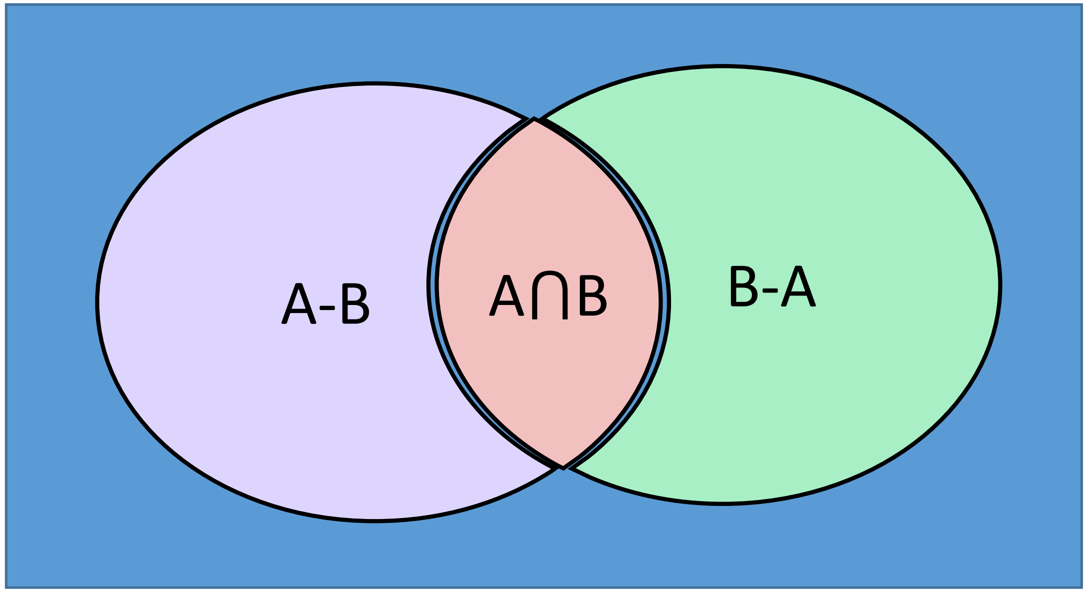
概率：文氏图

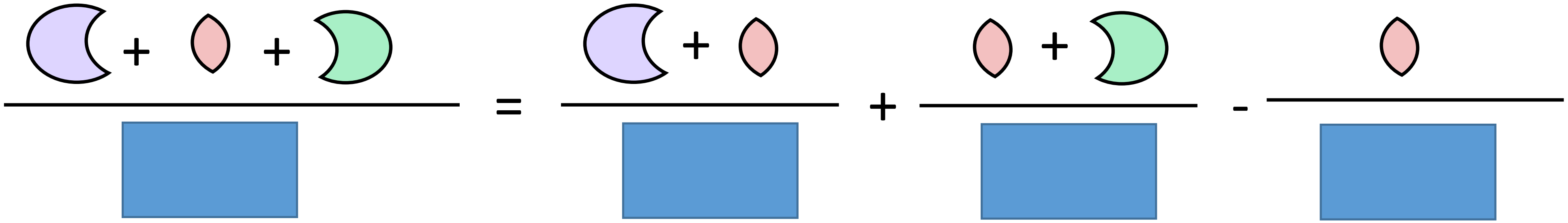
总事件 = 

事件 A = 

事件 B = 

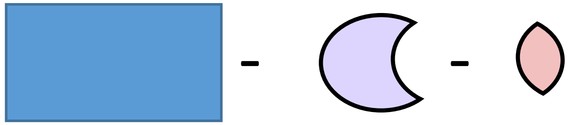
$p(A \cap B) =$ 

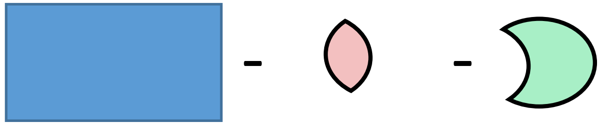


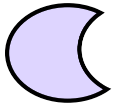
$p(A \cup B) =$ 

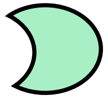
$= p(A) + p(B) - p(A \cap B)$

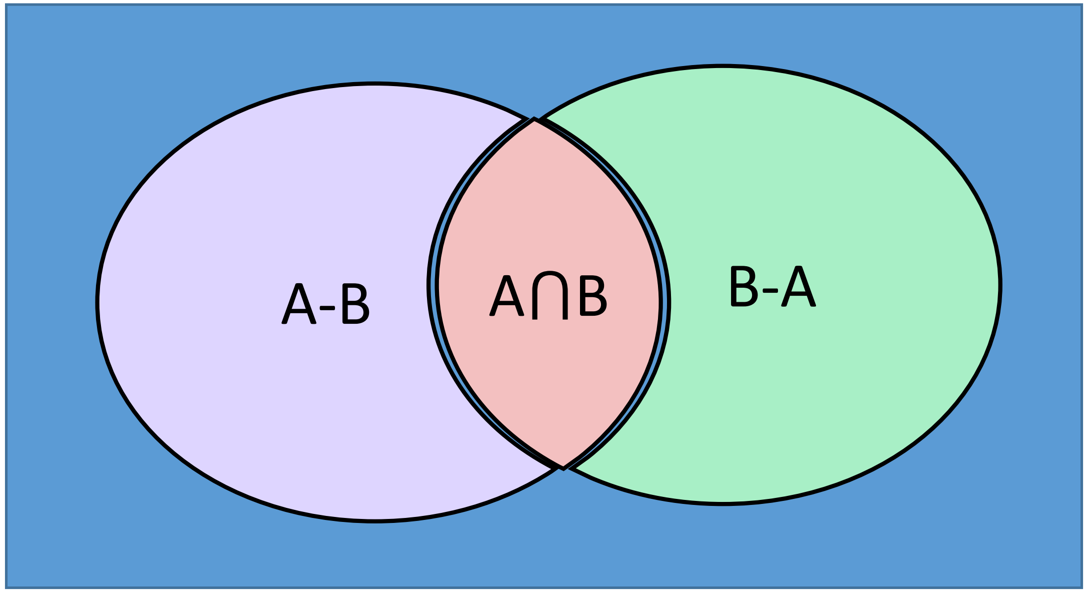
概率：文氏图

事件 $\bar{A}$ = 

事件 $\bar{B}$ = 

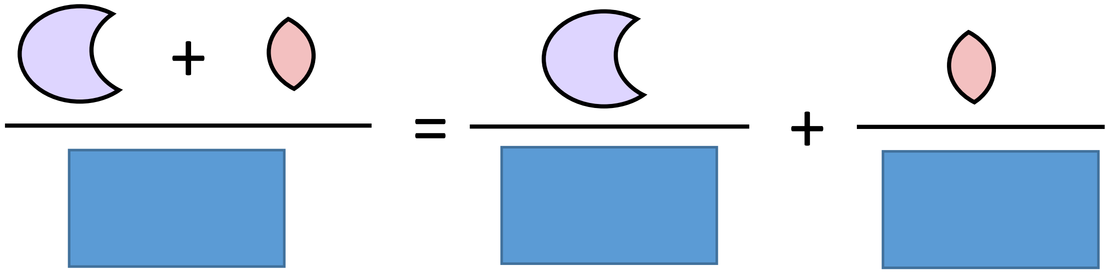
事件 $A \cap \bar{B}$ = 

事件 $B \cap \bar{A}$ = 

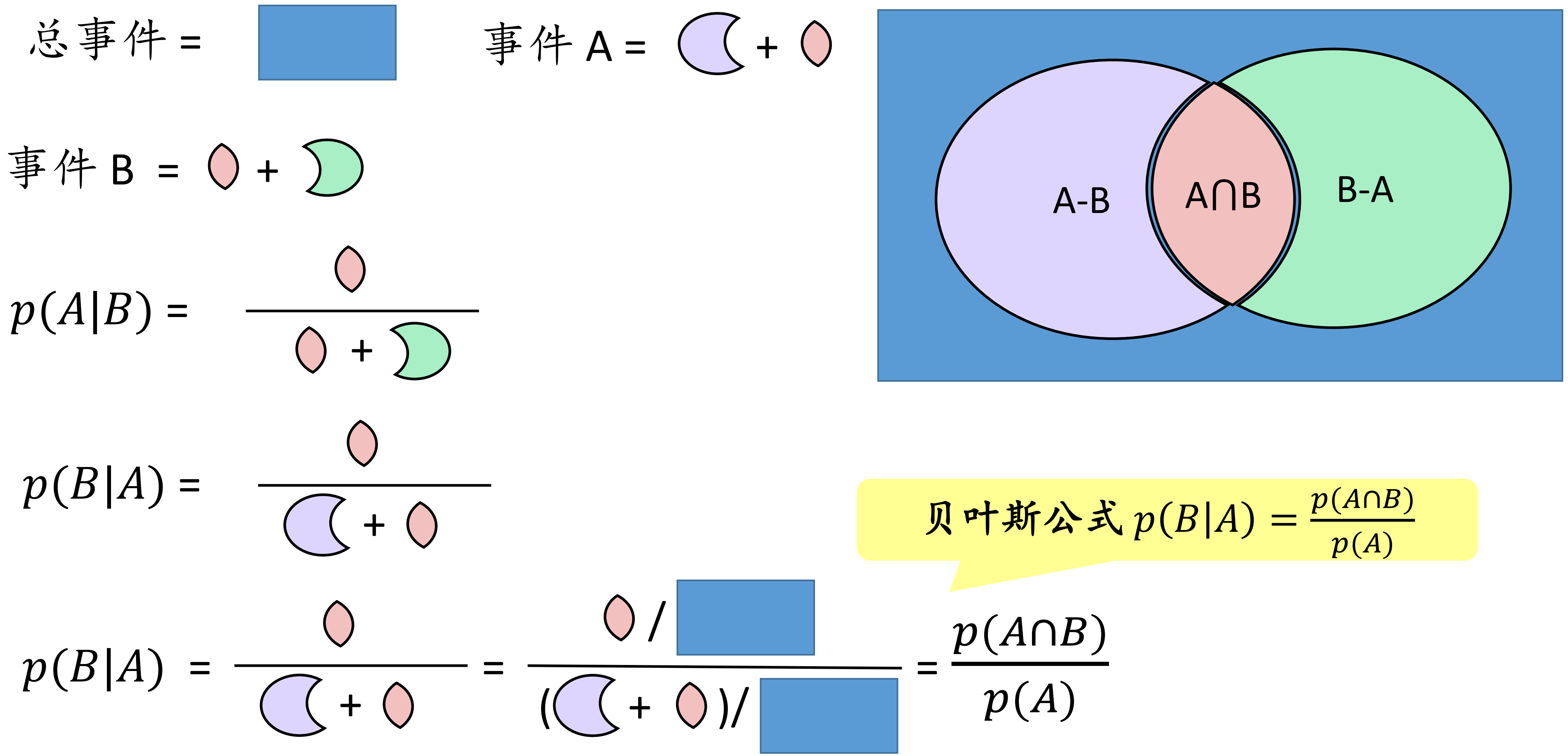


全概率公式

$$p(\mathrm{A}) = p(A \cap \bar{B}) + p(A \cap B)$$

$p(A)$ =  $= p(A \cap \bar{B}) + p(A \cap B)$

概率：文氏图


总事件 =
事件 A = +
事件 B = +
$p(A|B) =$
$p(B|A) =$
A-B
A⋂B
B-A
贝叶斯公式 $p(B|A) = \frac{p(A\cap B)}{p(A)}$
$p(B|A) = = \frac{p(A\cap B)}{p(A)}$

概率：文氏图

贝叶斯公式 $p(B|A) = \frac{p(A \cap B)}{p(A)}$

事件A和事件B相互独立，即这两个事件没有任何关系，有

$$p(B|A) = p(B) \quad \text{或} \quad p(A|B) = p(A)$$

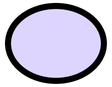
根据贝叶斯公式，可以得到

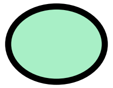
$$\underbrace{p(A \cap B) = p(A)p(B|A)} = p(A)p(B)$$

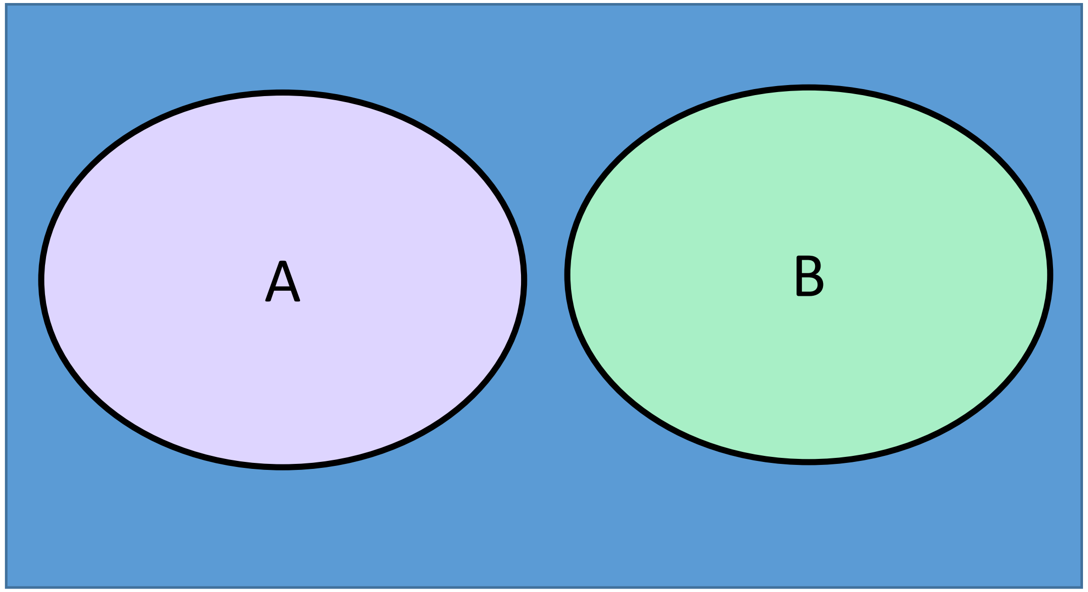
事件A和事件B同时发生等价于A先发生，然后B再发生。

概率：文氏图

总事件 = 

事件 A = 

事件 B = 



事件A和事件B没有任何交集，意味着这两个事件不会同时发生，有

$$p(A \cap B) = 0$$

概率：文氏图

文氏图

$p(A),p(B)$

$p(A\cap B)$

$p(A\cup B)$

贝叶斯公式

互斥

全概率公式

相互独立

概率

变量 vs 随机变量

- **小写斜体英文字母代表变量**。变量代表一个数值，虽然有时候我们不知道这个数值的大小，比如 $x=2$，$x+y=5$。
- **大写字母代表随机变量**。一般，随机变量的可能取值不唯一，每一个取值都对应一个概率。比如

$$X=\begin{cases}0 & p(X=0)=1/2\\ 1 & p(X=1)=1/4\\ 2 & p(X=2)=1/4\end{cases}\qquad \text{或} \qquad p(X=i)=\begin{cases}1/2 & i=0\\ 1/4 & i=1\\ 1/4 & i=2\end{cases}$$

