

# ECE-GY 6233 System Optimization Methods

## Mathematic basic

<https://www.bilibili.com/video/BV1Eg41177FC>

## 1 Linear algebra

### 1.1 Matrix and calculation

#### 1. Basic matrix concepts

我们先来看看什么是矩阵。如下,这是一个 $m \times n$ 的矩阵,  $A_{ij}$ 表示是第  $i$  行第  $j$  列的元素

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

如果是行向量(row vector), 那么就是 $m=1$ , 如果是列向量(column vector), 那么就是 $n=1$ .

Identity matrix就是对角线是1, 其余值为0. 而且, 这是一个square matrix

$$A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & 1 & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

然后还有diagonal matrix, 这个 $\lambda$ 的值可以是0, 也可以是其他值

$$A = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \lambda_i & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}$$

#### 2. Basic matrix calculation (addition, subtraction, multiplication)

矩阵的加加法都非常简单

$$A \pm B = \begin{bmatrix} a_{11} \pm b_{11} & a_{12} \pm b_{12} & \cdots & a_{1n} \pm b_{1n} \\ a_{21} \pm b_{21} & a_{22} \pm b_{22} & \cdots & a_{2n} \pm b_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} \pm b_{m1} & a_{m2} \pm b_{m2} & \cdots & a_{mn} \pm b_{mn} \end{bmatrix}$$

如果一个常数 $\lambda$ 跟矩阵相乘, 则是

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

如果是两个矩阵的相乘, 则必须是  $A_{m \times n} \cdot B_{n \times s}$ . 最后会得到一个 $m \times s$ 的矩阵

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1s} \\ b_{21} & b_{22} & \cdots & b_{2s} \\ \vdots & \vdots & \cdots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{ns} \end{bmatrix}$$
$$A \cdot B = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + \cdots + a_{1n}b_{n1} & a_{11}b_{12} + a_{12}b_{22} + \cdots + a_{1n}b_{n2} & \cdots & a_{11}b_{1s} + a_{12}b_{2s} + \cdots + a_{1n}b_{ns} \\ a_{21}b_{11} + a_{22}b_{21} + \cdots + a_{2n}b_{n1} & a_{21}b_{12} + a_{22}b_{22} + \cdots + a_{2n}b_{n2} & \cdots & a_{21}b_{1s} + a_{22}b_{2s} + \cdots + a_{2n}b_{ns} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}b_{11} + a_{m2}b_{21} + \cdots + a_{mn}b_{n1} & a_{m1}b_{12} + a_{m2}b_{22} + \cdots + a_{mn}b_{n2} & \cdots & a_{m1}b_{1s} + a_{m2}b_{2s} + \cdots + a_{mn}b_{ns} \end{bmatrix}$$

#### 3. Trace of a matrix

在square matrix中, trace就是对角线的和。这里呢, 有一个性质是 $tr(A_{m \times n} \cdot B_{n \times s}) = tr(A_{m \times n} \cdot B_{n \times s})$

#### 4. Transpose of a matrix

转置就是行跟列反过来

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & 1 \end{bmatrix} \quad A^T = \begin{bmatrix} 1 & 3 \\ 2 & -1 \\ 0 & 1 \end{bmatrix}$$

转置的一些性质

- $(A^T)^T = A$
- $(A + B)^T = A^T + B^T$
- $(\lambda A)^T = \lambda A^T$
- $(AB)^T = B^T A^T$

## 5. Symmetric matrix

对称矩阵满足  $A^T = A$ . 以对角线为对称轴, 对应的值相等。比如说

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

[https://www.cxyzjd.com/article/qg\\_24690701/81839016](https://www.cxyzjd.com/article/qg_24690701/81839016)

[https://www.cxyzjd.com/article/robert\\_chen1988/92437038](https://www.cxyzjd.com/article/robert_chen1988/92437038)

## 1.2 Determinant of Matrix

### 1. 行列式的引入

我们先看一下行列式是怎么来的。我们从解方程组的角度去看。假设有如下的二元线性方程组

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \end{cases}$$

其实, 这个二元线性方程也是可以写成矩阵形式的, 如下

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

然后我们要解这个方程组, 用消元法把  $x_1, x_2$  给求出来, 我们可以得到如下的式子

$$x_1 = \frac{b_1 a_{22} - a_{12} b_2}{a_{11} a_{22} - a_{12} a_{21}}, x_2 = \frac{a_{11} b_2 - b_1 a_{21}}{a_{11} a_{22} - a_{12} a_{21}}$$

假设有一个矩阵A, 那么A的行列式的表达就是  $|A|$ 。我们先看二阶行列式, 它的定义是

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, |A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

若记

$$D = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, D_1 = \begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}, D_2 = \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix},$$

所以上述  $x_1, x_2$  的解可以写成

$$x_1 = \frac{D_1}{D} = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}, x_2 = \frac{D_2}{D} = \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}$$

起初, 行列式就是用来替代多元方程式的解, 用其来表达。行列式就是一个值。

那么三阶甚至更高阶的行列式应该怎么写呢, 有什么规律呢? 规律就是一下三步

- 全排列
- 逆序数
  - 我们先来看下什么是逆序数, 所谓逆序数, 就是看这个数的前面有几个是比当前那个数大的。给定一个排列, 比如说32514, 这里一个有五个数, 分别是3,2,5,1,4。这五个数的逆序数分别是: 3, 排在首位, 逆序数为0。2, 前面有一个3, 比2大, 所以逆序数是1。5, 在这里是最大的数, 逆序数为0。1, 比1大的数有3,2,5, 有三个, 于是逆序数就是3。4, 前面比4大的就只有5, 于是逆序数就是1, 因为只有一个。然后这个排列32514的逆序数就是  $0+1+0+3+1=5$ 。
- 偶正奇负
  - 行列式把一项一项写出来后, 不是有负有正吗? 那什么时候是正, 什么时候是负呢? 就看逆序数, 如果是偶数, 则为正, 奇数, 则为负。

于是, 行列式的固定公式就是

假设是n阶

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} = \sum (-1)^t a_{1p_1} a_{2p_2} \cdots a_{np_n}$$

这里  $p_1 p_2 \cdots p_n$  是  $1 \dots n$  的全排列。相当于就是说每次去不同行不同列的数进行相乘。然后  $t$  的值就是  $p_1 p_2 \cdots p_n$  全排列后的逆序数。如果是  $n$  阶，那么就一共有  $n!$  项。

现在呢，我们主要是记住三阶的怎么写

假设如下matrix,

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

现在我们要写出三阶的determinant, 结果如下

$$\det(A) = a(ei - fh) - b(di - fg) + c(dh - eg)$$

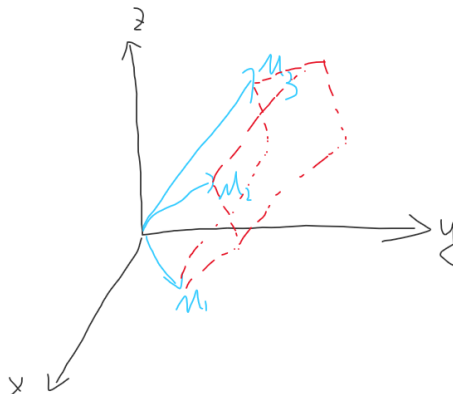
假设现在的三阶矩阵是

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & -2 & 3 \\ 0 & 5 & -1 \end{bmatrix}$$

那么，计算得到的行列式就是

$$\begin{aligned} \det(A) &= 1 \times ((-2 \times -1) - (3 \times 5)) - 2 \times ((4 \times -1) - (3 \times 0)) + 3 \times (4 \times 5 - (-2 \times 0)) \\ &= -13 + 8 + 60 \\ &= 55 \end{aligned}$$

还有，行列式的绝对值，其实就是向量所组成的体积，volumn。举个例子来说，假设给出三个vectors,  $\mu_1 = (1, 1, 0), \mu_2 = (1, 1, 1), \mu_3 = (0, 2, 3)$ . 我们要计算这三个向量所组成的体积，如下图所示



那么，我们就可以计算这三个向量的行列式的绝对值，顺序无所谓

$$V = \left| \det \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 2 & 3 \end{bmatrix} \right|$$

## 2. 行列式的重要性质

- if  $\det(A) = 0$ , then  $A$  is a singular matrix. Singular matrix does not have inverse. 就是不可逆
- $\det(A^T) = \det(A)$
- $\det(AB) = \det(A)\det(B)$
- $\det(A^{-1}) = (\det(A))^{-1}$
- 这个性质就需要画一下图，假设有如下这么一个矩阵，可以切成如下部分，这里要注意， $A$ 和 $D$ 在切的时候一定是square matrix, 对 $B$ 和 $C$ 就无所谓了， $C=0$ 的意思就是说，那一个block全部元素都为0. 那个这个矩阵的行列是就是  $\det(M) = \det(A)\det(D)$

$$M = \begin{bmatrix} A & B \\ C=0 & D \end{bmatrix}$$

举个例子来说

$$M = \begin{bmatrix} 2 & 3 & 4 & 7 & 8 \\ -1 & 5 & 3 & 2 & 1 \\ 0 & 0 & 2 & 1 & 5 \\ 0 & 0 & 3 & -1 & 4 \\ 0 & 0 & 5 & 2 & 6 \end{bmatrix}$$

我们可以进行如下的划分，这样我们就只需要求 $\det(A)\det(D)$ 。

$$M = \begin{array}{cc|ccc} & \text{A} & & \text{B} & & \\ \begin{array}{c} 2 \\ -1 \\ 0 \\ 0 \\ 0 \end{array} & \begin{array}{c} 3 \\ 5 \\ 0 \\ 0 \\ 0 \end{array} & \begin{array}{c} 4 \\ 3 \\ 2 \\ 3 \\ 5 \end{array} & \begin{array}{c} 7 \\ 2 \\ 1 \\ -1 \\ 2 \end{array} & \begin{array}{c} 8 \\ 1 \\ 5 \\ 4 \\ 6 \end{array} \\ \hline & \text{C} & & \text{D} & & \end{array}$$

当然了，我们有时候并不能直接看出这样的情况，可能要进行划分啊，交换啊之类的，才能得到这样的矩阵，化简。

3. 行列式按行(列)展开，代数余子式

4. 克莱姆法则

### 1.3 Inverse of matrix

1. 矩阵的逆

2. 矩阵逆的性质

### 矩阵的初等变换

### 矩阵的秩

## 1.6 Eigenvector and Eigenvalue

### 1. The inner product and norm

Inner product是两个向量之间的一种运算，结果是一个实数。举个例子

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$x \cdot y = x^T y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

这里有个不等式，Cauchy-Schwarz inequality，就是柯西不等式啦。这里 $[\cdot, \cdot]$ 表示的就是内积

$$[x, y]^2 \leq [x, x][y, y]$$

现在我们来讲范数，就是norm，其实norm可以表示成一下的形式

$$\|x\|_2 = \sqrt{[x, x]} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

其实，这就是 $l_2$  norm了。我们可以有两种表达，如下

$$\|x\|_2 = \sqrt{[x, x]} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\sum_{i=1}^n x_i^2}$$

OR

$$\|x\|_2^2 = [x, x] = x_1^2 + x_2^2 + \cdots + x_n^2 = \sum_{i=1}^n x_i^2 = x^T x$$

关于这个norm，有如下几个性质，假设 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 。就是说这是一个n维的向量，然后里边的数都是real number

- $\forall x \in \mathbb{R}^n, f(x) \geq 0$ . 就是说如果这个向量里边的数都是real number，那么它的norm就一定 $\geq 0$ 。(non-negative)
- $f(x) = 0$  if and only if  $x=0$ . 就是当向量x里边的数都是0的时候，那么norm也就是0
- $\forall x \in \mathbb{R}^n, t \in \mathbb{R}, f(tx) = |t|f(x)$ . 这个就是说 $\|\lambda x\| = |\lambda| \|x\|$  (Homogeneity)
- $\forall x, y \in \mathbb{R}^n, f(x, y) \leq f(x) + f(y)$  (triangle property). 这个也非常好理解，就是 $\|x + y\| \leq \|x\| + \|y\|$ .