概率

随机变量

- 离散型随机变量: 概率函数，积累概率函数
- 连续性随机变量: 概率密度函数，积累密度函数
- 均值
- 方差
- 协方差
- 相关系数

概率

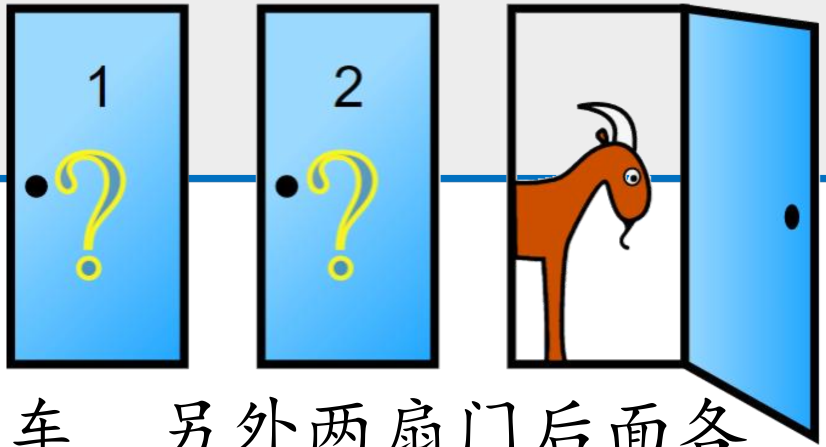
离散型随机变量

- 伯努利分布
- 二项式分布
- 泊松分布

连续性随机变量

- 指数分布
- 高斯分布
- 多变量高斯分布

趣味问题

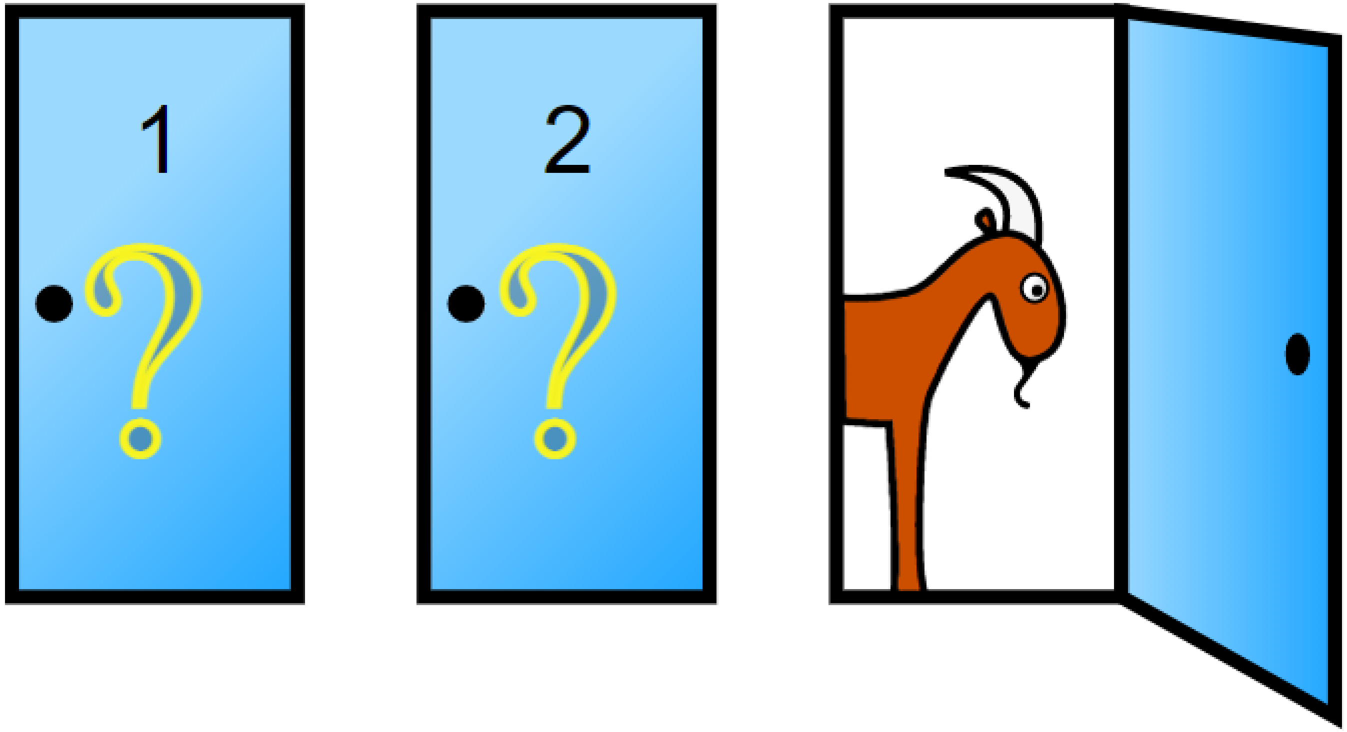


Monty Hall 问题

- 主持人在台上展示三扇门。一扇门后面有一辆汽车，另外两扇门后面各有一只羊。我们不知道那扇门后面有汽车。
- 主持人邀请我们选一扇门，如果这扇门后面是汽车，我们将赢走汽车；如果这扇门后面是羊，我们将赢走羊。
- 假设我们选了第一扇门，但还没有打开。
- 主持人为了节目的效果，打开了另外两扇门中的一个，假设第三扇门，展示给我们的是羊。
- 主持人问：**是坚守开第一扇门，还是换开第二扇门？**

注：主持人可以看到每扇门后面。主持人不是随机选择一扇门打开，而是故意打开后面有羊的门。

趣味问题



A 代表第一扇门后面是汽车

B 代表第二扇门后面是汽车

C 代表第三扇门后面是汽车

E 代表主持人打开了第三扇门，后面是羊

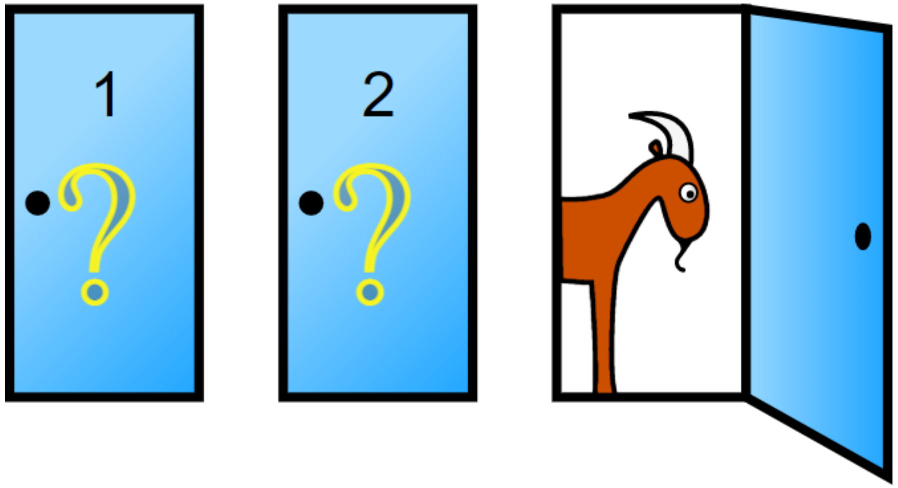
趣味问题

$$p(A) = p(B) = p(C) = \frac{1}{3}$$

$$\begin{aligned} p(E) &= p(E \cap A) + p(E \cap B) + p(E \cap C) \\ &= p(A)p(E|A) + p(B)p(E|B) + p(C)p(E|C) \\ &= \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 1 + \frac{1}{3} \times 0 \\ &= 1/2 \end{aligned}$$

$$p(A|E) = \frac{p(A \cup E)}{p(E)} = \frac{p(A)p(E|A)}{p(E)} = \frac{1/3 \times 1/2}{1/2} = \frac{1}{3}$$

$$p(B|E) = \frac{p(B \cup E)}{p(E)} = \frac{p(B)p(E|B)}{p(E)} = \frac{1/3 \times 1}{1/2} = \frac{2}{3}$$

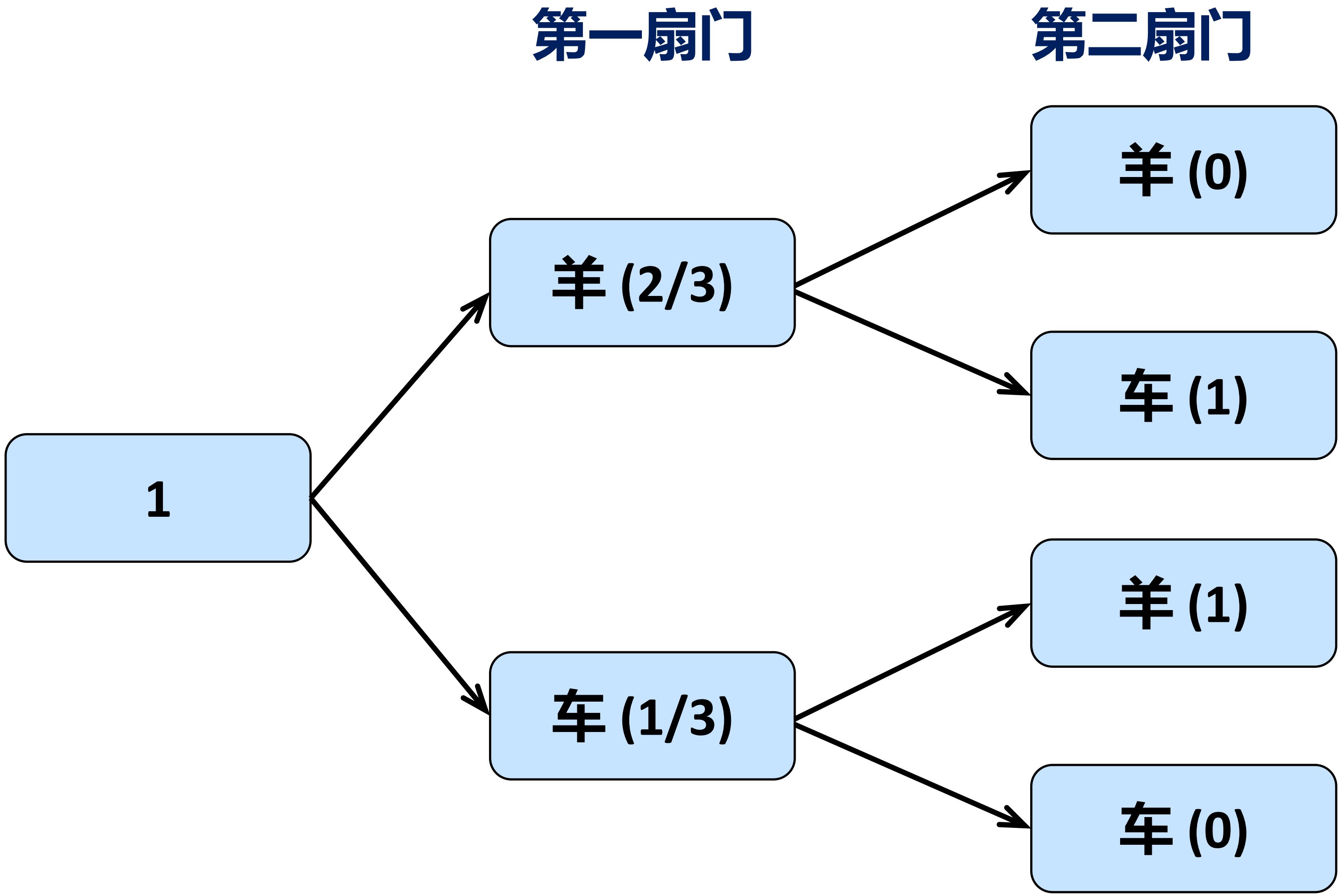


趣味问题

直观解决方案

第二扇门后面是羊的概率

$$\frac{2}{3} \times 1 + \frac{1}{3} \times 0 = \frac{2}{3}$$



趣味问题

三个囚犯问题

- 三个囚犯甲、乙、丙中的一个要被释放，但不知道谁会被释放。
- 囚犯甲问长官谁会被释放，长官说“我不能告诉你，但是我可以告诉你乙肯定不会被释放”。
- 囚犯甲听了之后非常开心，因为他认为自己被释放的概率从$\frac{1}{3}$ 增加到 $\frac{1}{2}$.
- **囚犯甲的想法对吗？**

向量与矩阵

一般用加黑小写英文字母表示向量。默认为向量为列向量。

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad \text{or} \quad \boldsymbol{x} = [x_1, x_2, \cdots x_m]^T$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad \text{or} \quad \boldsymbol{y} = [y_1, y_2, \cdots y_m]^T$$

向量与矩阵

向量乘法

内积:

$$\boldsymbol{x}^T\boldsymbol{y} = [x_1, x_2, \cdots x_m]\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = x_1y_1 + x_2y_2 + \cdots + x_my_m = \sum_{i=1}^{n} x_iy_i$$

外积: $\boldsymbol{x}\boldsymbol{y}^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}[y_1, y_2, \cdots y_m] = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_m \\ x_2y_1 & x_2y_2 & \cdots & x_2y_m \\ \vdots & \vdots & \vdots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_m \end{bmatrix}$

常规斜体英文字母表示变量.

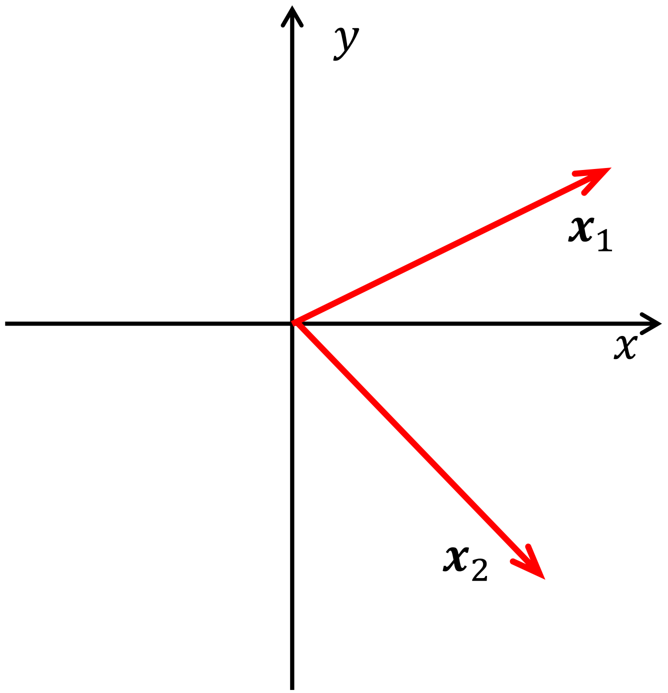
线性独立

对于一个向量集合一组向量$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n\}$，如果存在一组非全零系数$\{\omega_1, \omega_1, \cdots \omega_n\}$，使这些向量线性组合为零向量，即

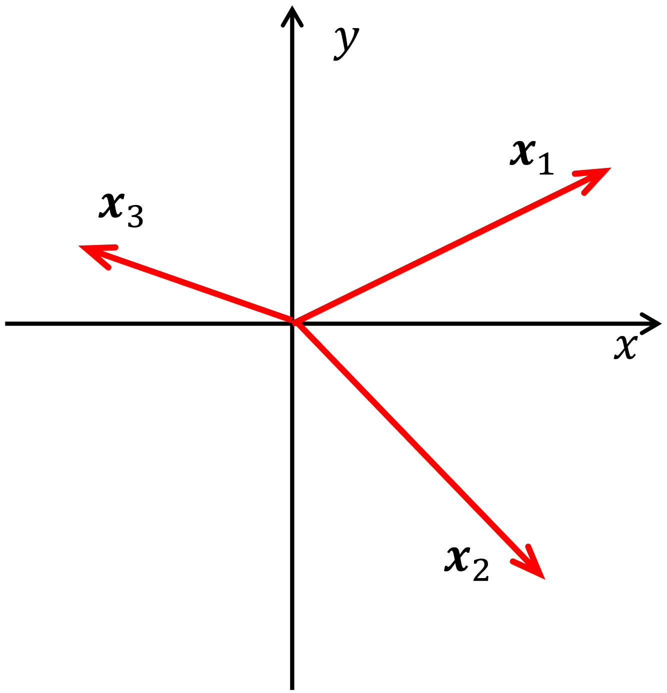
$$\omega_1\boldsymbol{x}_1 + \omega_2\boldsymbol{x}_2 + \cdots\cdots + \omega_n\boldsymbol{x}_n = \mathbf{0}$$

， 那么说这些向量线性独立。

是否线性独立?



y
x
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$



y
x
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_3$

向量与矩阵

向量空间(vector space)

在一个集合V中存在加法和标量乘法运算，而且下面8个性质成立，那么称集合V为向量空间

(1) $\boldsymbol{a}+\boldsymbol{b}=\boldsymbol{b}+\boldsymbol{a}$

(2) $(\boldsymbol{a}+\boldsymbol{b})+\boldsymbol{c}=\boldsymbol{a}+(\boldsymbol{b}+\boldsymbol{c})$

(3) $\boldsymbol{a}+\boldsymbol{0}=\boldsymbol{0}+\boldsymbol{a}=\boldsymbol{0}$

(4) $\boldsymbol{a}+(-\boldsymbol{a})=(-\boldsymbol{a})+\boldsymbol{a}=\boldsymbol{0}$

(5) $r(\boldsymbol{a}+\boldsymbol{b})=r\boldsymbol{a}+r\boldsymbol{b}$

(6) $(r+s)\boldsymbol{a}=r\boldsymbol{a}+r\boldsymbol{s}$

(7) $(rs)\boldsymbol{a}=r(s\boldsymbol{a})$

(8) $1\boldsymbol{a}=\boldsymbol{a}$

子空间(Subspace)

V为一个向量空间。W是V的一个子集。如果W满足一下两个性质，则称W是一个子空间。

- $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathrm{W}$,有$\boldsymbol{x}_1 + \boldsymbol{x}_2 \in \mathrm{W}$,
- 对任意标量c, 如果$\boldsymbol{x} \in W$，则$c\boldsymbol{x} \in \mathrm{W}$.

扩张空间(Span)

如果一个子空间V中的任何一个向量都可以由向量$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_n$线性组合表示，即

$$\forall \boldsymbol{x}\in V, \exists \omega_1,\omega_2\cdots\omega_n, \text{such that } \boldsymbol{x}=\omega_1\boldsymbol{x}_1+\omega_2\boldsymbol{x}_2+\cdots\cdots+\omega_n\boldsymbol{x}_n$$

，那么说空间V是向量$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_n$的扩张子空间，或者说向量$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_n$扩张成V。

The vectors $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_n$ in a vector space are said to span V if every vector in V is a linear combination of $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_n$. That is

$$\forall \boldsymbol{x}\in V, \exists \omega_1,\omega_2\cdots\omega_n, \text{such that } \boldsymbol{x}=\omega_1\boldsymbol{x}_1+\omega_2\boldsymbol{x}_2+\cdots\cdots+\omega_n\boldsymbol{x}_n$$

If $S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_n\}$, then we say that S spans V or V is spanned by S.

基底(Basis)

对于向量集合$S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots \boldsymbol{x}_n\}$和一个子空间V，如果以下两个性质成立，则称向量集合S是子空间V的基底：

- S的扩展空间为V，
- S中的向量都线性独立。

子空间V的维度为子空间V基底向量的个数。

The set of vectors $S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots \boldsymbol{x}_n\}$, in a vector space V is called a basis for V if

- S spans V
- Vectors in S are linearly independent

Standard Basis

向量与矩阵

向量乘法

内积:

$$\boldsymbol{x}^T\boldsymbol{y} = [x_1, x_2, \cdots x_m]\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = x_1y_1 + x_2y_2 + \cdots + x_my_m = \sum_{i=1}^{n} x_iy_i$$

外积或者tensor成积:

$$\boldsymbol{x}\boldsymbol{y}^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}[y_1, y_2, \cdots y_m] = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_m \\ x_2y_1 & x_2y_2 & \cdots & x_2y_m \\ \vdots & \vdots & \vdots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_m \end{bmatrix}$$

斜体小写字母代表变量。

向量与矩阵

黑体大写英文字母表示矩阵。一个 $m \times n$ 的矩阵为

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}$$

矩阵的秩定义为

- 线性不相关列的最大数目
- 或线性不相关行的最大数目

向量与矩阵

一个$m \times n$ 矩阵和一个 $n \times 1$ 向量相乘，得到一个 $m \times 1$ 的向量

$$\boldsymbol{Ax} = \underbrace{\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}}_{m \times n} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} A_{11}x_1 + A_{21}x_2 + \cdots + A_{1n}x_n \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n \\ \vdots \\ A_{m1}x_1 + A_{m1}x_2 + \cdots + A_{mn}x_n \end{bmatrix}}_{m \times 1}$$

向量与矩阵

一个$m\times n$ 矩阵和一个 $n\times 1$ 向量相乘，得到一个 $m\times 1$ 的向量

$$\boldsymbol{A}\boldsymbol{x}=\underbrace{\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}}_{m\times n}\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{n\times 1}=\underbrace{\begin{bmatrix} A_{11}x_1+A_{21}x_2+\cdots+A_{1n}x_n \\ A_{21}x_1+A_{22}x_2+\cdots+A_{2n}x_n \\ \vdots \\ A_{m1}x_1+A_{m1}x_2+\cdots+A_{mn}x_n \end{bmatrix}}_{m\times 1}$$

向量与矩阵

一个$m \times n$ 矩阵和一个 $n \times 1$ 向量相乘，得到一个 $m \times 1$ 的向量

$$\boldsymbol{A}\boldsymbol{x} = [\boldsymbol{A}_{:1} \quad \boldsymbol{A}_{:2} \quad \cdots \quad \boldsymbol{A}_{:n}] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \underbrace{x_1\boldsymbol{A}_{:1} + x_2\boldsymbol{A}_{:2} + \cdots + x_n\boldsymbol{A}_{:n}}$$

, 其中 $\boldsymbol{A}_{:i}$ 是矩阵A的第i列。

所有列的加权线性组合

向量与矩阵

$$\begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 1\times 5 + 2\times 6 \\ 4\times 5 + 3\times 6 \end{pmatrix} = \begin{pmatrix} 17 \\ 38 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} 5 \\ 6 \end{pmatrix} = 5\times\begin{pmatrix} 1 \\ 4 \end{pmatrix} + 6\times\begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 5 \\ 20 \end{pmatrix} + \begin{pmatrix} 12 \\ 18 \end{pmatrix} = \begin{pmatrix} 17 \\ 38 \end{pmatrix}$$

向量与矩阵

一个 $m \times n$ 矩阵和一个$n \times k$矩阵相乘得到一个$m \times k$矩阵

矩阵右乘一个向量

$$\boldsymbol{AB} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1k} \\ B_{21} & B_{22} & \cdots & B_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mk} \end{bmatrix} = [\boldsymbol{AB}_{:1} \quad \boldsymbol{AB}_{:2} \quad \cdots \quad \boldsymbol{AB}_{:k}]$$

$$\boldsymbol{AB} = \boldsymbol{A}_{:1}\boldsymbol{B}_{1:} + \boldsymbol{A}_{:2}\boldsymbol{B}_{2:} + \cdots + \boldsymbol{A}_{:n}\boldsymbol{B}_{n:} = \sum_{i=1}^{n} \boldsymbol{A}_{:i}\boldsymbol{B}_{i:}$$

一个列向量乘以一个行向量

，其中 $\boldsymbol{B}_{i:}$ 是矩阵$\boldsymbol{B}$的第i行。

向量与矩阵

$$\begin{pmatrix}1 & 2\\4 & 3\end{pmatrix}\begin{pmatrix}5 & 6\\7 & 8\end{pmatrix}=\begin{pmatrix}1\times5+2\times7 & 1\times6+2\times8\\4\times5+3\times7 & 4\times6+3\times8\end{pmatrix}=\begin{pmatrix}19 & 22\\41 & 48\end{pmatrix}$$

$$\begin{aligned}\begin{pmatrix}1 & 2\\4 & 3\end{pmatrix}\begin{pmatrix}5 & 6\\7 & 8\end{pmatrix}&=\begin{pmatrix}1\\4\end{pmatrix}\begin{pmatrix}5 & 6\end{pmatrix}+\begin{pmatrix}2\\3\end{pmatrix}\begin{pmatrix}7 & 8\end{pmatrix}\\&=\begin{pmatrix}5 & 6\\20 & 24\end{pmatrix}+\begin{pmatrix}14 & 16\\21 & 24\end{pmatrix}\\&=\begin{pmatrix}19 & 22\\41 & 48\end{pmatrix}\end{aligned}$$

哈德玛德乘积(Hadamard Product)

一个$m \times n$矩阵和一个$m \times n$矩阵的哈德玛德乘积为

$$\boldsymbol{A} \odot \boldsymbol{B} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \odot \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mn} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11}B_{11} & A_{12}B_{12} & \cdots & A_{1n}B_{1n} \\ A_{21}B_{21} & A_{22}B_{22} & \cdots & A_{2n}B_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1}B_{m1} & A_{m2}B_{m2} & \cdots & A_{mn}B_{mn} \end{bmatrix}$$

向量与矩阵

Kronecker乘积(Kronecker Product)

一个$m \times n$ 矩阵与一个 $p \times q$矩阵的Kronecker乘积为

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \otimes \boldsymbol{B}$$

$$= \begin{bmatrix} A_{11}\boldsymbol{B} & A_{12}\boldsymbol{B} & \cdots & A_{1n}\boldsymbol{B} \\ A_{21}\boldsymbol{B} & A_{22}\boldsymbol{B} & \cdots & A_{2n}\boldsymbol{B} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1}\boldsymbol{B} & A_{m2}\boldsymbol{B} & \cdots & A_{mn}\boldsymbol{B} \end{bmatrix}$$

向量与矩阵

$$\begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix} \odot \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \times 5 & 2 \times 6 \\ 4 \times 7 & 3 \times 8 \end{pmatrix} = \begin{pmatrix} 5 & 12 \\ 28 & 24 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix} \otimes \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \times \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} & 2 \times \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \\ 4 \times \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} & 3 \times \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 5 & 6 & 10 & 12 \\ 7 & 8 & 14 & 16 \\ 20 & 24 & 15 & 18 \\ 28 & 32 & 21 & 24 \end{pmatrix}$$

向量与矩阵

列空间：由矩阵A的列扩张成的空间。
Column space: space spanned by columns of the matrix $\boldsymbol{A}$.