这里再讲一些其他norm

- $l_1$  - norm: 就是 $\|x\|_1 = \sum_{i=1}^n |x_i|$ . 也就是绝对值相加
- $l_p$  - norm: 就是 $(\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ .
- $l_\infty$  - norm: 就是 $\|x\|_\infty = \max_i |x_i|$ . 就是绝对值中的最大。举个例子来说， $x = [3, 5]$ . 假设有个100次方来看看，那就是 $(3^{100} + 5^{100})^{0.01} = 5$ . 记住了，就是去绝对值中最大的那个数

接下来再讲一个Frobenius norm, 之前的norm是处理vector的, 这个是处理matrix的. 公式如下, 这里呢, tr就是trace的意思, 就是对角线的数相加, 之前讲过了

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} = \sqrt{\text{tr}(A^{-1}A)}$$

## 2. 特征值特征向量以及矩阵的相似

设  $A$  是  $n$  阶矩阵, 如果数  $\lambda$  和  $n$  维非零列向量  $x$  使如下关系式成立

$$Ax = \lambda x$$

那么, 这个数  $\lambda$  就是矩阵  $A$  的特征值, 非零向量  $x$  就是  $A$  的对应于特征值  $\lambda$  的特征向量

关于 Eigenvalue 和 eigenvector, 我们需要记住以下几个东西

- $\det(A - \lambda I) = 0$ . 因为  $(A - \lambda I)x = 0$ .  $x$  是非零的列向量
- $(A - \lambda I)$  is singular. 就是不能 inverse
- 一个  $n$  阶 matrix, 是有  $n$  个 eigenvalues 的.  $\lambda_1 + \lambda_2 + \dots + \lambda_n = a_{11} + a_{22} + \dots + a_{nn} = \text{tr}(A)$ . 也就是  $A$  的迹, 对角线之和就是 eigenvalue 的和
- $|A| = \lambda_1 \lambda_2 \dots \lambda_n$ . 也就是矩阵  $A$  的行列式等于 eigenvalues 的乘积

举个例子来看看, 我们要求矩阵  $A$  的特征值和特征向量

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

我们可以得到

$$\begin{aligned} \det(A - \lambda I) &= \det \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \\ &= (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0 \\ &\quad \downarrow \\ \lambda^2 - (a_{11} + a_{22})\lambda - a_{12}a_{21} &= 0 \end{aligned}$$

这个求解  $\lambda$  的方程, 就是 Characteristic equation, 我们求解出  $\lambda$  的值即可. 得到  $\lambda$  的值, 我们再把  $\lambda$  代进去从而求 eigenvector

求 eigenvectors 的方式是  $(A - \lambda I)v = 0$ .

这里来看个例子吧

## 1.7 Quadratic form and definiteness.

这里来讲一讲 Quadratic form, 就是二次方的形式. 就是任意一个二次方的形式都可以用  $x^T A x$ ,  $x \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{n \times n}$  来表示. 这个  $A$  一定是 symmetric 的. 我们用一个通用形式来表达

$$\begin{aligned} f &= a_{11}x_1^2 + a_{12}x_1x_2 + \dots + a_{1n}x_1x_n + \\ &\quad a_{21}x_2x_1 + a_{22}x_2^2 + \dots + a_{2n}x_2x_n + \\ &\quad \dots \\ &\quad + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \dots + a_{nn}x_n^2 \\ &= [x_1, x_2, \dots, x_n] \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \end{aligned}$$

举个例子来说吧, 就非常简单好理解

假设有如下方程式, 写出 symmetric matrix  $A$  的表示

$$ax^2 + by^2 + cz^2 + 2fyz + 2gzx + 2hxy$$

已知这是一个 quadratic form,  $A$  就很容易写出来了

$$A = \begin{bmatrix} a & h & g \\ h & b & f \\ g & f & c \end{bmatrix}$$

现在用一个实例来说明, Find the symmetric matrix  $A$  which gives the quadratic form  $3x^2 - 2y^2 + 8xy - 5yz + xz + z^2$

$$A = \begin{bmatrix} 3 & 4 & \frac{1}{2} \\ 4 & -2 & -\frac{5}{2} \\ \frac{1}{2} & -\frac{5}{2} & 1 \end{bmatrix}$$

现在我们来验证一下  $x^T A x$  是不是等于题目给出的 quadratic form

$$x^T A x = \begin{bmatrix} x \\ y \\ z \end{bmatrix}^T \begin{bmatrix} 3 & 4 & \frac{1}{2} \\ 4 & -2 & -\frac{5}{2} \\ \frac{1}{2} & -\frac{5}{2} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 3x^2 - 2y^2 + 8xy - 5yz + xz + z^2$$

关于这个  $x^T A x$ , 有一些变换我们是需要了解的。我们必须清楚,  $A$  is symmetric matrix

$$(x^T A x)^T = x^T A x = x^T A^T x = x^T \left( \frac{1}{2} A + \frac{1}{2} A^T \right) x$$

接下来就是看 definiteness, 也就是我们要了解的 positive definite, negative definite, positive semi-definite, negative semi-definite 了

就是给了一个 quadratic form, 我们可以写出  $f(x) = x^T A x$ 。于是, 我们就可以根据这个来判断  $A$  的 definiteness。

- we call the quadratic form  $f(x)$ ,  $A$  is positive definite, if  $f(x) > 0, \forall x \neq 0$ .
- we call the quadratic form  $f(x)$ ,  $A$  is positive semi-definite, if  $f(x) \geq 0$  for all  $x$ .
- we call the quadratic form  $f(x)$ ,  $A$  is negative definite, if  $f(x) < 0, \forall x \neq 0$ .
- we call the quadratic form  $f(x)$ ,  $A$  is negative semi-definite, if  $f(x) \leq 0$  for all  $x$ .
- we call the quadratic form  $f(x)$ ,  $A$  is indefinite, if there exists  $x_1, x_2$   $f(x_1) > 0 > f(x_2)$ .

We can use another way to check the definiteness by looking at the eigenvalues.

- $A$  is positive definite if  $\forall i, \lambda_i > 0$ .
- $A$  is positive semi-definite if  $\forall i, \lambda_i \geq 0$ .
- $A$  is negative definite if  $\forall i, \lambda_i < 0$ .
- $A$  is negative semi-definite if  $\forall i, \lambda_i \leq 0$ .
- $A$  is indefinite if there exists  $j, k, \lambda_j > 0 > \lambda_k$

还有更简单的方法来看 matrix 的 definiteness。这个方法也简单, 后面会总结一下。不过, 看这个方法之前, 先介绍两个概念

Principal minor (PM) and leading principal minor (leading PM)

- Principal minor
  - definition:  $A$  is a  $n \times n$  matrix, a  $k \times k$  submatrix of  $A$  attained by deleting  $n-k$  columns and  $n-k$  rows with same index (remove  $i^{th}$  row, then need to remove  $i^{th}$  col too). This submatrix is a principal submatrix of  $A$ . 就很简单理解, 就是同时删除相同 index 的 row 和 col (任意 row, 任意 col, 只要 index 相同就可以), 得到的 submatrix 就是 principal submatrix.
  - 刚刚给的概念是 principal submatrix, 现在我们要来理解 principal minor。这个的定义是: The determinant of a  $k^{th}$  order principal submatrix of  $A$  is called  $k^{th}$  order principal minor.

- 举个例子来说, 假设有如下矩阵

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

然后我们要求 first order principal minus。这也就意味着我们只能留一行一列。我们也就要删除两行两列, 于是我们就可以得到三个 principal submatrix

$$sub_1 = [a_{11}], sub_2 = [a_{22}], sub_3 = [a_{33}]$$

然后求它们的 determinant。所以  $A$  的 first order principal minors 就是

$$\det(sub_1) = a_{11}, \det(sub_2) = a_{22}, \det(sub_3) = a_{33}.$$

现在假设我们要求 second order principal minor。这也就意味着我们要留两行, 任意删除一行一列, 可以得到

$$sub_1 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, sub_2 = \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix}, sub_3 = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}$$

然后再求它们的 determinant 就行了。

- leading principal minor
  - definition: The leading  $k^{th}$  order principal submatrix of  $A$  is  $k \times k$  matrix attained by deleting the last  $n-k$  rows and cols. 跟刚刚不一样的地方就是这个只能从最后往前删, principal submatrix 是任意删
  - 举个例子来说, 假设有如下的矩阵, 我们要求 first order leading principal minor

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

first order, 也就意味着我们只能留一行一列, 然后从后往前删, 就得到

$$sub = [a_{11}]$$

然后first order leading principal minor就是 $\det(sub) = a_{11}$ .

现在假设要求second order principal minor. 那么就是保留两行, 从后往前删一行一列, 可以得到

$$sub = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

然后求determinant就行了

现在我们明白了leading principal minor的概念了, 我们就可以用这个方式来判断一个matrix的definiteness了。假设 $A$  is  $n \times n$  symmetric matrix, 判断方法如下

- $A$  is positive definite if and only if **all** leading principal minors are greater than 0
- $A$  is negative definite if and only if **all** leading principal minors are less than 0
- if **all** leading principal minors are non-zero, but not in pattern above, then indefinite.

我们来看个例子, 假设有如下矩阵

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}$$

我们就可以计算其leading P.M. 可以得到 $\det[2] = 2 > 0$  and  $\det[A] = 9 > 0$ . So, it is a positive definite.

我们再来看个例子, 假设有如下矩阵

$$A = \begin{bmatrix} -3 & -3 & 0 \\ -3 & -10 & -7 \\ 0 & -7 & -5 \end{bmatrix}$$

我们可以计算all leading P.M. 还有一种方法是计算 $-A$ . 也就是

$$-A = \begin{bmatrix} 3 & 3 & 0 \\ 3 & 10 & 7 \\ 0 & 7 & 5 \end{bmatrix}$$

我们可以得到

$$\det[3] = 3 > 0, \det \begin{bmatrix} 3 & 3 \\ 3 & 10 \end{bmatrix} = 21 > 0, \det[-A] = 21 > 0$$

所以, 我们知道 $-A$ 是positive definite. 这也就说明 $A$ 是negative definite.

## 1.8 Range and nullspace

## 1.9 The Covariance matrix

### 矩阵对角化二次型

### SVD分解证明

### SVD的应用与多元线性回归

## 2 Calculus

Taylor's expansions

## 3 Probability

### 3.1 Basic probability knowledge

#### 1. Random experiment, sample space, random event

Random experiment就是在相同条件下对某随机现象进行大量的重复观测。举个例子来说，投个硬币啊，投个筛子啊之类的。

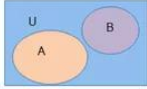
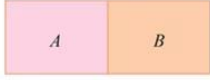
Sample space就是随机试验得到的所有可能结果的集合。比如说，投一枚硬币，那么就会出现两个结果，<H,T>,那么，这个就是sample space

Random event就是sample space中任意一个子集就是random event。

complementary event(互斥事件)

mutual exclusive(对立事件)，不是发生A就是发生B

对立事件一定是互斥事件

	互斥事件	对立事件
定义	若 $A \cap B$ 为不可能事件( $A \cap B = \emptyset$ )，那么称事件A与事件B互斥	若 $A \cap B$ 为不可能事件， $A \cup B$ 为必然事件，那么称事件A与事件B互为对立事件
含义	事件A与事件B在任何一次试验中不会同时发生	事件A与事件B在任何一次试验中有且仅有一个发生
图示	与两个集合的交集为空集类比 	与两个集合的补集类比，即事件A的对立事件是全集中由事件A包含的结果组成的集合的补集，也可以记作 $B = \bar{A}$ 
示例	在掷骰子实验中，定义 $A = \{\text{出现1点}\}$ ， $B = \{\text{出现2点}\}$ ，则事件A与事件B互斥	在掷骰子实验中，定义 $A = \{\text{出现的点数为偶数}\}$ ， $B = \{\text{出现的点数为奇数}\}$ ，则 $A \cap B$ 为不可能事件， $A \cup B$ 为必然事件，所以A与B互为对立事件

#### 2. Conditional probability and multiplication formula

条件概率就是在A发生的前提下发生B的概率，有如下公式

$$P(B|A) = \frac{P(AB)}{P(A)}$$

$P(AB)$ 就是指AB同时发生的概率，就是 $P(A \cap B)$ 。

还有就是乘法公式

$$P(AB) = P(B|A)P(A)$$

$$P(ABC) = P(C|AB)P(B|A)P(A)$$

$$P(A_1 A_2 \cdots A_n) = P(A_n | A_1 A_2 \cdots A_{n-1}) P(A_{n-1} | A_1 A_2 \cdots A_{n-2}) \cdots P(A_2 | A_1) P(A_1)$$

#### 3. Law of total probability and Bayesian formula

#### 4. 独立性

### 3.2 随机变量与多维随机变量

### 3.3 期望与方差

### 3.4 参数的估计

## Unit 1 Introduction

SIR模型

## Unit 2 Review basic

<https://zhuanlan.zhihu.com/p/51127402>



## 2.1 矩阵求导

向量和矩阵求导在ML、图像处理、optimization中都是非常重要的。比如说，在多元线性回归中损失函数是一个标量，每个输入都有多个属性，计算权重 $w$ 时就需要用到标量对向量的求导。在计算神经网络梯度时，相比于求解参数矩阵里的单独每一个元素的梯度，使用向量和矩阵求导能极大增加效率。要去看matrix cookbook

## 2.2 Jacobian matrix

Jacobian matrix 可作为矩阵求导的基本模块

## 2.3 Hessian Matrix

<https://machinelearningmastery.com/a-gentle-introduction-to-hessian-matrices/>

Why Is The Hessian Matrix Important In Machine Learning?