行空间：由矩阵A的行扩张成的空间。
Row space: space spanned by rows of the matrix $\boldsymbol{A}$.

秩：矩阵行空间维度或者列空间维度。
Rank: dimension of column space or row space.

- ✓ $\mathrm{rank}(\boldsymbol{A}) \leq \min(m, n)$
- ✓ $\mathrm{rank}(\boldsymbol{AB}) = \min(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B}))$

向量与矩阵

零空间：方程$\boldsymbol{Ax} = \boldsymbol{0}$ 的解扩张成的空间。 Nullity为零空间的维度。
Null space: space spanned by solutions to $\boldsymbol{Ax} = \boldsymbol{0}$. Nullity is the dimension of the null space

秩-Nullity定理：秩 + Nullity = 矩阵列的个数。
Rank-Nullity theorem: rank + nullity = number of columns

特征值与特征向量

理解矩阵右乘向量运算中矩阵的作用

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

矩阵作为向量的转换操作，将一个向量转换为另一个向量。转换包括

- 改变向量方向
- 改变向量长度
- 改变向量维度

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$为一个线性变换。什么是线性变换?

特征值与特征向量

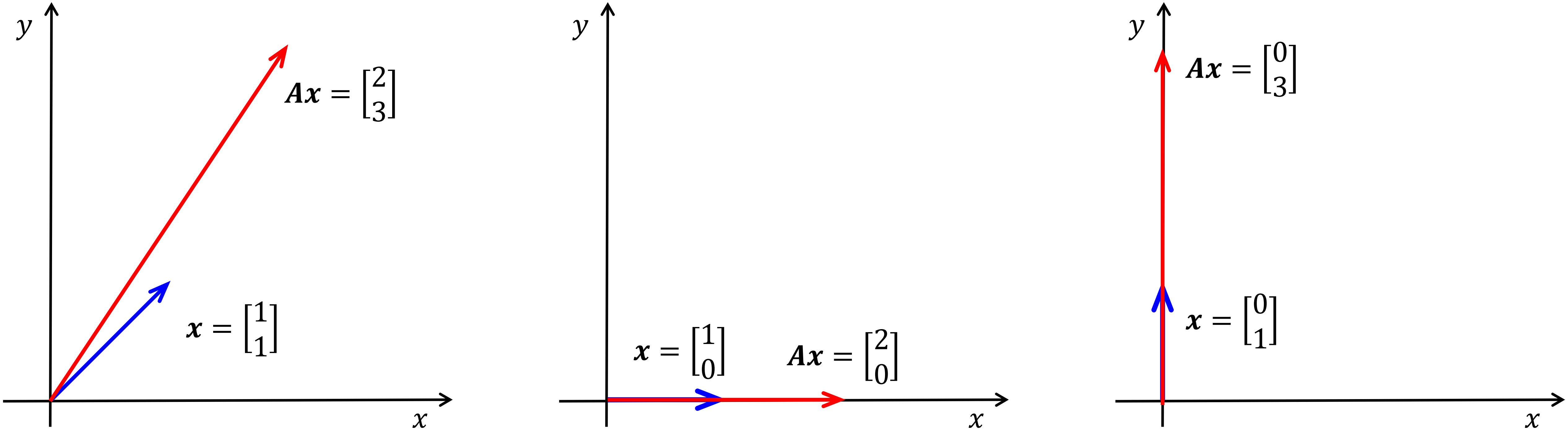
令$\boldsymbol{A}$为一个方阵。存在一个向量，矩阵A右乘这个向量只改变向量的长度，却保持向量方向不变或反向，即

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

这样的向量被称为矩阵$\boldsymbol{A}$的特征向量，对应的λ被称作特征向量对应的特征值。

特征值与特征向量

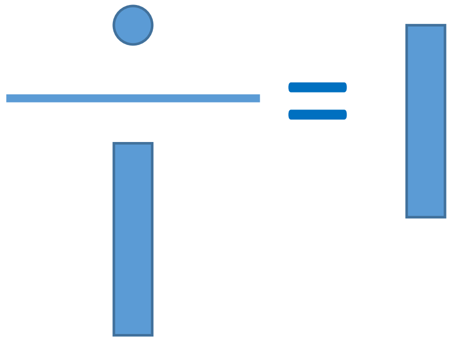
$A=\begin{bmatrix}2 & 0\\0 & 3\end{bmatrix}$，特征向量为 $\begin{bmatrix}0\\1\end{bmatrix}$，$\begin{bmatrix}1\\0\end{bmatrix}$，对应的特征值为2和3。



标量对向量导数

y是n维向量$\boldsymbol{x} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$一个标量函数。$y$对$\boldsymbol{x}$的导数是一个如下的列向量

$$\frac{\partial y}{\partial \boldsymbol{x}} = \begin{bmatrix} \dfrac{\partial y}{\partial x_1} \\ \dfrac{\partial y}{\partial x_2} \\ \vdots \\ \dfrac{\partial y}{\partial x_n} \end{bmatrix}$$

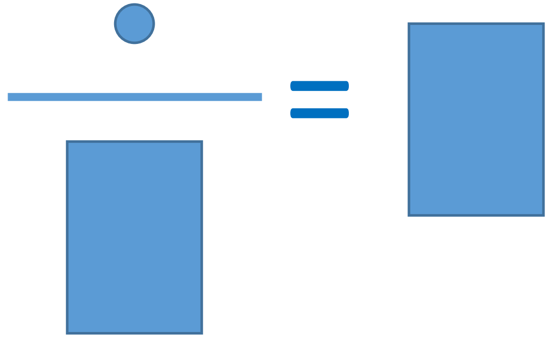


这个导数也被称作梯度。

标量对矩阵的导数

y是一个m行n列矩阵$\boldsymbol{X}$的标量函数。y对$\boldsymbol{X}$的导数也是一个如下的m行n列的矩阵，

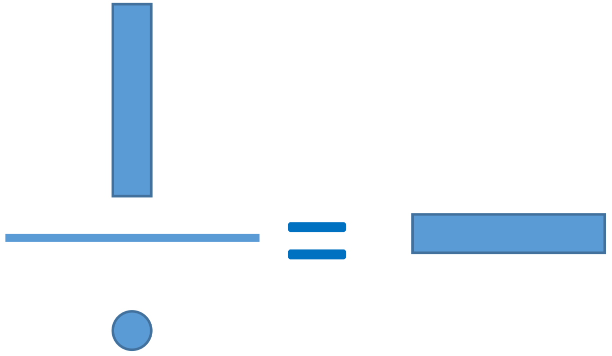
$$\frac{\partial y}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial y}{\partial X_{11}} & \frac{\partial y}{\partial X_{12}} & \cdots & \frac{\partial y}{\partial X_{1n}} \\ \frac{\partial y}{\partial X_{21}} & \frac{\partial y}{\partial X_{22}} & \cdots & \frac{\partial y}{\partial X_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial X_{m1}} & \frac{\partial y}{\partial X_{m2}} & \cdots & \frac{\partial y}{\partial X_{mn}} \end{bmatrix}$$



向量对标量的导数

$\boldsymbol{y} = [y_1 \quad y_2 \quad \cdots \quad y_m]^T$是一个$m$维的向量，而且每个元素都是标量$x$的函数。$\boldsymbol{y}$对$x$的导数是如下$m$维的行向量

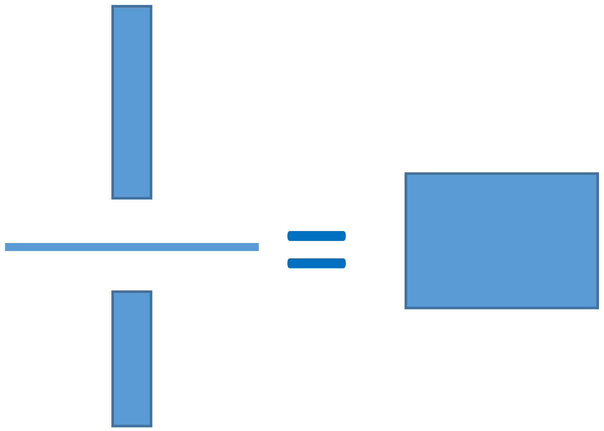
$$\frac{\partial \boldsymbol{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix}$$



向量对向量导数

m维的向量$\boldsymbol{y} = [y_1 \quad y_2 \quad \cdots \quad y_m]^T$是$n$维向量$\boldsymbol{x} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$的函数。$\boldsymbol{y}$对$\boldsymbol{x}$的导数是如下$n$行$m$列的矩阵

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \dfrac{\partial \boldsymbol{y}}{\partial x_1} \\ \dfrac{\partial \boldsymbol{y}}{\partial x_2} \\ \vdots \\ \dfrac{\partial \boldsymbol{y}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \dfrac{\partial y_2}{\partial x_1} & \cdots & \dfrac{\partial y_m}{\partial x_1} \\ \dfrac{\partial y_1}{\partial x_2} & \dfrac{\partial y_2}{\partial x_2} & \cdots & \dfrac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial y_1}{\partial x_n} & \dfrac{\partial y_2}{\partial x_n} & \cdots & \dfrac{\partial y_m}{\partial x_n} \end{bmatrix}$$



，等号右侧矩阵也被称为雅克比矩阵(Jacobian matrix).

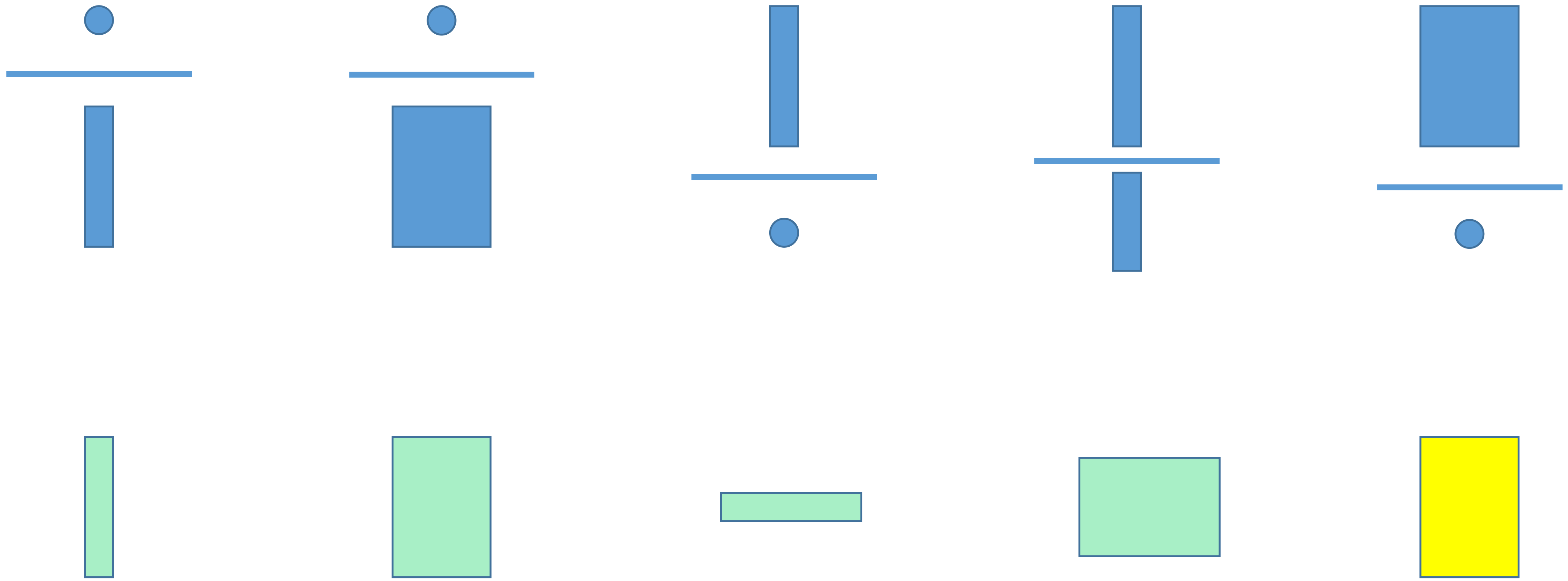
导数

方程 变量	标量	向量	矩阵
标量	$\frac{\partial y}{\partial x}$	$\frac{\partial \boldsymbol{y}}{\partial x}$	$\frac{\partial Y}{\partial x}$
向量	$\frac{\partial y}{\partial \boldsymbol{x}}$	$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$	
矩阵	$\frac{\partial y}{\partial \boldsymbol{X}}$		

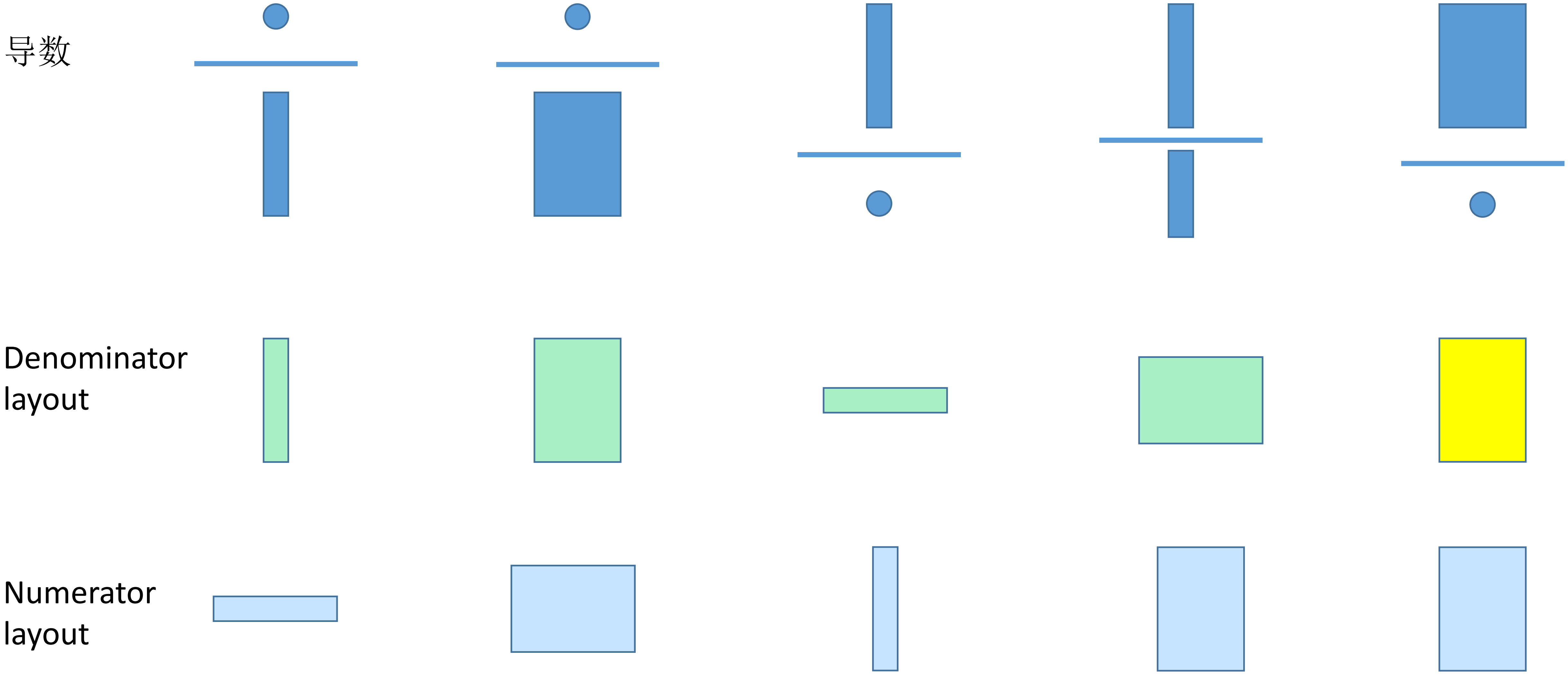
导数

导数

结果布局



导数



H. Lutkepohl, Handbook of Matrices, John Wiley & Sons, 1996.

链式法则

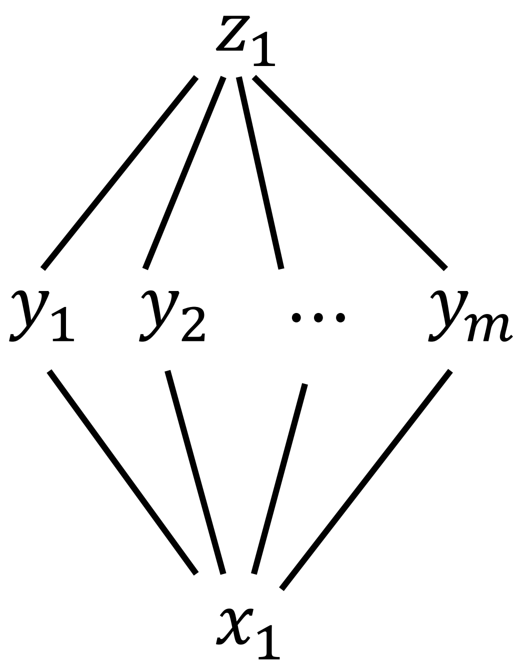
$$\boldsymbol{z}=[z_1 \quad z_2 \quad \cdots \quad z_k]^T$$
$$\boldsymbol{y}=[y_1 \quad y_2 \quad \cdots \quad y_m]^T$$
$$\boldsymbol{x}=[x_1 \quad x_2 \quad \cdots \quad x_n]^T,$$
$$\boldsymbol{z}=f(\boldsymbol{y}),\boldsymbol{y}=g(\boldsymbol{x})$$

$\mathbf{z}$对$\boldsymbol{x}$的导数是一个n行k列的矩阵

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}=\begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_k}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_k}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_1}{\partial x_n} & \frac{\partial z_2}{\partial x_n} & \cdots & \frac{\partial z_k}{\partial x_n} \end{bmatrix}$$

链式法则

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_k}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_k}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_1}{\partial x_n} & \frac{\partial z_2}{\partial x_n} & \cdots & \frac{\partial z_k}{\partial x_n} \end{bmatrix}$$



$$\frac{\partial z_1}{\partial x_1} = \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial z_1}{\partial y_2}\frac{\partial y_2}{\partial x_1} + \cdots + \frac{\partial z_1}{\partial y_m}\frac{\partial y_m}{\partial x_1} = \frac{\partial z_1}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_1}$$

链式法则

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_2}{\partial x_1} & \cdots & \frac{\partial z_k}{\partial x_1} \\ \frac{\partial z_1}{\partial x_2} & \frac{\partial z_2}{\partial x_2} & \cdots & \frac{\partial z_k}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_1}{\partial x_n} & \frac{\partial z_2}{\partial x_n} & \cdots & \frac{\partial z_k}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial z_1}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_1} & \frac{\partial z_2}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_1} & \cdots & \frac{\partial z_k}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_1} \\ \frac{\partial z_1}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_2} & \frac{\partial z_2}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_2} & \cdots & \frac{\partial z_k}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_1}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_n} & \frac{\partial z_2}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_n} & \cdots & \frac{\partial z_k}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_k} \end{bmatrix}$$

$$\frac{\partial z_1}{\partial x_1} = \frac{\partial z_1}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial z_1}{\partial y_2}\frac{\partial y_2}{\partial x_1} + \cdots + \frac{\partial z_1}{\partial y_m}\frac{\partial y_m}{\partial x_1} = \frac{\partial z_1}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_1}$$

链式法则

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial z_1}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_1} & \frac{\partial z_2}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_1} & \cdots & \frac{\partial z_k}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_1} \\ \frac{\partial z_1}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_2} & \frac{\partial z_2}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_2} & \cdots & \frac{\partial z_k}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_1}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_n} & \frac{\partial z_2}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_n} & \cdots & \frac{\partial z_k}{\partial \boldsymbol{y}}\frac{\partial \boldsymbol{y}}{\partial x_k} \end{bmatrix} = \begin{bmatrix} \frac{\partial \boldsymbol{y}}{\partial x_1} \\ \frac{\partial \boldsymbol{y}}{\partial x_2} \\ \vdots \\ \frac{\partial \boldsymbol{y}}{\partial x_n} \end{bmatrix} \begin{bmatrix} \frac{\partial z_1}{\partial \boldsymbol{y}} & \frac{\partial z_1}{\partial \boldsymbol{y}} & \cdots & \frac{\partial z_k}{\partial \boldsymbol{y}} \end{bmatrix} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}}$$

链式法则

对于标量变量和标量函数，有

$$\frac{\partial z}{\partial x} = \frac{\partial y}{\partial x}\frac{\partial z}{\partial y} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

对于向量变量和向量函数，有

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}}$$

如果$\boldsymbol{x}$是$\boldsymbol{w}$的函数，有

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{w}} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{w}}\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}}$$

涉及向量与矩阵的导数运算

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^T$$

$$\frac{\partial \boldsymbol{x}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$$

$$\frac{\partial \boldsymbol{a}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^T\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{\partial \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^T\boldsymbol{x}$$

$$\frac{\partial \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{A}} = \boldsymbol{y}\boldsymbol{x}^T$$

“The Matrix Cookbook” from Petersen and Pedersen (2012) gives an even larger list.

优化

无约束最小化问题为

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

- $\boldsymbol{x}$为变量.
- $f(\boldsymbol{x})$ 被称作目标函数.
- 最大化问题一般转换为最小化问题

$$\max_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \Rightarrow \quad \min_{\boldsymbol{x}} -f(\boldsymbol{x})$$

优化

有约束条件的最小化问题

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{s.t.}\quad c_i(\boldsymbol{x}) \le 0,\quad i = 1,2,\cdots k$$

- $c_i(\boldsymbol{x}) \le 0$ 为约束条件。
- 满足所有约束约束条件的$\boldsymbol{x}$构成一个可行域（feasible region）。
- 在可行域内，找到一个 $\boldsymbol{x}$ 最小化目标函数。

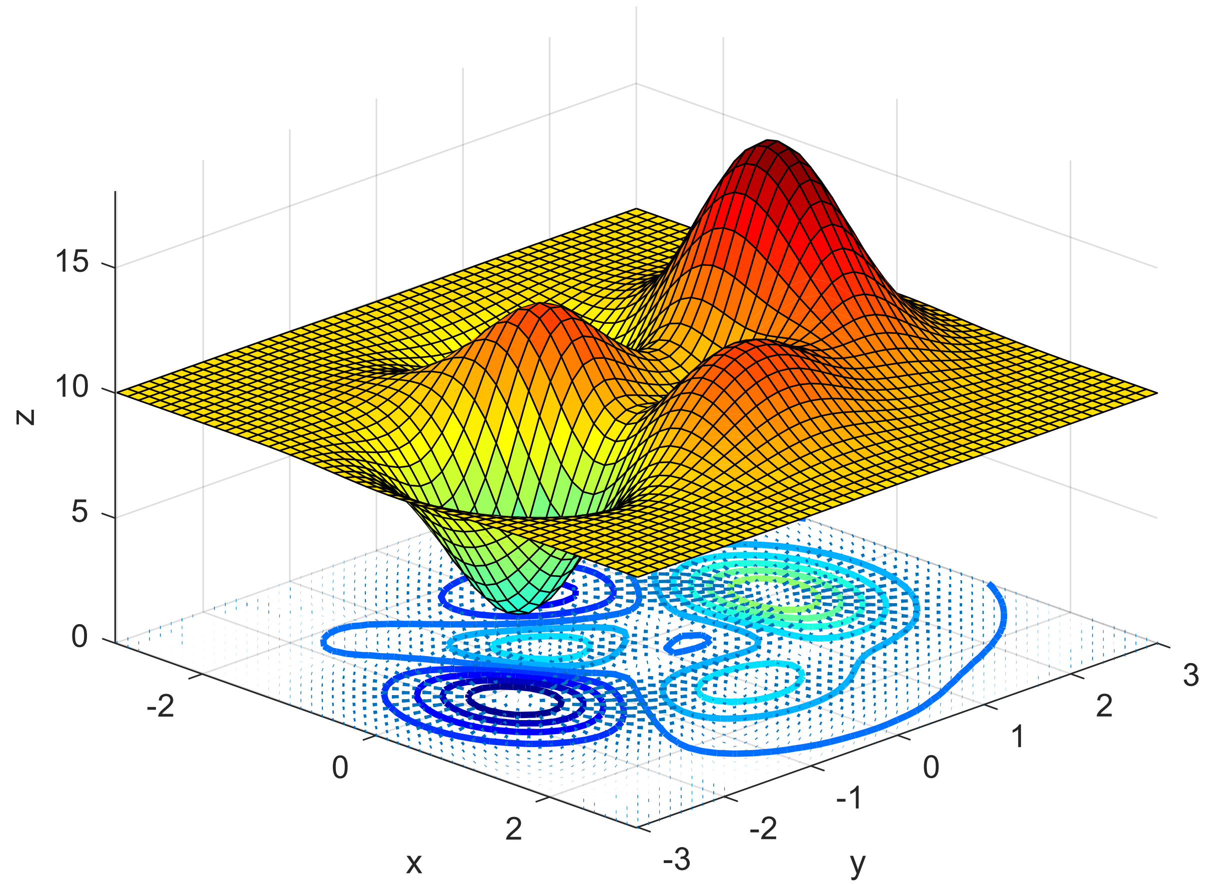
优化：梯度下降

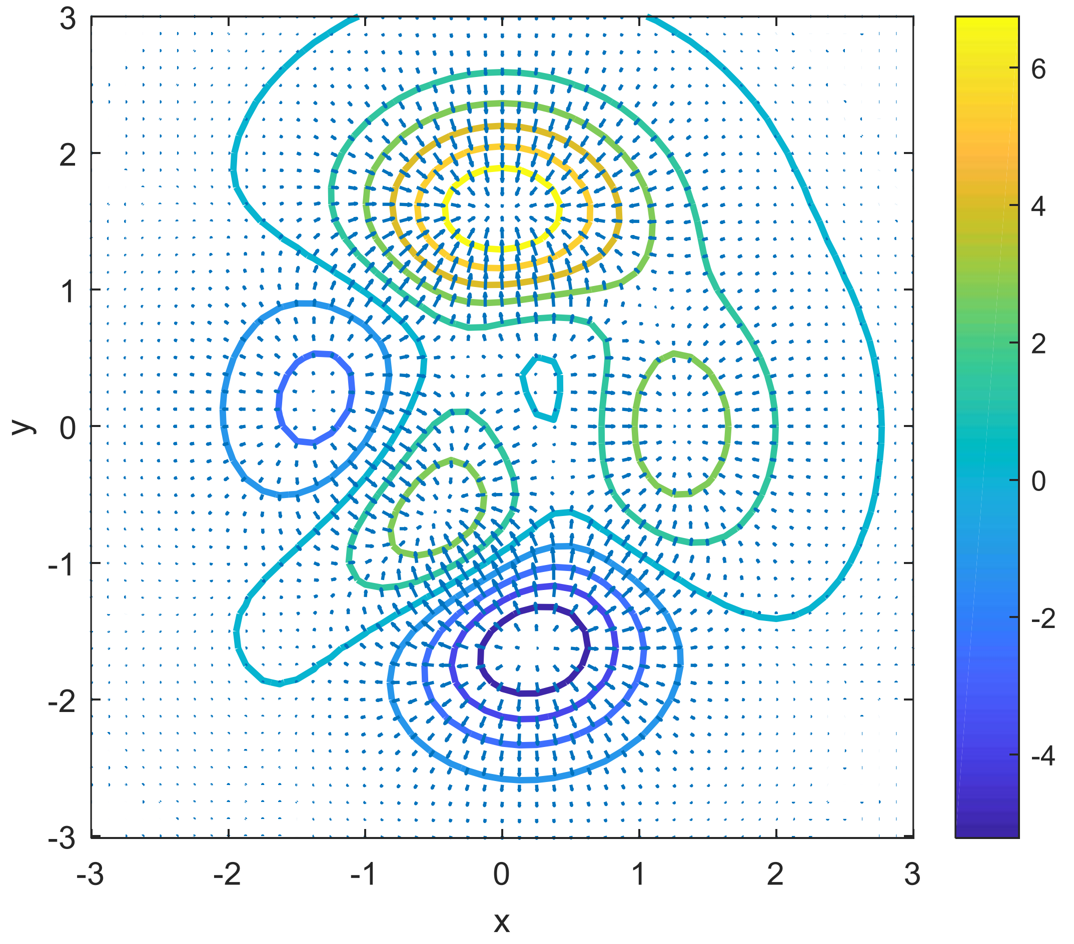
函数$f(\boldsymbol{x})$的梯度为函数$f(\boldsymbol{x})$的偏导数

$$\nabla f(\boldsymbol{x}) = \frac{df(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