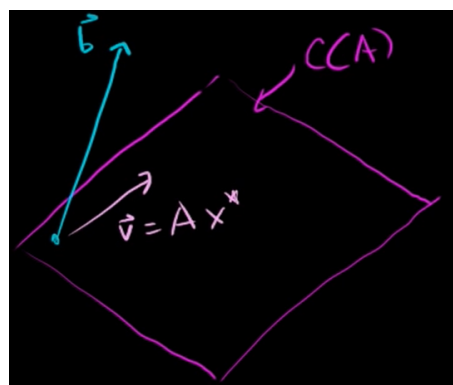
The Hessian matrix plays an important role in many machine learning algorithms, which involve optimizing a given function. While it may be expensive to compute, it holds some key information about the function being optimized. It can help determine the saddle points, and the local extremum of a function. It is used extensively in training neural networks and deep learning architectures.

## 2.4 Least square problem

Least square problem就是说给出很多点，其实我们要找出一条线，使得所有点到这条线的距离的总和最短，其实就是regression啦。比方说，现在有三个点，它们是 $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ 。然后我们要求一条线，使得所有点到这条线的距离总和最短，那不就是求Least square了嘛。于是我们就有 $Cx_1 + D = y_1, Cx_2 + D = y_2, Cx_3 + D = y_3$ 。写成矩阵就是

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

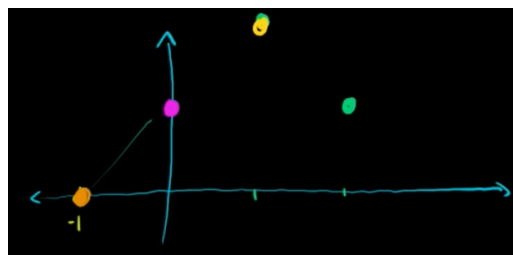
就是这种情况，我们把上述矩阵通常是写成 $Ax = b$ 。这种情况通常是无解的，因为方程式比未知参数还多。这里方程是有三个，未知参数有2个，那就是C和D。如果是有很多点，那么方程式就更多了。所以，我们要找出C和D，尽可能地去缩小这个距离。如果换成是矩阵描述的话，如下图，我们其实是希望，找到一个solution  $x^*$ ，使得 $Ax$ 更加靠近 $b$ 。



那么怎么求这个 $x^*$ 呢？有两个方法，一个是通过linear algebra，另一个是通过calculus。这里呢，先用linear algebra很简单，用公式就行了，不用知道是怎么来的，套公式，这其实也是machine learning那里学过的。公式就是

$$A^T A x^* = A^T b$$

举个例子来看，假设我们现在有4个点，分别是 $(-1,0), (0,1), (1,2), (2,1)$ 。visualize 这四个点，我们仍旧是要找到一条线，使得所有点到这条线的距离总和最短



我们假设有一条线，就是 $y = mx + b$ 。于是乎，我们有

$$\begin{aligned}
 y &= f(x) = mx + b \\
 y_1 &= f(-1) = -m + b = 0 \\
 y_2 &= f(0) = b = 1 \\
 y_3 &= f(1) = m + b = 2 \\
 y_4 &= f(2) = 2m + b = 1
 \end{aligned}$$

我们可以写成矩阵的形式就是

$$\underbrace{\begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}}_A \underbrace{\begin{bmatrix} m^* \\ b^* \end{bmatrix}}_{x^*} = \underbrace{\begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix}}_b$$

于是我们可以用公式  $A^T A x^* = A^T b$  来求

$$A^T A = \underbrace{\begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}}_{A^T} \underbrace{\begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}}_A = \begin{bmatrix} 6 & 2 \\ 2 & 4 \end{bmatrix}$$

然后我们现在再来求  $A^T b$  结果如下

$$A^T b = \underbrace{\begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}}_{A^T} \underbrace{\begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix}}_b = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$$

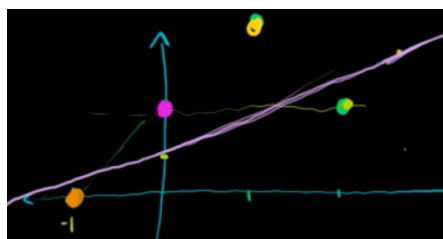
于是乎我们就可以得到

$$\begin{aligned}
 \begin{bmatrix} 6 & 2 \\ 2 & 4 \end{bmatrix} \underbrace{\begin{bmatrix} m^* \\ b^* \end{bmatrix}}_{x^*} &= \begin{bmatrix} 4 \\ 4 \end{bmatrix} \\
 \downarrow \\
 6m^* + 2b^* &= 4 \\
 2m^* + 4b^* &= 4
 \end{aligned}$$

这样我们就可以求出  $m^* = \frac{2}{5}, b^* = \frac{4}{5}$ . 于是, 我们得到的线就是

$$y = \frac{2}{5}x + \frac{4}{5}$$

画上去大概就是这么一条线, 如下图所示



现在呢, 来看看calculus的方法于是, 我们现在有  $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ . 于是我们就要 minimize  $\|Ax - b\|_2^2$ . 那怎么最小化呢? 很简单, 用公式就行了, 不用知道是怎么来的, 套公式, 这其实也是在machine learning那里学过的。这里我们提一下transpose的一个性质, 就是  $(AB)^T = B^T A^T$ , 所以公式推导就是

$$\begin{aligned}
 f(x) &= \|Ax - b\|_2^2 = (Ax - b)^T (Ax - b) \\
 &= x^T A^T A x - x^T A^T b - b^T A x + b^T b \\
 &= x^T A^T A x - 2b^T A x + b^T b
 \end{aligned}$$

这里呢,  $x^T A^T b = b^T A x$ . 假设,  $x \in \mathbb{R}^n$ , 也就是一个  $n \times 1$  的列向量, 然后  $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ . 所以  $x^T A^T b$  的最后矩阵维度是  $1 \times 1$ ,  $b^T A x$  的最后矩阵维度也是  $1 \times 1$ , 这二者是  $1 \times 1$  的互为transpose, 也就是相等。然后  $b^T b$  也是  $1 \times 1$  的矩阵, 也是常数。

因为我们要 minimize 这个error, 也就是要对  $x$  求导, 这里, 我们先说一条规则, 就是  $\frac{\partial x^T B x}{\partial x} = (B + B^T)x$ . 还有一条规则是  $\frac{\partial x^T b}{\partial x} = \frac{\partial b^T x}{\partial x} = b$  于是, 可以得出

$$\begin{aligned}
\nabla_x f(x) &= \frac{\partial}{\partial x} (x^T A^T A x - 2b^T A x + b^T b) = 0 \\
&= (A^T A + A^T A)x - 2(b^T A)^T \\
&= 2A^T A x - 2A^T b = 0 \\
&\quad \downarrow \\
x &= (A^T A)^{-1} A^T b
\end{aligned}$$

这里呢,  $A^T A$ 一定是symmetric matrix, 而且, 一定得是可逆的才行。这个公式我们要记住, 这个公式也是在ML中应用的。这里举个例子。仍旧用上面的几个点, 假设我们现在有4个点, 分别是(-1,0), (0,1), (1,2), (2,1), 那么, 我们可以得到

$$\begin{aligned}
&\underbrace{\begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}}_A, \underbrace{\begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix}}_b \\
&\quad \downarrow \\
&x = (A^T A)^{-1} A^T b \\
&= \left( \underbrace{\begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}}_{A^T} \underbrace{\begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}}_A \right)^{-1} \underbrace{\begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}}_{A^T} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 & 2 \\ 2 & 4 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{5} & -\frac{1}{10} \\ -\frac{1}{10} & \frac{3}{10} \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -\frac{3}{10} & -\frac{1}{10} & \frac{1}{10} & \frac{3}{10} \\ \frac{4}{10} & \frac{3}{10} & \frac{2}{10} & \frac{1}{10} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} \frac{2}{5} \\ \frac{4}{5} \end{bmatrix}
\end{aligned}$$

我们可以看到, linear algebra和calculus的方法, 得到的答案是一样的

### 2.4.1 Regularization

## 2.5 Log Sum Exp function

这里, 我们来明白两个概念, 分别是underflow和overflow

- Overflow: 值超过了该类型所能表示的最大值
- Underflow: 值低于该类型所能表示的最小值

在神经网络的最后, 我们会用softmax function来进行分类, 公式是

$$softmax = \frac{e^{x_i}}{\sum_{i=1}^N e^{x_i}}$$

我们知道,  $x_i \in (-\mathbb{R}, \mathbb{R})$ , 所以  $e^{x_i} \in (0, \infty)$ . 当  $e^{x_i}$  的值都为非常非常非常小的数时, 就会发生underflow, 此时,  $e^{x_i}$  和  $\sum_{i=1}^N e^{x_i}$  都会被认为是0, softmax函数就无法计算, 也就无法进行预测了。当  $e^{x_i}$  的值都为非常非常非常大的数时, 就会发生overflow, softmax也无法进行预测了, 所以, 为了解决这个underflow和overflow的问题, 就需要对softmax进行变形, 得到

$$softmax(x_i - M) = \frac{e^{x_i - M}}{\sum_{i=1}^N e^{x_i - M}}$$

这里,  $M = \max(x_1, \dots, x_n)$ . 这样变换是不会改变softmax的值的, 但是却有效地避免了overflow和underflow。因为

$$0 < e^{x_i - M} \leq e^{M - M} = 1$$

所以, 我们可以知道  $1 \leq \sum_{i=1}^N e^{x_i - M} \leq N$ , 所以, 有效地避免了overflow, 但当数据非常非常多的时候, 还是有可能造成underflow的。

因为softmax和log经常连用,  $softmax(x_i) = 0$  会导致  $\log^{softmax(x_i)} = -\infty$ .

现在, 我们来了解一下log-sum-exp function, 它长下面这样

$$f(x) = \log \sum_{i=1}^N \exp(a_i^T x + b_i)$$

这里定义我们的matrix  $A$  and bias  $b$

$$A = \begin{bmatrix} \cdots & a_1^T & \cdots \\ \cdots & a_2^T & \cdots \\ & \vdots & \\ \cdots & a_n^T & \cdots \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

如果让  $y_i = a_i^T x + b_i$ , 那么就是  $f(y) = \log \sum_{i=1}^N \exp(y_i)$ , 我们对其进行求导就是

$$\nabla_y f(y) = \begin{bmatrix} \frac{\partial f(y)}{\partial y_1} \\ \frac{\partial f(y)}{\partial y_2} \\ \vdots \\ \frac{\partial f(y)}{\partial y_n} \end{bmatrix} = \frac{1}{\sum_{i=1}^n \exp(y_i)} \begin{bmatrix} \exp(y_1) \\ \exp(y_2) \\ \vdots \\ \exp(y_n) \end{bmatrix}$$

同理, 如果不变, 那么就是matrix  $A$ , 然后对  $x$  进行求导就是, 这就是公式, 这就是一阶导

$$\nabla_x f(x) = \frac{1}{\mathbf{1}^T z} A^T z$$

这里呢,

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} \exp(a_1^T x + b_1) \\ \exp(a_2^T x + b_2) \\ \vdots \\ \exp(a_n^T x + b_n) \end{bmatrix}, \mathbf{1}^T = [1, 1, \dots, 1], \mathbf{1}^T z = \sum_{i=1}^N z_i$$

看完了一阶导, 我们再来看看二阶导

<https://www.bilibili.com/video/BV1jt411p7jE?p=12>

we got  $\nabla_x f(x) = \frac{1}{\mathbf{1}^T z} A^T z$ . We need to take the derivative again. We can get

$$\nabla_x^2 f(x) = \text{diag}(\nabla_x f(x)) - \nabla_x f(x) \nabla_x f(x)^T$$

Then we can yield

$$\nabla_x^2 f(x) = A^T \left( \frac{1}{\mathbf{1}^T z} \text{diag}(z) - \frac{1}{(\mathbf{1}^T z)^2} z z^T \right) A$$

where  $z_i = \exp(a_i^T x + b_i), i = 1, \dots, m$ .