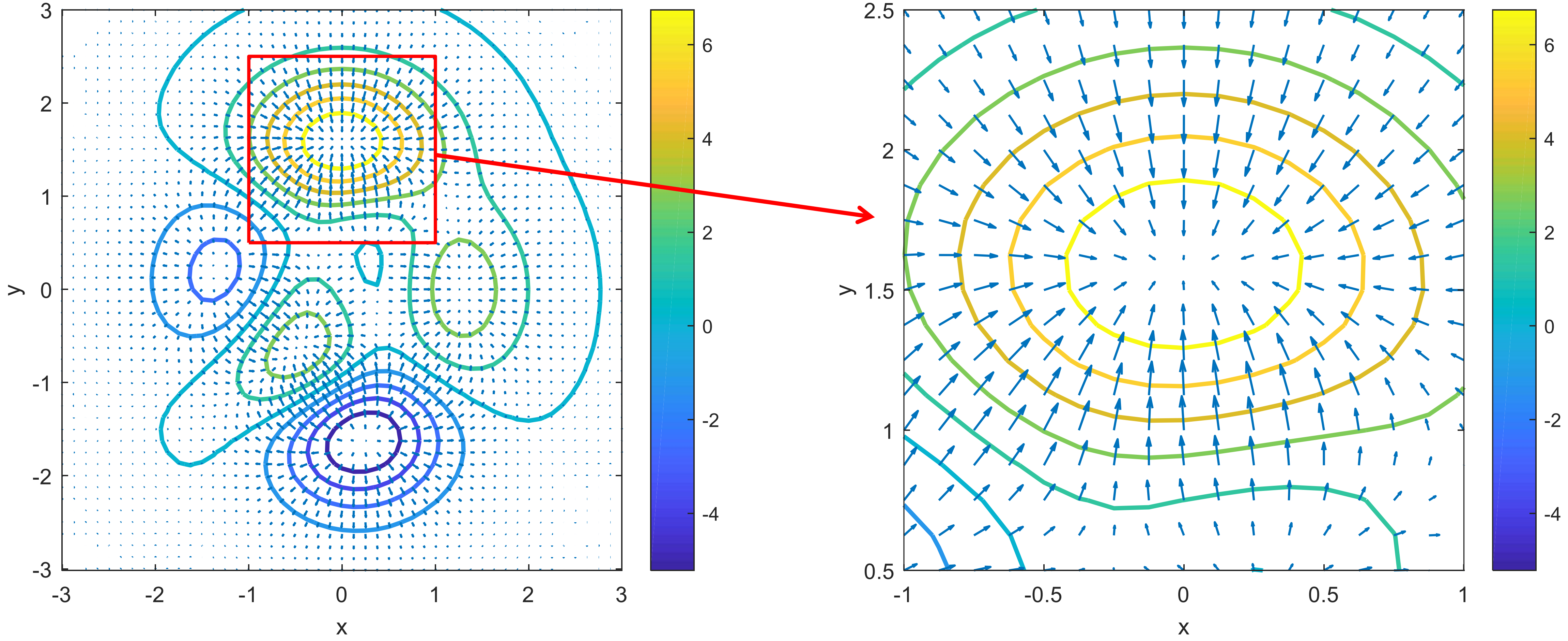
- 梯度一般写为$\nabla f(\boldsymbol{x})$ 或者 grad$(f(\boldsymbol{x}))$，
- 梯度反向指向函数值增加最快的方向。梯度的负方向指向函数值减小最快的方向。
- 梯度方向与函数的等高线垂直。

优化：梯度





优化：梯度



优化：梯度

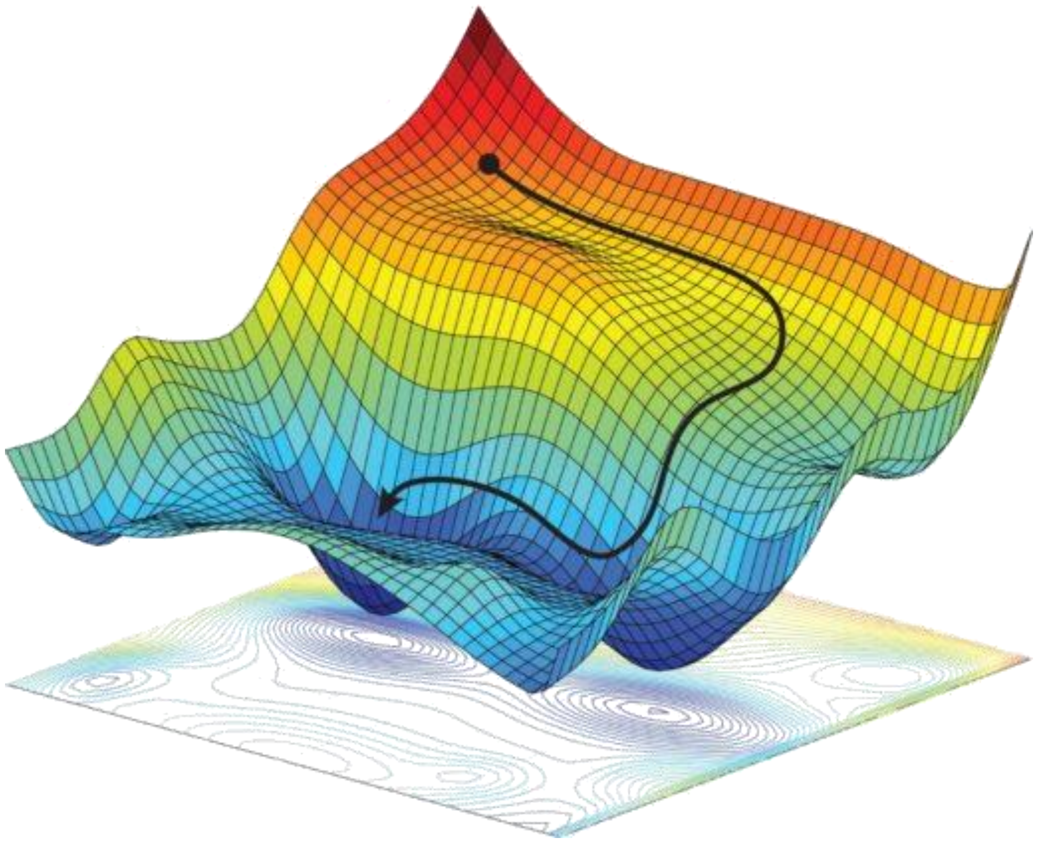
解决如下优化问题

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

梯度下降迭代地沿着负梯度方向更新变量

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \gamma \nabla f(\boldsymbol{x}_i)$$

，这里 γ是步长系数。



Trajectory by gradient descent

The pic for “trajectory by gradient descent” is from https://machinelearningknowledge.ai/keras-optimizers-explained-with-examples-for-beginners/.

优化：梯度下降

如何确定步长?

- 选择一个很小γ，比如0.001.
- 解决如下优化问题，得出 γ

$$\min_{\gamma} f(\boldsymbol{x}_i - \gamma \nabla f(\boldsymbol{x}_i))$$

迭代何时停止?

- 当负梯度幅值接近零。

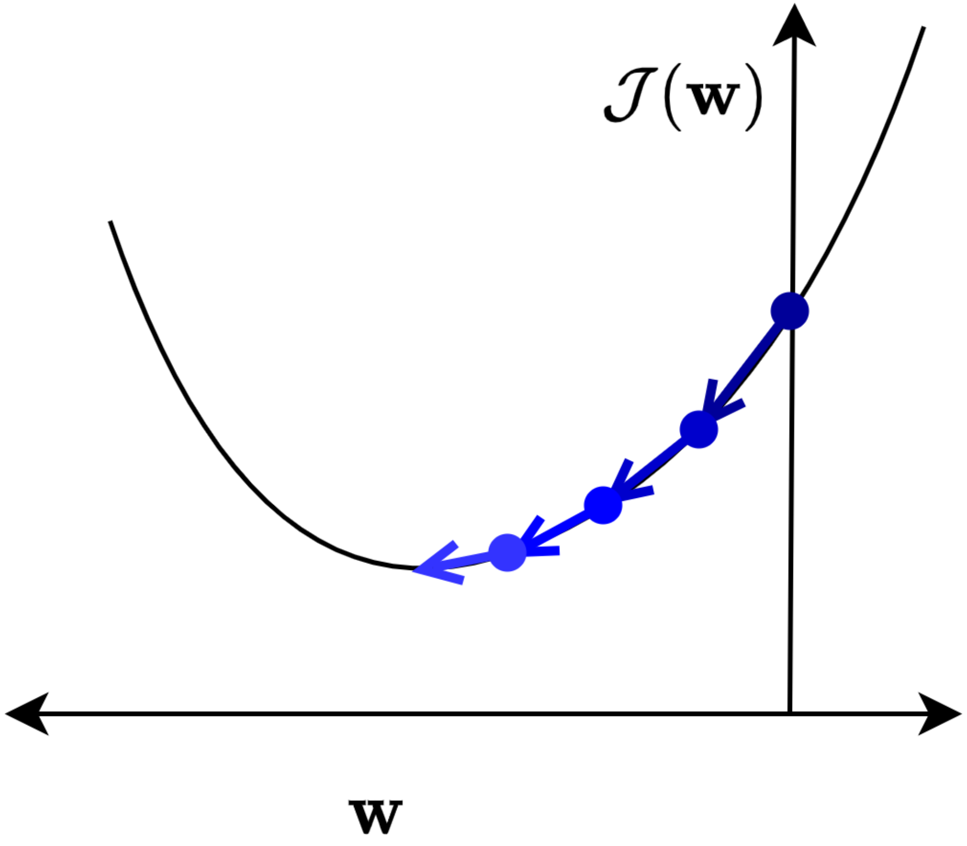
优化：梯度下降

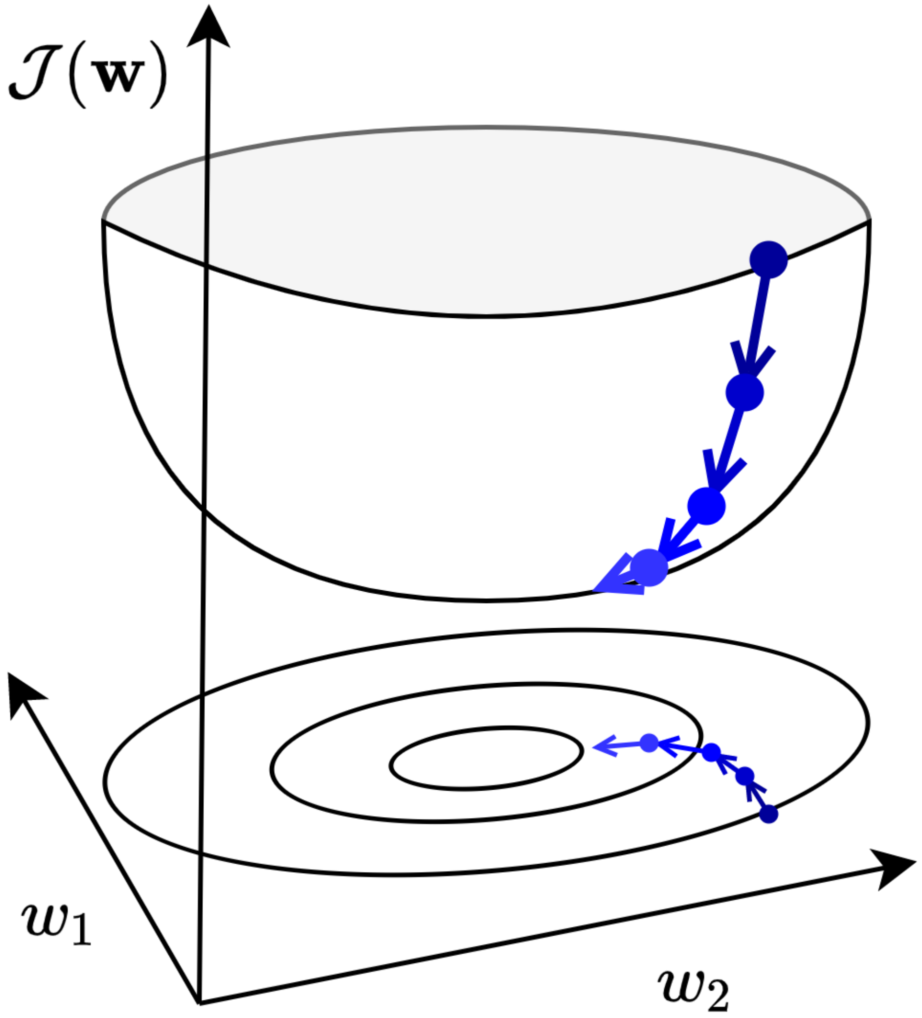
梯度下降伪代码

输入： 目标函数$f(\boldsymbol{x})$，步长γ.
输出： $\boldsymbol{x}$.

(1) **初始化：** 选择$\boldsymbol{x}_0$.
(2) **循环至停止条件满足：**
(3) 计算梯度 $\nabla f(\boldsymbol{x}_{i-1})$
(4) $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_{i-1} - \gamma \nabla f(\boldsymbol{x}_{i-1})$.
(5) 返回 $\boldsymbol{x}_i$.

优化：梯度下降

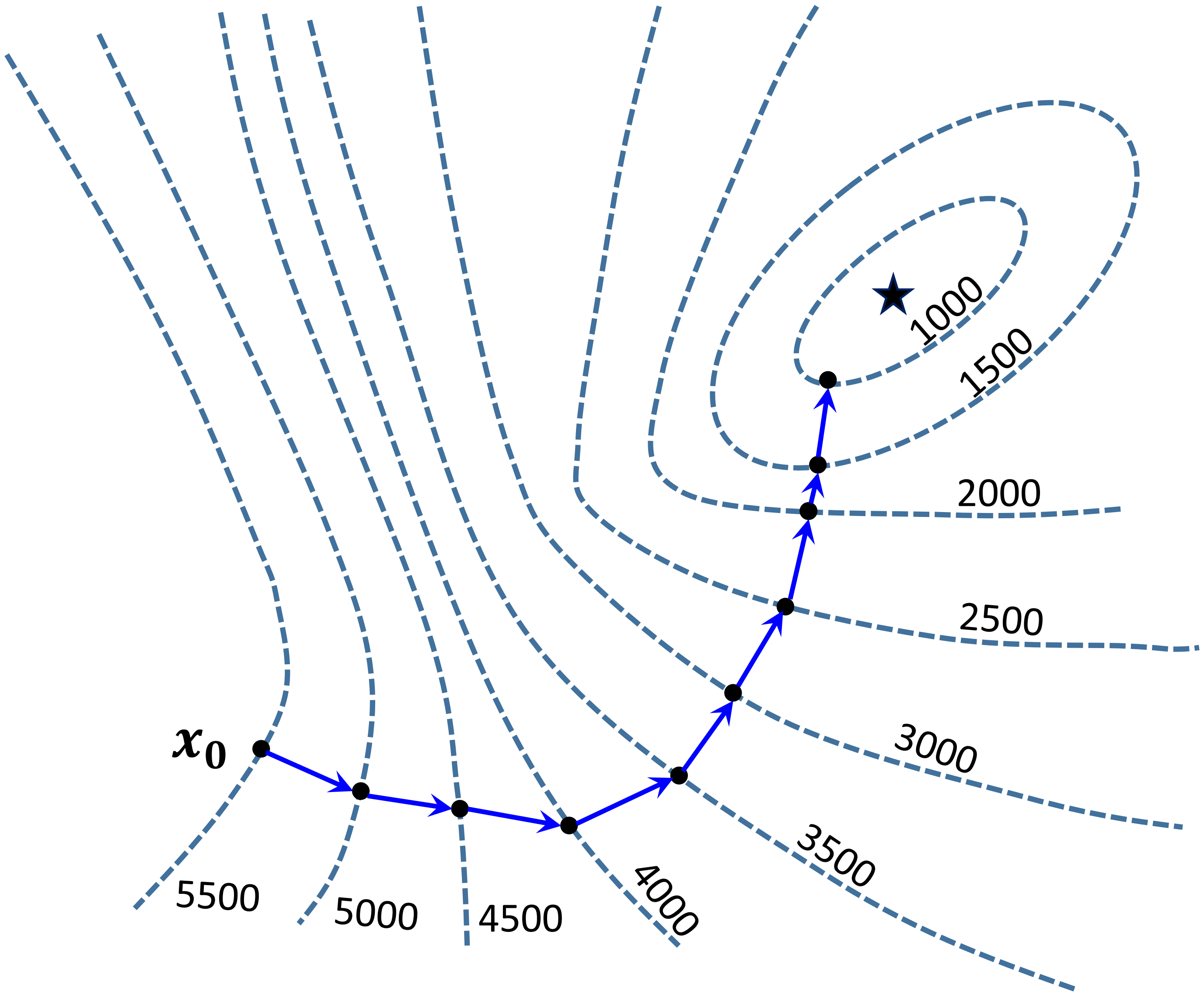


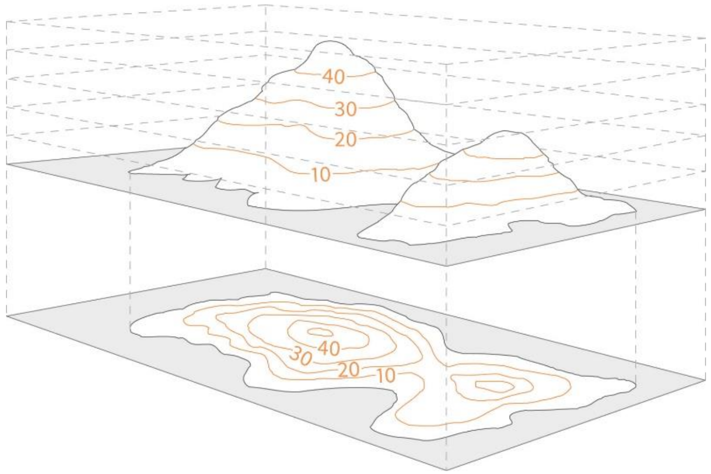


Pics are from CSC311 Introduction to machine learning 2020 Uuniversity of Toronto

优化：梯度下降

梯度下降





等高线示例

来自
https://getoutside.ordnancesurvey.co.uk/guides/understanding-map-contour-lines-for-beginners/

优化：牛顿法

函数在最小值处有

$$\nabla f(\boldsymbol{x}) = \mathbf{0}$$

在某一点 $\boldsymbol{x}_0$处，泰勒展开为

$$f(x) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2)$$

忽略二次及以上项，有

$$f(x) \approx f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)$$

上式两边同时计算梯度，有

$$\nabla f(x) \approx \nabla f(\boldsymbol{x}_0) + \nabla^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)$$

如果$\nabla f(x) = 0$, 有

$$\boldsymbol{x} = \boldsymbol{x}_0 - (\nabla^2 f(\boldsymbol{x}_0))^{-1} \nabla f(\boldsymbol{x}_0)$$

可以看出 $\boldsymbol{x}_0$ 更新为 $\boldsymbol{x}_0 - (\nabla^2 f(\boldsymbol{x}_0))^{-1} \nabla f(\boldsymbol{x}_0)$.

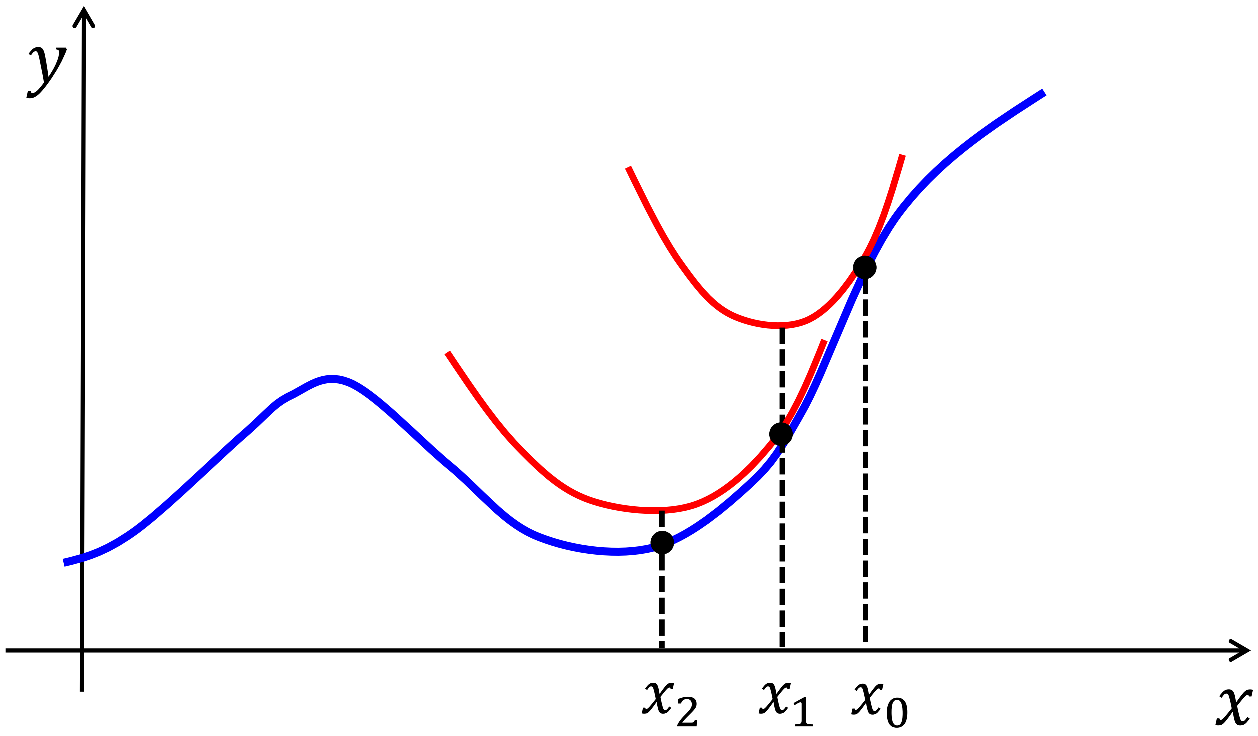
优化：牛顿法

牛顿法下降伪代码

输入： 目标函数 $f(\boldsymbol{x})$.
输出： $\boldsymbol{x}$.

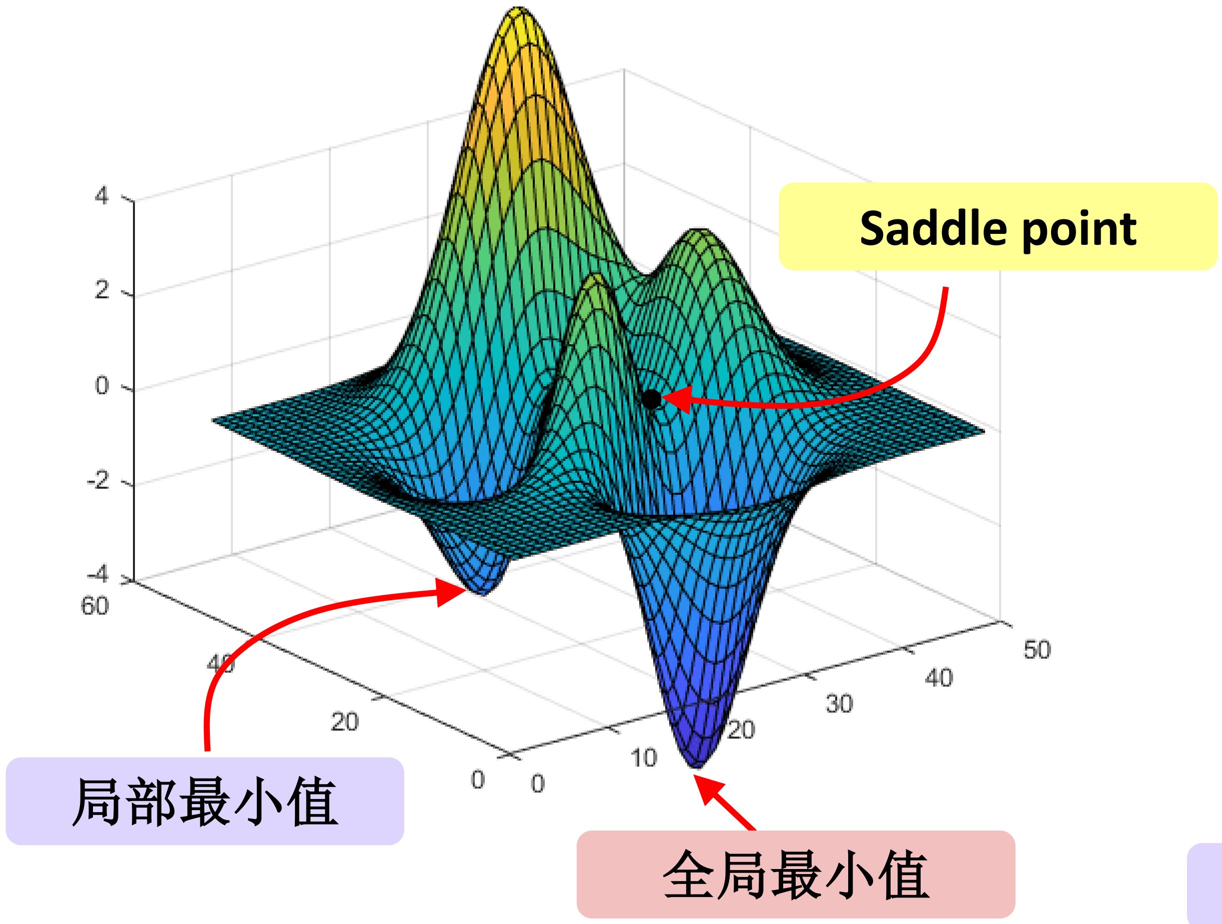
(1) **初始化：** 选择一个$\boldsymbol{x}_0$.
(2) **循环至终止条件满足：**
(3) 计算梯度 $\nabla f(\boldsymbol{x}_{i-1})$ 和Hessian矩阵 $\nabla^2 f(\boldsymbol{x}_0)$.
(4) $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_{i-1} - (\nabla^2 f(\boldsymbol{x}_{i-1}))^{-1} \nabla f(\boldsymbol{x}_{i-1})$
(5) 返回 $\boldsymbol{x}_i$.