Log Sum Exp function 是convex的。证明过程如下

If the function is convex, we need to show that for any arbitrary vector  $v$ , we have

$$v^T \nabla^2 f(x) v \geq 0$$

That is said

$$v^T \nabla^2 f(x) v = v^T A^T \left( \frac{1}{\mathbf{1}^T z} \text{diag}(z) - \frac{1}{(\mathbf{1}^T z)^2} z z^T \right) A v \geq 0$$

Since  $v$  is arbitrary, we can let  $u = Av$  as a new vector, which is also arbitrary. It follows from the Cauchy-Schwarz inequality. Therefore, the Log\_Sum\_Exp function is convex.

## 2.6 Functions

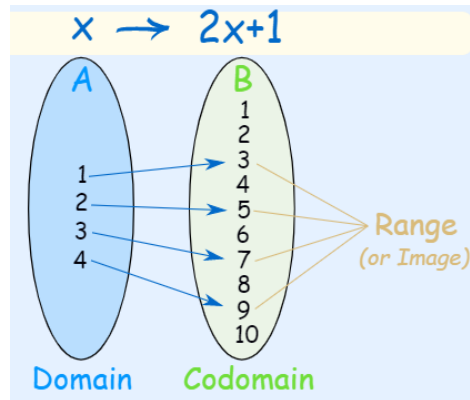
这个其实就是叫我们理解什么是function, 比如说  $f: A \rightarrow B$  代表的意思是,  $A$  是function的domain,  $B$  是function的codomain.

我们现在来理解三个概念

- domain: 假设有个函数,  $f(x)$ , domain的意思就是 $x$ 的取值范围, 比如说  $f(x) = \ln(x - 3)$  的domain就是  $(x \in \mathbb{R} | x > 3)$ .
- codomain: 函数的所有可能输出

- range: 函数的实际输出, 比如说What is the range for the function  $f(x) = e^x + 2$ ? 它的range就是  $(f(x) \in \mathbb{R} | f(x) > 2)$ .

下面这张图就是阐述以上三个概念的关系



然后还有vector, matrix的说一下, 比如说有  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  表示的是f map n-vectors to m-vectors。

## 2.7 Continuity

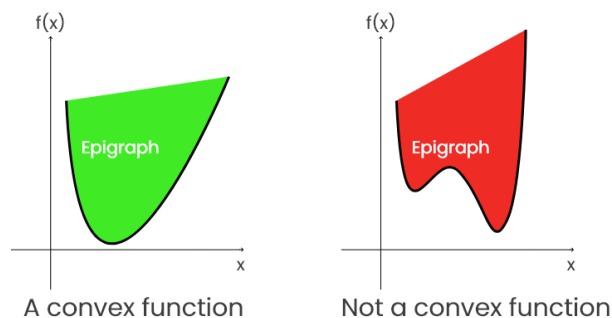
这里是判断函数是否连续

一个函数是否连续, 就得看这个函数在其domain中是不是都连续

## 2.8 Closed function

## 2.9 Epigraph

Epigraph就是函数上面的部分



# Unit 3 Unconstrained optimization

当我们在考虑optimization problem的时候, 需要考虑一下四点

- existence and uniqueness. 如果不存在还怎么优化, 如果是唯一的那就更好了, 如果不是, 还得去找
- characterization (optimality condition)
  - necessary: 什么是必要条件, 就是, 如果没有A, 就一定没有B。有A, 也不一定有B。
  - sufficient: 什么是充分条件, 就是, 如果没有A, 可能有B, 也可能没有B。有A, 就一定有B。
  - necessary and sufficient
- computation and algorithm
- heuristic

我们先理解一个概念, 就是feasible set, 意思就是符合所有约束条件的可行解

feasible problem

<https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/optimizing-multivariable-functions/a/maximums-minimums-and-saddle-points>

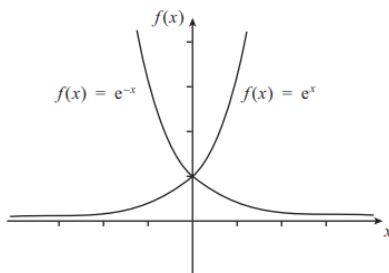
<https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/optimizing-multivariable-functions/a/second-partial-derivative-test>

<https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/quadratic-approximations/a/the-hessian>

### 3.1 Optimization

这里呢，我们的目标就是 $\min_{x \in D} f(x)$ . 就是要最小化 $f(x)$ ,  $D$ 是feasible set, 寻找最优解。

- $x^*$ 表示的是optimal point, 我们在寻找最优解的过程中, 我们先得assume  $x^*$ 的存在, 你看像 $e^{-x}$ , 它就不存在最小值。 $e^{-x}$ 的图像如下图所示



- $f^* = f(x^*)$ . 就是当我们找到最优解 $x^*$ 之后, 看看它所对应的值是啥

- jkk

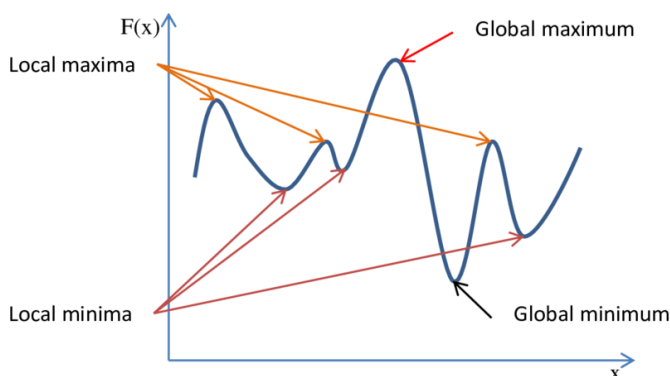
- 假设我们的optimal points有很多个

$$x_{opt} = \{x \in D; f(x) = f^*\}$$

如果 $x_{opt} \neq \emptyset$ , 就是表示找到了最优解的集, 这也就表示这个问题能够在 $x \in x_{opt}$ 中找到最优解

- Global V.S. Local optimization

如下图所示



接下来, 我们来看看是怎么定义local minimum and global minimum的

Global 可以说是local中的一个, 但是如果问题是convex的, 说明local就是global, 否则的话就不是.

我们先来看下怎么定义local minimum, 如果存在一个 $\exists \epsilon > 0$ , 使得满足所有 $|x - x^*| < \epsilon$ 的 $x$ 都有 $f(x^*) \leq f(x)$ , 我们就把点 $x^*$ 对应的函数值 $f(x^*)$ 称为一个函数 $f$ 的局部最小值。

还有一个概念是strict local minimum, 如果存在一个 $\exists \epsilon > 0$ , 使得满足所有 $|x - x^*| < \epsilon$ 的 $x$ 都有 $f(x^*) < f(x)$ , 我们就把点 $x^*$ 对应的函数值 $f(x^*)$ 称为一个函数 $f$ 的strict local minimum。

global minimum的定义, 就是, 如果 $x^*$ 对于任意的 $x$ 都满足 $f(x^*) \leq f(x)$ , 则称 $f(x^*)$ 是global minimum

<https://zhuanlan.zhihu.com/p/45028557>

<https://blog.csdn.net/u010182633/article/details/74998344>

我们先来说一下什么是feasible direction, 也就是一个定义。A vector  $d$  is said to be a feasible direction at  $x$  if there exists an  $\epsilon_0 > 0$  such that

$$x + \epsilon d \in D, \forall \epsilon \in (0, \epsilon_0)$$

then, this vector is a feasible direction. 就是说 $x$ 往某个方向移动一点点, 还落在 $D$ 中, 也就是feasible set中, 那么, 这个 $d$ 就是feasible direction。

这里来做一个关于feasible direction的练习

比如说, 一个优化问题的feasible set是

$$R = \{x : x_1 \geq 2, x_2 \geq 0\}$$

对于点

$$x_1 = \begin{bmatrix} 4 \\ 1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

现在有一些方向向量, 分别是

$$d_1 = \begin{bmatrix} -2 \\ 2 \end{bmatrix}, \quad d_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad d_3 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

现在问以上的哪些方向向量是 $x_1, x_2, x_3$ 的feasible direction?

因为 $\epsilon$ 是一个非常小的值, 我们假设 $\epsilon_0 = 1$ , 对于所有的,  $\forall \epsilon \in (0, \epsilon_0)$ .

我们先来看 $x_1$ , 它有 $x_1 + \epsilon d_1 \in R$ , 所以 $d_1$ 是 $x_1$ 的feasible direction。同理, 判断 $d_2, d_3$ , 都有

$$x_1 + \epsilon d_2 \in R, \quad x_1 + \epsilon d_3 \in R$$

所以,  $d_2, d_3$ 也是 $x_1$ 的feasible direction。也是用同样的方法来判断point  $x_2, x_3$ 。

### 3.1.1 Conditions for Local minimum

判断某个点是Local minimum的必要条件。目标函数要想有极小值, 必须满足两个条件, 也就是一阶与二阶条件

**定理:** 极小值的一阶**必要条件**, if  $x^*$  is a local minimum of  $f$  over  $D$ , then  $x^*$  satisfies gradient  $\nabla^T f(x^*)d \geq 0$  for all feasible directions  $d$  at  $x^*$ .

从这个定理中, 我们得知两个信息,

- 一个是local minimum的必要条件,
- 一个是feasible direction。

我们现在来证明这个定理, 就是如果 $x^*$  is local minimum, then  $x(\epsilon) = x^* + \epsilon d \in R, \forall \epsilon \in (0, \epsilon_0)$ .

令 $f(x(\epsilon)) = g(\epsilon)$ , by Taylor Theorem, 我们可以得到

$$g(\epsilon) = g(0) + \epsilon g'(0) + O(\epsilon)$$

这里的 $O(\epsilon)$ 的意思是 $\lim_{\epsilon \rightarrow 0} \frac{O(\epsilon)}{\epsilon} = 0$ . 是一个非常小的值。

当 $\epsilon$ 非常小的时候, 趋近于0的时候

$$g(0) = f(x(0)) = f(x^*)$$

$$g'(0) = g'(\epsilon)|_{\epsilon=0}$$

↓

$$\begin{aligned} g'(\epsilon) &= \sum_i \frac{\partial g}{\partial x_i} \frac{\partial x_i}{\partial \epsilon} \\ &= \sum_i \frac{\partial f}{\partial x_i} d_i \\ &= \nabla^T f(x) d \end{aligned}$$

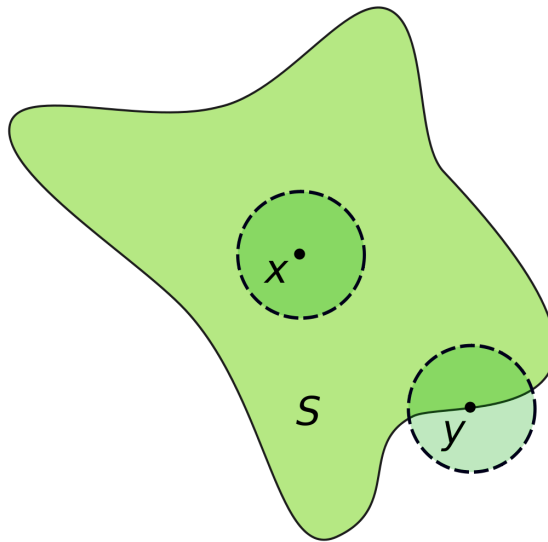
我们就可以得到  $g'(0) = \nabla^T f(x^*)d$ .

$$f(x^* + \epsilon d) = f(x^*) + \epsilon \nabla^T f(x^*)d + O(\epsilon) \geq f(x^*)$$

因为 $\epsilon \in (0, \epsilon_0)$ , 如果 $\nabla^T f(x^*)d < 0$ 的话, 就说明 $f(x^*)$ 不是极小值, 这与我们的假设矛盾, 所以 $\nabla^T f(x^*)d \geq 0$ .

所以,  $x^*$ 为极小值的必要条件就是 $\nabla^T f(x^*)d \geq 0$ .

还有一个推论(Corollary), 但是在介绍这个推论之前, 我们先讲一个概念, 即使Interior point。如下图所示,  $x$ 就是 $S$ 的interior point,  $y$ 就是 $S$ 的boundary point.



另外一个**推论**就是: if  $x^* \in \text{Int}(D)$ , then  $\nabla f(x^*) = 0$ .  $\text{Int}$ 就是Interior point的意思。

证明如下:

如果说 $x^*$ 是 $D$ 的interior point, let  $d$  be a feasible direction at  $x^* \in \text{Int}(D)$ . 从上面的定理可知,  $\nabla^T f(x^*)d \geq 0$ . 同理, 对于反方向来说, 也就是 $-d$ , 也是 $-\nabla^T f(x^*)d \geq 0$ . 所以, 我们可以得到 $\nabla f(x^*) = 0$ . 因此, 在这种情况下,  $x^*$ 是极小值的必要条件就是 $\nabla f(x^*) = 0$ . 这也就是为什么我们要求一阶导, 令其等于0的原因了。

接下来我们要讲second order sufficient condition, **二阶充分条件**, 就是

<https://zhuanlan.zhihu.com/p/45028557>

for any interior point  $x^*$ , if

- $\nabla f(x^*) = 0$
- $\nabla^2 f(x^*) > 0$ , 就是正定矩阵

Then  $x^*$  is local minimum

所以说白了就是, 一阶导, 令其等于0, 然后求出一些值, 然后再进行二阶导, 如果二阶导是一个positive definite, 那么就是local minimum, 如果是negative definite, 那么就是local maximum

这里, 我们来看一个**example**, 就是 $f(x) = x_1^2 - x_1x_2 + x_2^2 - 3x_2$ , find the local minimum. Is this local minimum also a global minimum?

我们先求一阶导, 令其等于0, 就是

$$\begin{aligned}\nabla_{x_1} f(x) &= 2x_1 - x_2 = 0 \\ \nabla_{x_2} f(x) &= -x_1 + 2x_2 - 3 = 0 \\ &\downarrow \\ \begin{bmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - 3 \end{bmatrix} &= 0 \\ &\downarrow \\ x_1 = 1, x_2 = 2\end{aligned}$$

接下来我们要求二阶导, 看其是否为positive definite

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

然后这是一个positive definite, 所以 $x_1 = 1, x_2 = 2$ 是一个local minimum。但是, 与此同时, 它是不是也是global minimum呢?

这里我们先说一个实对称矩阵的相关**性质**,  $\lambda_{min}$ 就是 $A$ ,也可以说是 $\nabla^2 f(x)$ 的最小特征值,  $\lambda_{max}$ 就是最大特征值, 这个记住就好了

$$\forall y \in \mathbb{R}^n, \lambda_{min} \|y\|^2 \leq y^T A y \leq \lambda_{max} \|y\|^2$$

于是, 我们可以把 $f(x) = x_1^2 - x_1x_2 + x_2^2 - 3x_2$ 中的 $x_1^2 - x_1x_2 + x_2^2$ 写成quadratic form, 就是 $f(x) = x^T A x - 3x_2$ , 我们可以得到



$$A = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} > 0$$

我们可以知道，这里的 $A$ 是一个positive definite。而且，是一个对称矩阵， $A = A^T$ 。根据刚刚的性质，我们可以得到

$$f(x) = x^T A x - 3x_2 \geq \lambda_{\min} \|x\|^2 - 3x_2$$

$$\downarrow$$

$$\lambda_{\min}(x_1^2 + x_2^2) - 3x_2$$

根据positive definite的性质，它的特征值都是恒大于0的。因为

$$\lambda_{\min}(A) > 0$$

所以， $f(x) \rightarrow +\infty$ ，所以 $\|x\| \rightarrow +\infty$ 。所以，这也是一个global minimum， $f(x) - f(x^*) \geq 0$ 。

我们再来看另外一个example

$$J(x) = (x_1 - x_2^2)(x_1 - 2x_2^2), \text{ find the local minimum if any.}$$

我们先把它展开，可以得到， $J(x) = x_1^2 - 2x_1x_2^2 - x_1x_2^4 + 2x_2^4$ 。然后分别对 $x_1, x_2$ 求一阶导，可以得到

$$\nabla_{x_1} J(x) = 2x_1 - 3x_2^2$$

$$\nabla_{x_2} J(x) = -6x_1x_2 + 8x_2^3$$

令其等于0，我们可以得到 $(x_1, x_2) = (0, 0)$ 。

然后，我们再进行二阶求导，可以得到

$$H(x) = \nabla^2 J(x) = \begin{bmatrix} 2 & -6x_2 \\ -6x_2 & -6x_1 + 24x_2^2 \end{bmatrix} \Big|_{(x_1, x_2) = (0, 0)} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

然后我们看 $H(x)$ 的leading principal minors，可以知道 $H(x) \geq 0$ ，所以它是positive semi-definite的。所以，我们是不知道它是不是local minimum的。sufficient condition是说要大于0，这个是大于等于0，所以无法判断。

现在来看另外一个condition，**second order necessary condition**，就是

If  $x^*$  is a local minimum, then

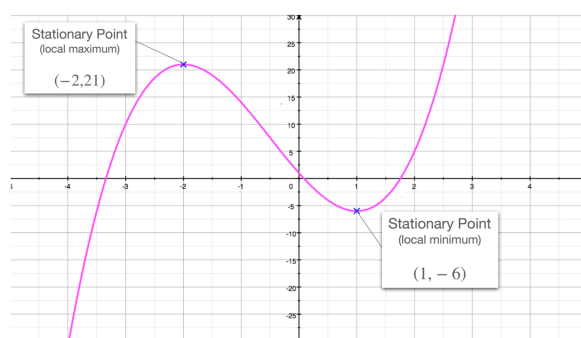
- $\nabla f(x^*) = 0$
- $\nabla^2 f(x^*) \geq 0$ . A positive semi-definite

necessary condition，是只能正推，不能反推。我们来看一个例子，比如说 $y = x^3$ ，当我们求出一阶导等于0时，可以得到 $x = 0$ 。但是在 $x = 0$ 那个地方并不是local minimum

我们来看一个example

$$J(x) = \frac{1}{2}x^T Q x + b^T x + C, \text{ find stationary points and determine conditions so that it is a local minimum.}$$

这里先说一些什么是stationary point，就是一阶导为0的地方，就是stationary point，如下图所示



接下来，我们来看看怎么解这道题吧。先进行求一阶导

$$\nabla_x J(x) = Qx + b = 0$$

$$\downarrow$$

$$x^* = -Q^{-1}b$$

assume  $Q$  is non-singular

然后进行求二阶导，可以得到 $H(x) = Q$ , if  $Q > 0$ , a positive definite, then  $x^*$  is local minimum.

## Unit 4 Convex optimization

$\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$ . non negative real numbers.

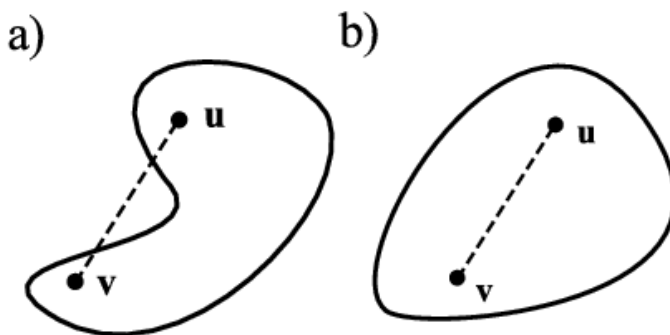
$\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$ . positive real

## 4.1 Convex set

先来看看什么是Affine set? 就是现在有两个点, 分别是 $x_1 \neq x_2 \in \mathbb{R}^n$ , 现在存在一个 $\lambda \in \mathbb{R}$ . 于是 $x_1, x_2$ 之间就可以连成一条线, 那么这之间的集合的表达就是 $\lambda x_1 + (1 - \lambda)x_2$ . 这条线段就是的集合就是Affine set. 这里需要限制 $\lambda \in [0, 1]$ . 要不然就跑出 $x_1, x_2$ 之间的范围了。

再来讲讲什么是convex set?

A set  $C$  is said to be convex set if for every  $x \in C, y \in C, \lambda \in [0, 1]$ , then  $z \triangleq \lambda x + (1 - \lambda)y \in C$ . 也就是说对于任意的 $x, y \in C$  与任意的 $\lambda \in [0, 1]$  有 $\lambda x + (1 - \lambda)y \in C$ . 也就是集合中的两点连成线仍属于集合。我们看下面的图。a) 就不是convex set, b) 就是convex set。



**Convexity preserving**, 保留凸性。

- suppose that  $C, D$  are convex set, then  $C \cap D$  is convex set too.
- non-negative weighted sum:  $w_i \geq 0, \sum_{i=1}^k w_i f_i$  is convex.
- composition with monotone convex function.  $g(f(x))$  is convex if and only if  $f$  is convex,  $g$  is convex and non-decreasing.

接下来就是证明这个定理

因为function  $f$  是convex的, for some  $\lambda \in [0, 1]$ . 所以我们可以得到

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

又因为 $g$ 是increasing的, 所以我们可以得到

$$g(f(\lambda x + (1 - \lambda)y)) \leq g(\lambda f(x) + (1 - \lambda)f(y))$$

又因为 $g$ 也是convex的, 所以我们可以得到

$$g(\lambda f(x) + (1 - \lambda)f(y)) \leq \lambda g(f(x)) + (1 - \lambda)g(f(y))$$

最终, 我们是可以得到

$$g(f(\lambda x + (1 - \lambda)y)) \leq \lambda g(f(x)) + (1 - \lambda)g(f(y))$$

for some  $\lambda \in [0, 1]$ , 因此呢,  $g(f(x))$  is convex if and only if  $f$  is convex,  $g$  is convex and non-decreasing.

- Affine transformation

If  $C \subseteq \mathbb{R}^n$  is a convex set,  $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ , then  $AC + b = \{Ax + b | x \in C\} \subseteq \mathbb{R}^m$  is also convex.

这里举个例子来说, 假设function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is convex, 那么

$$f(x) = \|Ax - b\|$$

where  $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ , and  $\|\cdot\|$  is a norm on  $\mathbb{R}^m$ .

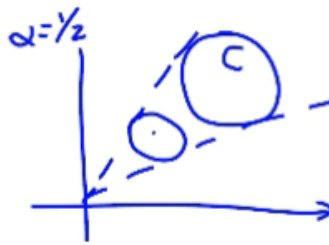
这种情况下,  $f(x)$ 也是convex的。因为norm本身就是convex的, 然后又是affine transformation, 所以就是convex的。

- translation

$C+b$ , is also convex

- scaling

$\alpha C$ , here,  $\alpha$  is a real number. 放大缩小如下图所示



- sum of sets

$C_1, C_2$  are convex sets, So,  $c_1 + c_2 = \{C_1 + C_2 | c_1 \in C_1, c_2 \in C_2\}$ .

常见的convex set

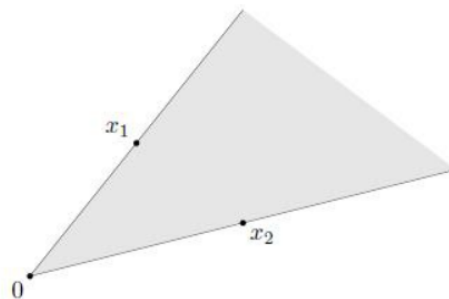
- $\mathbb{R}^n$  space
- subspace of  $\mathbb{R}^n$  space
- 任意直线、线段
- 球、椭球、多面体

convex set的性质

现在来看看**convex cone**

这是cone的定义,  $C$  是cone, 然后有  $\forall x \in C, \theta \geq 0$ , 有  $\theta x \in C$ .

这是convex cone的定义,  $\forall x_1, x_2 \in C, \lambda_1, \lambda_2 \geq 0$ , 有  $\lambda_1 x_1 + \lambda_2 x_2 \in C$ . 这里跟convex set不一样的点是, 不要求  $\lambda_1 + \lambda_2 = 1$ , 只需要是非负数即可。凸锥从几何上来说, 可以看作是从原点出发, 经过点  $x_1$  和  $x_2$  的平面 (直线) 围成的区域, 凸锥的重要性质之一是它一定包含**原点**。如下图所示

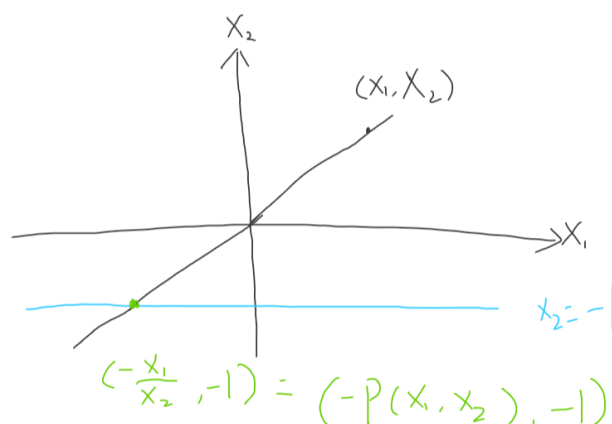


这里还要讲一个**perspective function**

假设有个函数  $P: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ , 这个function是从  $n+1$  维的空间映射到  $n$  维的空间, 也就相当于是做个降维。这个函数定义域  $\text{dom}$  为  $\mathbb{R}^n \times \mathbb{R}_{++}$ . 这个定义域的意思就是前  $n$  维中可以在  $\mathbb{R}$  中取任意数, 但是最后一维必须是正数。现在我们来定义这个perspective function, 就是  $P(z, t) = \frac{z}{t}, z \in \mathbb{R}^n, t \in \mathbb{R}_{++}$ . 也就是说  $z$  是任意的  $n$  维向量, 但是  $t$  必须是正数。perspective function做的事情就是假设有一个  $n+1$  维的向量, 最后一维必须是正数, 然后每个数除以最后一个数, 那么最后一维的数就变成了1, 我们把那个1给舍弃掉, 就变成了  $n$  维向量, 也就起到了降维的作用。公式的描述就是

$$P(y) = P(x_1, x_2, \dots, x_n, t) = \left( \frac{x_1}{t}, \dots, \frac{x_n}{t} \right) \in \mathbb{R}^n$$

这里的  $y$  呢就是  $y = (x, t), x \in \mathbb{R}^n, t \in \mathbb{R}_{++}$ . 然后这个  $y$  也是在convex set的一个element. perspective function我们可以用小孔成像这个例子来理解。如下图所示



相当于是说  $(x_1, x_2)$  经过原点的透射, 变成  $-\frac{x_1}{x_2}$ , 从二维变一维。大概就这么理解

任意一个convex set, 经过perspective function之后, 仍旧是convex set.