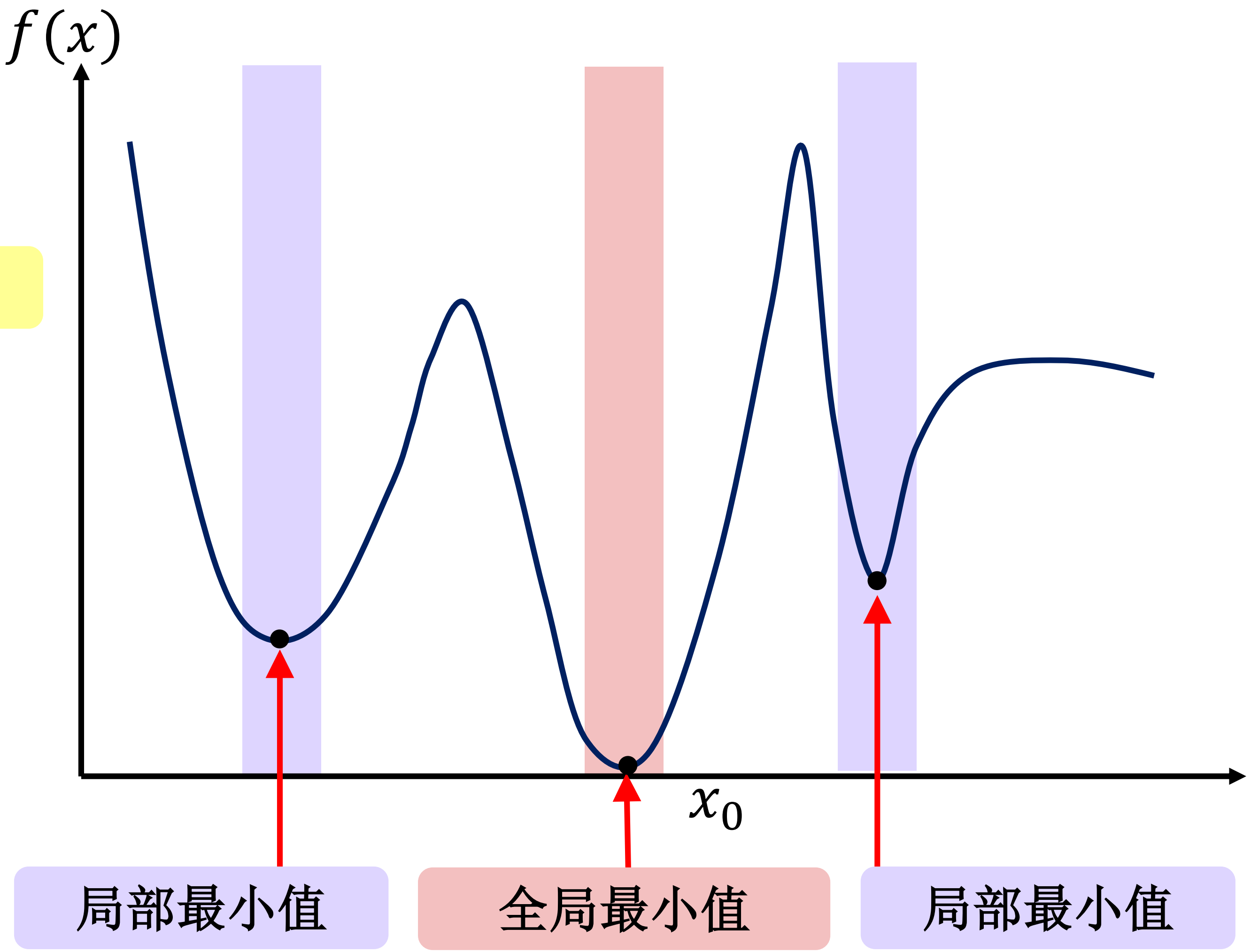
优化：牛顿法



如果目标函数是二次函数，牛顿法迭代一步即可得到最优解。

优化：全局最小值和局部最小值





Pic on left is from https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/.

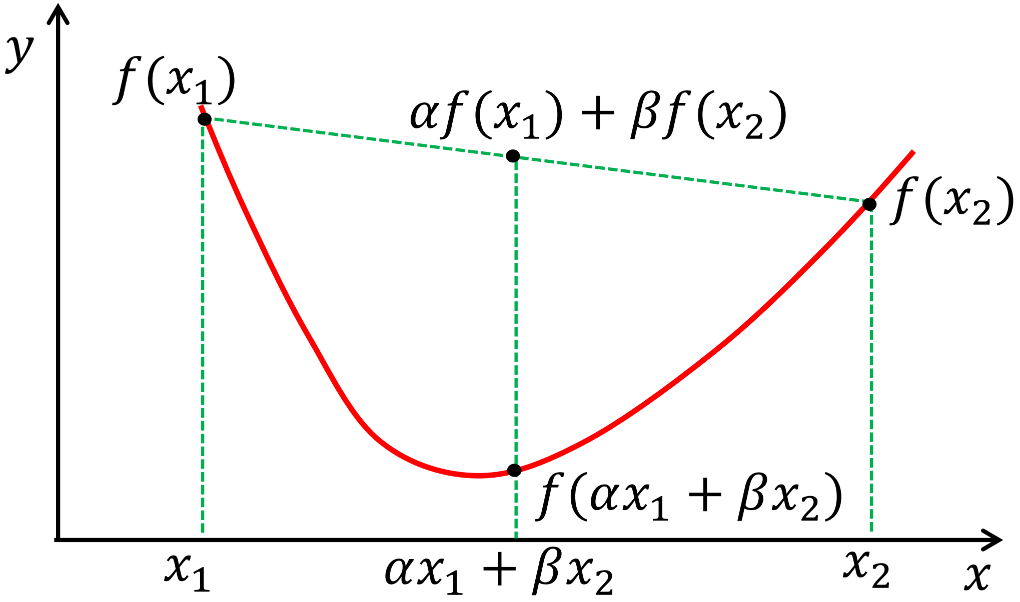
凸优化

- **凸函数**

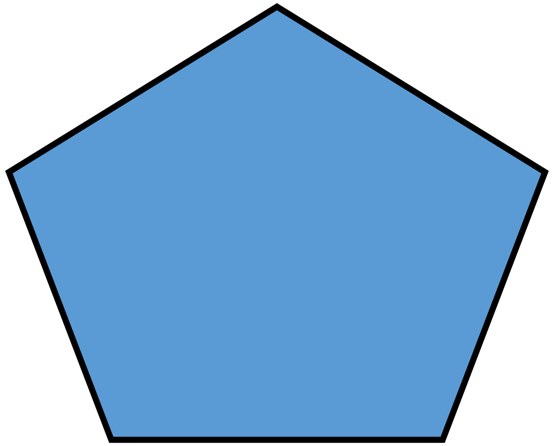
$$f(\alpha x_1 + \beta x_2) \le \alpha f(x_1) + \beta f(x_2) \qquad \alpha + \beta = 1, \alpha > 0, \beta > 0$$

- **凸集合**

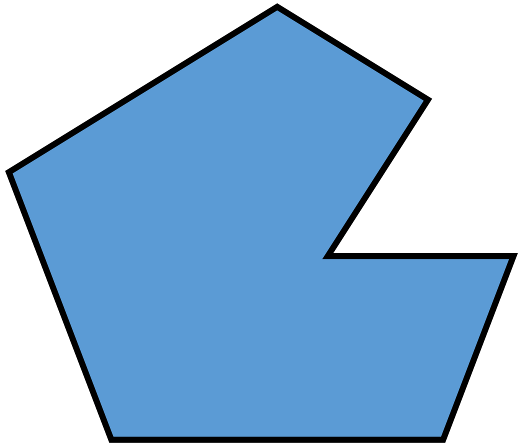
集合内任何两点连线也在集合内部。



凸函数



凸区域



非凸区域

凸优化

如果一个有约束优化问题的目标函数是**凸函数**，可行域是**凸区域**，那么这个优化问题为**凸优化问题。**

凸优化问题有且仅有一个解。换而言之，全局最优解也是局部最优解。

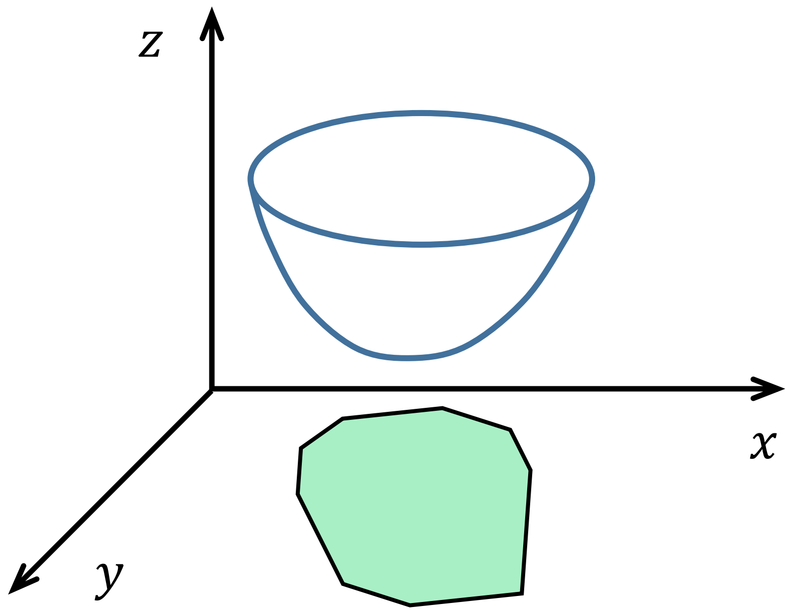
An optimization problem is a convex optimization problem is the objective function is convex and the feasible region is also convex.

For convex optimization problem, there is only one solution. In other words, local optimal solution is also the global optimal solution.

凸优化

如果一个有约束优化问题的目标函数是**凸函数**，可行域是**凸区域**，那么这个优化问题为**凸优化问题**。

凸优化问题有且仅有一个解。换而言之，全局最优解也是局部最优解。



优化：拉格朗日法

拉格朗日法用于解决等式约束的优化问题

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{s.t.} \quad c_i(\boldsymbol{x}) = 0, \quad i = 1,2,\cdots k$$

构造如下拉格朗日方程

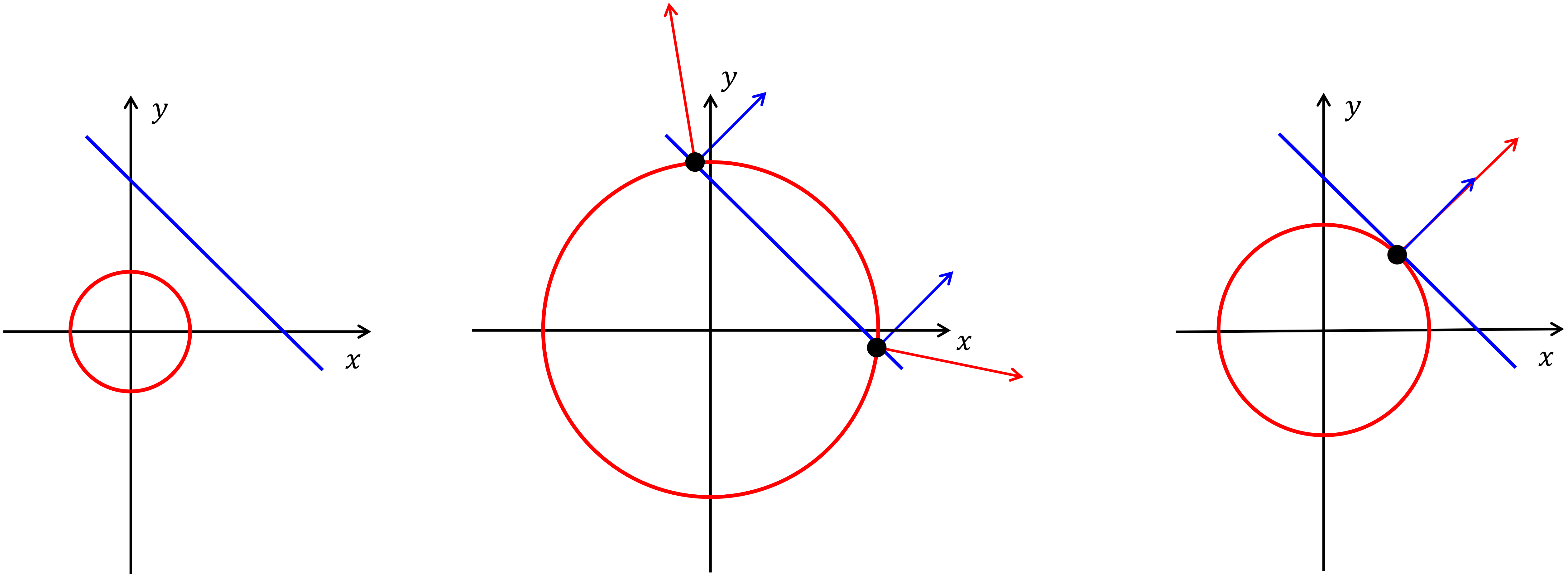
$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{k} \lambda_i c_i(\boldsymbol{x})$$

，其中 λ_i 被称作拉格朗日乘子(multipliers)，为新的未知数。最小化拉格朗日方程，可以得到最优解。

通过转换成拉格朗日方程，将一个有约束优化问题转换为无约束优化问题。代价是变量增多。

优化：拉格朗日法

$$\min_{\boldsymbol{x}} x_1^2 + x_2^2 \qquad s.t.\ x_1 - x_2 = 5$$



优化：KKT条件

有约束优化问题

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{s.t. } c_i(\boldsymbol{x}) \leq 0, i = 1,2,\cdots,p$$
$$g_i(\boldsymbol{x}) = 0, i = 1,2,\cdots,q$$

拉格朗日方程为

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i c_i(\boldsymbol{x}) + \sum_{i=1}^{q} \mu_i g_i(\boldsymbol{x})$$

其中$\boldsymbol{\lambda}$和$\boldsymbol{\mu}$被称为KKT乘子(multipliers).

优化：KKT条件

最优解$\boldsymbol{x}^*$ 满足如下KKT条件

$$\nabla L(\boldsymbol{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}$$
$$\lambda_i \geq 0$$
$$\lambda_i c_i(\boldsymbol{x}) = 0$$
$$c_i(\boldsymbol{x}) \leq 0$$
$$g_i(\boldsymbol{x}) = 0$$

KKT 条件仅仅是最优解的必要但不充分条件。