假设  $C \subseteq \mathbb{R}^n \times \mathbb{R}_{++}$ , and suppose that  $C$  is a convex set. Show that  $P(C) \in \mathbb{R}^n$  is also convex.  $P$  is perspective function.

以下是证明

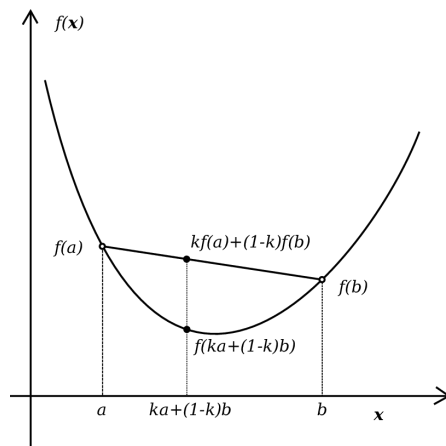
## 4.2 Convex function

我们先来看看什么是convex function。首先,  $C \subseteq \mathbb{R}^n$ .  $C$ 首先得是convex set.

### Definition

A function is convex if  $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$  for every  $x, y \in \mathbb{R}^n$  and  $\lambda \in [0, 1]$ . 我们先来看看图像是怎样的

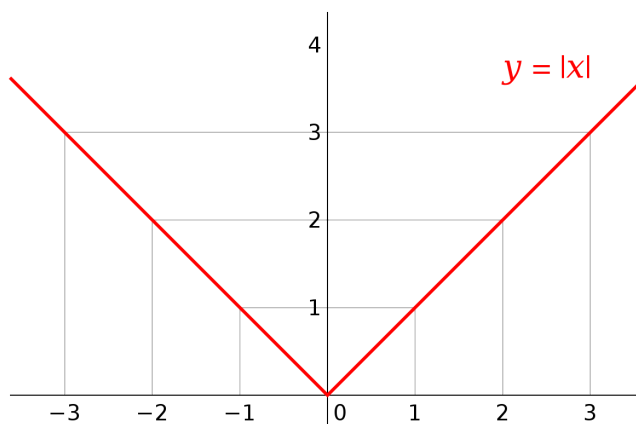
图片中的k就是 $\lambda$ .



convex function 一定是连续的

此外, 我们还得讲一讲strictly convex, 就是A function is strictly convex if  $f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y)$  for every  $x, y \in \mathbb{R}^n$  and  $\lambda \in (0, 1)$ .

不管是strictly convex还是就普通的convex, 都没有考虑这个function是否differentiable。比如说, 下面这个函数, 在 $x=0$ 处不可导, 但是它是convex的。



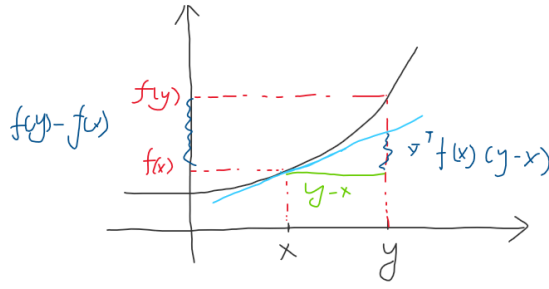
现在来看看一些常见的凸函数

- 线性函数:  $f(x) = a^T x + b$
- 二次函数:  $f(x) = x^T Q x + a^T x + b$ ,  $Q \in S_+^n$ . 这里呢,  $S_+^n$ 是指positive semi-definite的意思.
- 最小二乘函数:  $f(x) = \|Ax - b\|_2^2$ .
- p-norm:  $f(x) = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ ,  $p \geq 1$ .
- exponential function:  $f(x) = e^{ax}$ ,  $a \in \mathbb{R}$ .
- power function:  $f(x) = x^a$ ,  $x \in \mathbb{R}_{++}$ . 当  $a \geq 1$  or  $a \leq 0$  is convex. otherwise concave.
- negative log function:  $f(x) = -\log(x)$
- $f(x) = x \log(x)$  is convex.
- log-sum-exp function:  $f(x) = \log(e^{x_1} + e^{x_2} + \dots + e^{x_n})$ ,  $x \in \mathbb{R}^n$ .

现在我们来讲两个Theorems, 非常重要

1. Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  be a differentiable function,  $f$  is convex if and only if  $\nabla f(x)(y - x) \leq f(y) - f(x)$ ,  $\forall x, y \in \mathbb{R}^n$ .

如下图所示，就是一阶可导的情况下



现在来看看怎么证明，因为我们要证明这个函数是convex的，所以需要证明

$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ . 我们已知的条件就是  $\nabla f(x)(y - x) \leq f(y) - f(x), \forall x, y \in \mathbb{R}^n$ .

$$\nabla^T f(z)(y - z) \leq f(y) - f(z)$$

$$\nabla^T f(z)(x - z) \leq f(x) - f(z)$$

我们让第一行的式子乘以  $\lambda$ , 第二行的式子乘以  $(1 - \lambda)$ . 然后, 我们可以得到

$$\lambda \nabla^T f(z)(y - z) + (1 - \lambda) \nabla^T f(z)(x - z) \leq \lambda(f(y) - f(z)) + (1 - \lambda)(f(x) - f(z))$$

↓

$$\nabla^T f(z)(\lambda(y - z) + (1 - \lambda)(x - z)) \leq \lambda f(y) - \lambda f(z) + f(x) - f(z) - \lambda f(x) + \lambda f(z)$$

↓

$$\nabla^T f(z)(\lambda y + (1 - \lambda)x - z) \leq \lambda f(y) + (1 - \lambda)f(x) - f(z)$$

此时, 我们让  $z = \lambda y + (1 - \lambda)x$ , 于是, 我们可以得到

$$0 \leq \lambda f(y) + (1 - \lambda)f(x) - f(z)$$

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

得证。

2. For  $f$  that is twice differentiable,  $f$  is convex if and only if the Hessian is positive semi-definite. 也就是说, 如果  $f$  是二阶可导的话, 那么, 只要它的 Hessian matrix 是 positive semi-definite 的, 那么,  $f$  就是 convex 的。

这个证明涉及到 Taylor expansion, 比较复杂。我们来看个例子吧

#### Example

$$f(x_1, x_2, x_3) = 2x_1^2 + x_1x_3 + x_2^2 + 2x_2x_3 + \frac{1}{2}x_3^2. \text{ 问是否为convex?}$$

我们先求它的一阶导, 就是

$$\nabla_{x_1} f(x_1, x_2, x_3) = 4x_1 + x_3.$$

$$\nabla_{x_2} f(x_1, x_2, x_3) = 2x_2 + 2x_3.$$

$$\nabla_{x_3} f(x_1, x_2, x_3) = x_1 + 2x_2 + x_3.$$

然后再求 Hessian matrix,

$$\nabla^2 f(x_1, x_2, x_3) = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

We can know that it is not positive semi-definite. Therefore, it is not convex and not concave either.

这里总结下结论

- 如果 Hessian 是 positive semi-definite, 那么就是 convex
- 如果 Hessian 是 negative semi-definite, 那么就是 concave
- 如果 Hessian 是 positive definite, 那么就是 strictly convex
- 如果 Hessian 是 negative definite, 那么就是 strictly concave
- 如果都不是, 那么就既不是 convex, 也不是 concave.

以上的结论都要好好记住。

接下来我们会看一堆的例子

### 4.2.1 Examples

#### Example 1

Let  $f$  be a convex function over a convex set  $C$ , given  $r \in \mathbb{R}$ ,  $S = \{x \in C : f(x) \leq r\}$ , show that  $S$  is a convex set.

要证明 convex set, 我们就要记住一个 convex set 的结论, 就是

$$\lambda x + (1 - \lambda)y \in C.$$

我们先来看看怎么证明

$$x_1, x_2 \in S, \lambda \in [0, 1].$$

因为  $f(x_1) \leq r, f(x_2) \leq r$ , 所以,  $x_1, x_2 \in S$ , 相当于是个交集啦。

所以  $\lambda x_1 + (1 - \lambda)x_2 \in S$ , 所以  $S$  is a convex set.

### Example 2

$S = \{x \in \mathbb{R}^2 : x_1 > 0, x_2 > 0 \text{ and } x_1 \log x_1 + x_2 \log x_2 \leq 2\}$ , 证明  $S$  是 convex set.

看到有 function 的情况下, 先证明这个 function 是 convex function, 也就是求 Hessian。而且, 我们看到这里有 and, 说明得是交集。

也就是说  $S = S_1 \cap S_2$ .

令  $f(x) = x_1 \log x_1 + x_2 \log x_2$ .

$$\nabla f(x) = \begin{bmatrix} 1 + \log x_1 \\ 1 + \log x_2 \end{bmatrix}, \quad \nabla^2 f(x) = \begin{bmatrix} x_1^{-1} & 0 \\ 0 & x_2^{-1} \end{bmatrix} > 0, \forall x \in C$$

所以,  $S = \{x \in C : x_1 \log x_1 + x_2 \log x_2 \leq 2\}$ , where  $C = \{x \in \mathbb{R}^2 : x_1 > 0, x_2 > 0\}$

所以,  $S$  is a convex set.

### Example 3

Let  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}, (i = 1, \dots, k)$  convex, 也就说这里有  $k$  个 function, 都是 convex 的。show that  $g(x) = \max_{i \in \{1, \dots, k\}} f_i(x)$  is convex.



这个  $g(x)$  就是这么个意思。其实, 这个函数的定义域是  $\text{dom } f = \text{dom } f_1 \cap \text{dom } f_2$ . 是一个交集, 所以说这个 function 的定义域是 convex set.

关于 convex function 的判断, 就是  $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ .

有

$$\begin{aligned} g(\lambda x_1 + (1 - \lambda)x_2) &= \max_{i \in \{1, \dots, k\}} f_i(\lambda x_1 + (1 - \lambda)x_2) \\ &\leq \max_{i \in \{1, \dots, k\}} [\lambda f_i(x_1) + (1 - \lambda)f_i(x_2)] \\ &\leq \max_{i \in \{1, \dots, k\}} [\lambda f_i(x_1)] + \max_{i \in \{1, \dots, k\}} [(1 - \lambda)f_i(x_2)] \\ &= \lambda g(x_1) + (1 - \lambda)g(x_2) \end{aligned}$$

上述过程的第二行是因为  $f$  是一个 convex function。

于是, 得证。  $g(x)$  is convex.

### Example 4

Define  $g : \mathbb{R}^n \rightarrow \mathbb{R}, g(x) = \text{sum of the } r \text{ largest elements of } x$ . Show  $g$  is a convex function. 就是前  $r$  大的数

$$X = \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix}$$

Let  $S$  be any subset of size  $r$  of  $\{1, 2, \dots, n\}$ .

$f_S(x) = \sum_{j \in S} x_j$  is convex function

$g(x) = \max_S f_S(x)$ , 然后这个是max over convex function。所以 $g$ 是convex function。这相当于是convexity preserving operation

<https://www.bilibili.com/video/BV1jt411p7jE?p=13>, 证明过程在38分钟左右。

### Example 5

The epigraph  $\text{epi}(f)$  of a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is a subset of  $\mathbb{R}^{n+1}$  defined as

$$\text{epi}(f) = \{(x, t) : x \in \mathbb{R}^n, f(x) \leq t\}$$

show that a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is convex if and only if its epigraph is a convex set.

## 4.3 Convex optimization

**Affine function**, 即最高次数为1的多项式函数。常数项为零的仿射函数称为线性函数。

the convex problem has three additional requirements:

- the objective function must be convex
- the inequality constraint functions must be convex
- the equality constraint functions  $h_i(x) = a_i^T x - b_i$  must be affine

**KKT条件** (Karush-Kuhn-Tucker Conditions), 它是在满足一些有规则的条件下, 一个非线性规划问题能有最优化解法的一个必要和充分条件

[https://www.bilibili.com/video/BV1Qt411f79U?from=search&seid=8598333746733307293&spm\\_id\\_from=333.337.0.0](https://www.bilibili.com/video/BV1Qt411f79U?from=search&seid=8598333746733307293&spm_id_from=333.337.0.0)

[https://www.bilibili.com/video/BV1HP4y1Y79e?from=search&seid=8598333746733307293&spm\\_id\\_from=333.337.0.0](https://www.bilibili.com/video/BV1HP4y1Y79e?from=search&seid=8598333746733307293&spm_id_from=333.337.0.0)

### Theorem

现在有一个convex problem (CP)

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, i = 1, \dots, m \\ & i \in \mathcal{P} = \{1, \dots, m\} \end{aligned}$$

我们现在assume 这里所有的function都是differentiable的, 然后也是符合slater's condition的. Then, a feasible point  $x^*$  是最优解if and only if 符合KKT 条件, 也就是

$$\begin{aligned} \nabla f(x^*) + \sum_{i \in \mathcal{P}} \mu_i \nabla f_i(x^*) &= 0 \\ \mu_i f_i(x^*) &= 0, \forall i \in \mathcal{P} \\ \mu_i &\geq 0 \end{aligned}$$

我们现在来看看如何将这个CP Lagrangian, 也就是拉格朗日化。方式就是

$$\begin{aligned} L(x, \mu) &= f(x) + \sum_{i \in \mathcal{P}} \mu_i f_i(x) \\ \mu_i &\geq 0 \end{aligned}$$

然后这个 $\mu$ 就是 Lagrange multiplier. 正常来说, 我们要做的就是 $\nabla L(x^*, \mu) = 0$ .

我们来看一个例子, **example**

$$\begin{aligned} \min_{x=(x_1, x_2, x_3) \in \mathbb{R}^3} \quad & \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{s.t.} \quad & x_1 + x_2 + x_3 \leq -3 \end{aligned}$$

先来看看怎么求, 先得到Lagrangian function, 也就是

$$L(x, \lambda) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + \lambda(x_1 + x_2 + x_3 + 3)$$

And we take the derivative here  $\nabla_x L(x, \lambda) = 0$ , we can get

$$\begin{pmatrix} x_1 + \lambda \\ x_2 + \lambda \\ x_3 + \lambda \end{pmatrix} = 0 \Leftrightarrow x_1 = x_2 = x_3 = -\lambda$$

And now, we need to calculate the  $f^*$ .

$\lambda(x_1 + x_2 + x_3 + 3) = 0$ , we need to discuss the case

- $\lambda = 0, x_1 + x_2 + x_3 < -3$ . we can get  $x_1 = x_2 = x_3 = 0$ , it does not satisfy the constraint.
- $x_1 + x_2 + x_3 = -3, \lambda > 0$ . we can get  $x_1 = x_2 = x_3 = -1, \lambda = 1$ .

So, the  $x^* = (x_1^*, x_2^*, x_3^*) = (-1, -1, -1)$ . we can get  $f^* = \frac{1}{2}((-1)^2 + (-1)^2 + (-1)^2) = \frac{3}{2}$ .

这里来总结一下写KKT conditions的套路

我们的式子是

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p \end{aligned}$$

我们第一步都是先写出它的Lagrangian function的, 就是

$$L(x, \mu, \lambda) = f_0(x) + \sum_{i=1}^m \mu_i f_i(x) + \sum_{i=1}^p \lambda_i h_i(x)$$

第二步就是对 $x$ 求导

以上式子的KKT conditions就是

$$\begin{aligned} f_i(x^*) &\leq 0 \\ h_i(x^*) &= 0 \\ \mu_i &\geq 0 \\ \mu_i f_i(x) &= 0 \\ \nabla L(x, \mu, \lambda) &= 0 \end{aligned}$$

我们来看一个写KKT conditions的**example**

这个式子是

$$\begin{aligned} \min \quad & \frac{1}{2}x^T P x + q^T x + r \\ \text{s.t.} \quad & A x = b \end{aligned}$$

where  $P \in S_{+}^n$ . 然后这个式子的KKT conditions就是. 第一步, 先写Lagrangian function, 也就是

$$L(x, \lambda) = \frac{1}{2}x^T P x + q^T x + r + \lambda(Ax - b)$$

然后对 $x$ 求导, 可得

$$\nabla_x L(x, \lambda) = P x^* + q + A^T \lambda = 0$$

所以, 这个式子的KKT conditions就是

$$\begin{aligned} A x^* &= b \\ P x^* + q + A^T \lambda &= 0 \end{aligned}$$

我们接着来看**example**, 这是书上的**water-filling**问题



现在有一个CP问题，就是

$$\begin{aligned} \min \quad & -\sum_{i=1}^n \log(\alpha_i + x_i) \\ \text{s.t.} \quad & x \succeq 0 \\ & 1^T x = 1 \end{aligned}$$

首先第一步，我们仍旧是写出Lagrangian function, 这里我们要注意，要小心，它的约束，这里是大于等于0，要小心

$$L(x, \mu, \lambda) = -\sum_{i=1}^n \log(\alpha_i + x_i) - \mu_i x_i + \lambda(1^T x - 1)$$

这个式子的KKT conditions就是

$$\begin{aligned} x &\succeq 0 \quad (1) \\ 1^T x &= 1 \quad (2) \\ \mu_i &\geq 0 \quad (3) \\ \mu_i x_i &= 0, i = 1, \dots, n \quad (4) \\ \nabla L(x, \mu, \lambda) &= -\frac{1}{\alpha_i + x_i} - \mu_i + \lambda = 0, i = 1, \dots, n \quad (5) \end{aligned}$$

从第五个列式我们可以得到， $\mu_i = \lambda - \frac{1}{\alpha_i + x_i}$ . 于是，我们就可以删去 $\mu_i$ , 我们的KKT conditions就变成了

$$\begin{aligned} x &\succeq 0 \quad (1) \\ 1^T x &= 1 \quad (2) \\ x_i(\lambda - \frac{1}{\alpha_i + x_i}) &= 0, i = 1, \dots, n \quad (4) \\ \lambda &\geq \frac{1}{\alpha_i + x_i}, i = 1, \dots, n \quad (5) \end{aligned}$$

这里呢，我们就分情况讨论

- 如果 $\lambda \geq \frac{1}{\alpha_i}$ , 那么 $\lambda - \frac{1}{\alpha_i + x_i} > 0$ , 也就有 $x_i^* = 0$ .
- 如果 $\lambda < \frac{1}{\alpha_i}$ , 那么，我们就得让 $\lambda - \frac{1}{\alpha_i + x_i} = 0$ , 我们可以求出 $x_i^* = \frac{1}{\lambda} - \alpha_i$ .

我们来看另外一个example

$$\begin{aligned} \min_x \quad & x_1^2 + 2x_2^2 \\ \text{s.t.} \quad & -x_1 - x_2 + 3 \leq 0 \\ & -x_2 + x_1^2 + 1 \leq 0 \end{aligned}$$

第一步，就要先把Lagrangian function给写出来

$$L(x_1, x_2, \mu_1, \mu_2) = x_1^2 + 2x_2^2 + \mu_1(-x_1 - x_2 + 3) + \mu_2(-x_2 + x_1^2 + 1)$$

此时，我们就可以把它的KKT conditions给写出来

$$\begin{aligned} \nabla_{x_1} L(x_1, x_2, \mu_1, \mu_2) &= 2x_1 - \mu_1 + 2\mu_2 x_1 = 0 \quad (1) \\ \nabla_{x_2} L(x_1, x_2, \mu_1, \mu_2) &= 4x_2 - \mu_1 - \mu_2 = 0 \quad (2) \\ \mu_1(x_1 + x_2 - 3) &= 0 \quad (3) \\ \mu_2(x_2 - x_1^2 - 1) &= 0 \quad (4) \\ \mu_1, \mu_2 &\geq 0 \quad (5) \\ -x_1 - x_2 + 3 &\leq 0 \quad (6) \\ -x_2 + x_1^2 + 1 &\leq 0 \quad (7) \end{aligned}$$

接下来我们要求 $f^*$ . 我们就要分情况讨论

case 1:  $\mu_1 = 0, \mu_2 = 0$

这种情况下，我们可以得到 $2x_1 = 0, 4x_2 = 0$ , 从而推出 $x_1 = 0, x_2 = 0$ . 但是这个违反了KKT条件的(6), (7). 所以，这个是不可取的。

case 2:  $\mu_1 \geq 0, \mu_2 = 0$

这种情况下，我们可以得到

$$\begin{aligned} 2x_1 - \mu_1 &= 0 \quad (1) \\ 4x_2 - \mu_1 &= 0 \quad (2) \\ x_1 + x_2 - 3 &= 0 \quad (3) \end{aligned}$$

于是, 我们可以求出  $x_1 = 2, x_2 = 1, \mu_1 = 4$ . 但是, 这里违反了KKT条件的(7).

case 3:  $\mu_1 = 0, \mu_2 \geq 0$

这种情况下, 我们可以得到

$$\begin{aligned} 2x_1 + 2\mu_2 x_1 &= 0 \quad (1) \\ 4x_2 - \mu_2 &= 0 \quad (2) \\ x_2 - x_1^2 - 1 &= 0 \quad (3) \end{aligned}$$

于是, 我们可以得到  $x_1 = 0, x_2 = 1, \mu_2 = 4$ . 但是, 这里违反了KKT条件的(6).

case 4:  $\mu_1, \mu_2 > 0$

这种情况下, 我们可以得到

$$\begin{aligned} -x_1 - x_2 + 3 &= 0 \quad (1) \\ -x_2 + x_1^2 + 1 &= 0 \quad (2) \end{aligned}$$

我们可以得到  $x_1 = -2, x_2 = 5$  或者是  $x_1 = 1, x_2 = 2$ .

当  $x_1 = -2, x_2 = 5$ , 可以求出  $\mu_1 = 28, \mu_2 = -8$ . 当  $x_1 = 1, x_2 = 2$ , 可以求出  $\mu_1 = 6, \mu_2 = 2$ .

此时, 我们要求  $f^*$ , 而且是最小值, 我们可以得到  $f^* = 9$ .

刚刚的优化中, 只有不等式约束, 我们可能还会有等式约束, 就是  $h_j(x) = 0, j = 1, \dots, n$ .

也就是说我们现在变成这样的了

$$\begin{aligned} \min \quad & f_o(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, n \end{aligned}$$

我们要秒写出它的Lagrangian function, 也就是

$$L(x, \mu, \lambda) = f_o(x) + \sum_{i=1}^m \mu_i f_i(x) + \sum_{i=1}^n \lambda_i h_i(x)$$

不等式约束的  $\mu_i \geq 0$ , 但是等式约束的  $\lambda_i$  是没有符号限制的。然后这里的  $\mu, \lambda$  就是Lagrange multiplier.

下一个要讲的topic是**Duality**

在我们写出来Lagrangian function之后, 就要顺手写出Lagrange dual function或者直接就叫dual function。很简单, 就是

$$d(\mu, \lambda) = \inf_{x \in D} L(x, \mu, \lambda)$$

这里的  $D$  指的是  $D = \cap_{i=1}^m \text{dom} f_i \cap \cap_{i=1}^n \text{dom} h_i$ .

一般来说, 这里就要求  $d^*$  了, 求法就是

$$d^* = \max_{\mu, \lambda} d(\mu, \lambda)$$

这里提两个性质

- dual function为concave function
- $\forall \lambda \geq 0, \forall \mu, g(\lambda, \mu) \leq f^*$ .

证明:

设  $x^*$  是原问题最优解, 则有  $f_i(x^*) \leq 0, h_i(x^*) = 0$ , 必然满足约束

当  $\forall \lambda \geq 0, \forall \mu$ , 我们有

$$\sum_{i=1}^m \mu_i f_i(x^*) + \sum_{i=1}^n \lambda_i h_i(x^*) \leq 0$$

因为  $\sum_{i=1}^m \mu_i f_i(x^*) \leq 0$ , 所以总体小于等于0.

$$L(x, \mu, \lambda) = f_o(x) + \sum_{i=1}^m \mu_i f_i(x) + \sum_{i=1}^n \lambda_i h_i(x) \leq f^*$$

又因为dual problem是最小化，所以必然有

$$d(\mu, \lambda) \leq f^*.$$

这里来看一个example

$$\begin{aligned} \min_{x=(x_1, x_2, x_3) \in \mathbb{R}^3} \quad & \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{s.t.} \quad & x_1 + x_2 + x_3 \leq -3 \end{aligned}$$

我们立马写出Lagrangian function，也就是

$$L(x, \lambda) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + \lambda(x_1 + x_2 + x_3 + 3)$$

然后这个dual problem就是,这是不等式约束，需要有 $\lambda \geq 0$

$$\begin{aligned} d(\lambda) &= \min_x L(x, \lambda) \\ &= \min_x \left( \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + \lambda(x_1 + x_2 + x_3 + 3) \right) \end{aligned}$$

And we take the derivative here  $\nabla_x L(x, \lambda) = 0$ , we can get

$$\begin{pmatrix} x_1 + \lambda \\ x_2 + \lambda \\ x_3 + \lambda \end{pmatrix} = 0 \Leftrightarrow x_1 = x_2 = x_3 = -\lambda$$

于是，我们可以得到

$$\begin{aligned} d(\lambda) &= \frac{1}{2}[(-\lambda)^2 + (-\lambda)^2 + (-\lambda)^2] + \lambda((- \lambda) + (-\lambda) + (-\lambda) + 3) \\ &= -\frac{3}{2}\lambda^2 + 3\lambda \end{aligned}$$

此时，我们就要去求 $d^*$ ，也就是 $\max_{\lambda \geq 0} d(\lambda)$ 。于是有

$$\max_{\lambda \geq 0} -\frac{3}{2}\lambda^2 + 3\lambda$$

求导，令其等于0，可以得到 $-3\lambda + 3 = 0 \Leftrightarrow \lambda = 1 \Rightarrow x_1 = x_2 = x_3 = -1$ 。

我们可以得到 $d^* = \frac{3}{2}$ 。从最上面的example我们可以知道， $d^* = f^* = \frac{3}{2}$ 。

现在，我们需要知道

when  $f^* = d^*$ , then it is **strong duality**.

when  $d^* \leq f^*$ , then it is **weak duality**.  $d^* \leq f^*$ 是always true的。

这里呢，就要来讲一个**slater condition**，这个slater condition是用来回答什么时候 $f^* = d^*$ 的，这是一个充分条件，不是一个充要条件。也就是说重要满足这个条件，就一定有 $f^* = d^*$ 。

若有convex problem,

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0 \\ & Ax = b \end{aligned}$$

其中 $f_i(x)$ 为convex,  $\forall i$ .

当 $\exists x \in \text{relint} D$ , 使 $f_i(x) < 0, i = 1, \dots, m$ , 还有  $Ax = b$  满足时,  $f^* = d^*$ . 这里要注意,  $f_i(x) < 0$ , 这里没有等号。

我们再来看一个example

$$\begin{aligned} \min \quad & x^T x \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

where  $x \in \mathbb{R}^n, b \in \mathbb{R}^p, A \in \mathbb{R}^{p \times n}$ . 我们一看式子，就要知道这是一个等式约束，我们要立马写出Lagrangian function和dual function来。Lagrangian function就是

$$L(x, \lambda) = x^T x + \lambda^T (Ax - b)$$

dual function 就是

$$\begin{aligned} d(\lambda) &= \inf_{x \in D} L(x, \lambda) \\ &= \inf_{x \in D} x^T x + \lambda^T (Ax - b) \end{aligned}$$

因为要minimize这个dual function，我们要做的就是求导，可以得到

$$\nabla_x d(\lambda) = 2x + A^T \lambda \rightarrow x = -\frac{A^T \lambda}{2}$$

然后再把这个 $x$ 代进Lagrangian function，完全变成跟 $\lambda$ 相关的函数，就是

$$\begin{aligned} \frac{\lambda^T A A^T \lambda}{4} - \frac{\lambda^T A A^T \lambda}{2} - \lambda^T b \\ = -\frac{\lambda^T A A^T \lambda}{4} - b^T \lambda \end{aligned}$$

这是一个concave function

接着再看一个example

$$\begin{aligned} \min_x \quad & \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{s.t.} \quad & x_1 + x_2 + x_3 + 3 \leq \epsilon \end{aligned}$$

这里的 $\epsilon$ 是一个很小的数。我们首先把Lagrangian function和dual function给写出来

Lagrangian function就是

$$L(x, \lambda) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + \lambda(x_1 + x_2 + x_3 + 3 - \epsilon)$$

dual function就是

$$d(\lambda) = \min_x \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + \lambda(x_1 + x_2 + x_3 + 3 - \epsilon)$$

这里呢，分别对 $x_1, x_2, x_3$ 求偏导，我们是可以得知 $x_1 = x_2 = x_3 = -\lambda$ 的，于是dual function就变成了

$$d(\lambda) = \min_x -\frac{3\lambda^2}{2} + \lambda(3 - \epsilon)$$

上面提到了strong duality，也就是 $f^* = d^*$ ，这里就来讲讲其四种解释

- 几何解释

我们要minimize这个函数

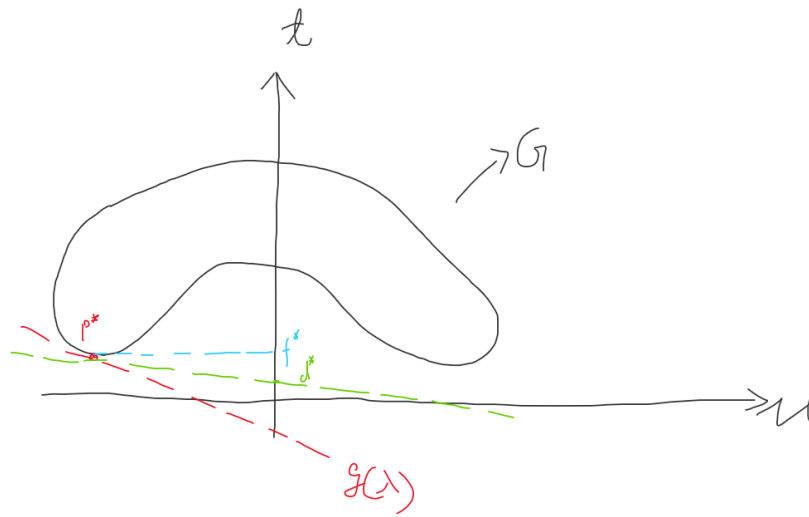
$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_1(x) \leq 0 \end{aligned}$$

现在假设有一个二维平面上有点，也就是 $G = \{(f_1(x), f_0(x)) | x \in D\}$ .

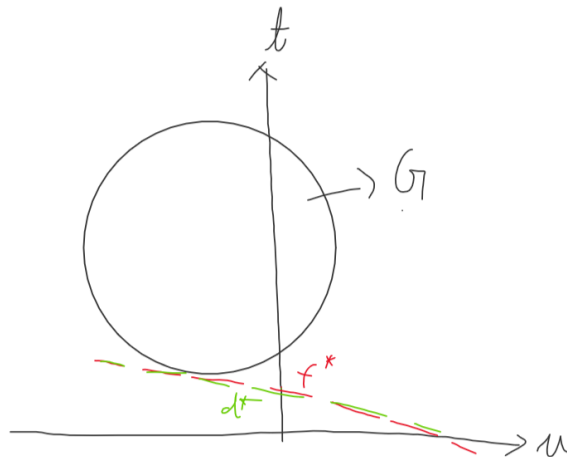
那么，我们的 $f^*$ 就是 $f^* = \inf\{t | (u, t) \in G, u \leq 0\}$ . 这个呢，其实就是上述方程，是一样的， $u = f_1(x), t = f_0(x)$ .

$g(\lambda) = \inf\{\lambda u + t | (u, t) \in G\}$ , 这个就是Lagrangian function。

我们的几何解释就是，我们看下图，看到绿色的那条线，我们既要通过 $p^*$ 的点，又要使得 $d^*$ 最大化，因为对偶问题解就是要最大化嘛。



当  $f^* = d^*$  时, 说明对这个图形是有要求的

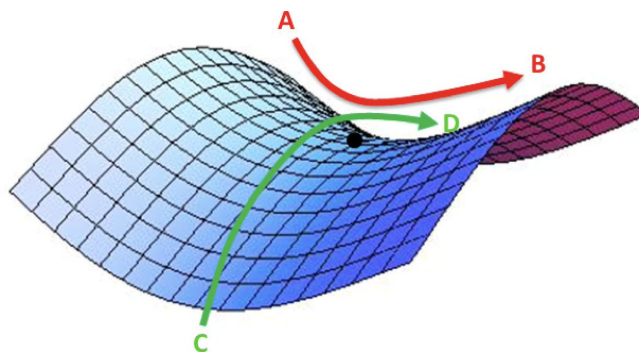


#### • 鞍点解释

saddle point, 可以写成

$$d^* = \inf_{x \in D} \sup_{\lambda \geq 0} L(x, y) = f^* = \sup_{\lambda \geq 0} \inf_{x \in D} L(x, y)$$

$\inf$  表示对  $x$  求极小值,  $\sup$  表示对  $\lambda$  求极大值, 看上述的式子, 就表示这是一个鞍点



saddle point  $p^* = d^*$ , 我们可以得到  $(x^*, \lambda^*)$  是 primal and dual problem 的最优解。

- 多目标优化解释
- 经济学角度解释

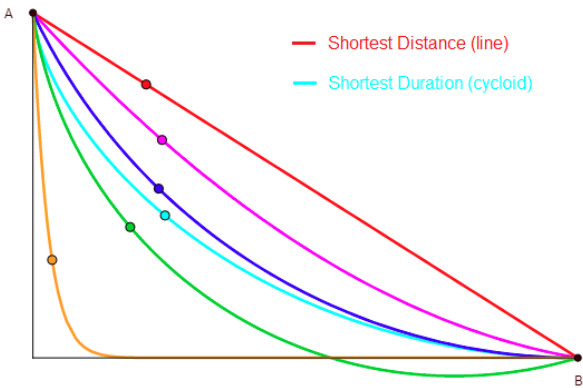
## Unit 5 Problems in data science

## Unit 6 Gradient descent

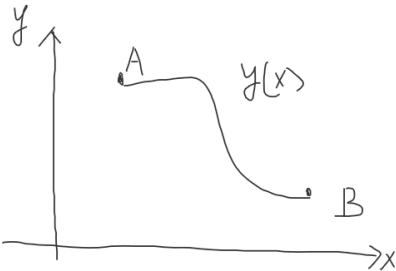
# Unit 7 Calculus of Variations

## Euler-Lagrange equation

这章学的东西，有点难度，calculus of variation，就是变分法。这里呢，最重要的就是Euler-Lagrange 方程。我们先来看一个例子，一般都是用最速降线 (Brachistochrone curve) 的问题来举例，就是固定两个端点，然后一个球从A点出发，看哪条线能够最快到达B点，然后我们要求的就是这条线。



再看一个例子，就是如下图所示，有A,B两个端点，我们要找连接A,B两点的线，求极小值



于是我们可以列出方程

$$\begin{aligned} L &= \int dl \\ &= \int \sqrt{dx^2 + dy^2} \\ &= \int \sqrt{\left(\frac{dy}{dx}\right)^2 + 1} dx \\ &= \int \sqrt{y'^2 + 1} dx \end{aligned}$$

上面得出的式子，其实就是泛函(functional), 我们要求的是 $y$ . 也就是说functional的目的是，我们要求一条函数，使得某个目标最大/小化. functional，它的输入是函数输出是实数。

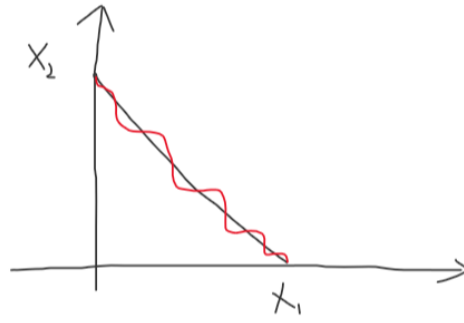
那么，怎么求这个函数呢？于是就有了Euler-Lagrange 方程

$$\frac{\partial L}{\partial f(x)} - \frac{d}{dx} \left( \frac{\partial L}{\partial f'(x)} \right) = 0$$

我们就来看看这个玩意到底是个啥，我们回到最速下降问题那里，我们有

$$A = \int L(f(x), f'(x), x) dx$$

我们想要求 $f(x)$ . 也就是我们想要求这个函数到底是什么。我们之前是在一个函数中求极值。现在是我们要找到一个函数，使得 $A$ 是一个极值。现在的方式就是引入一个任意函数 $\eta(x)$ , 但是这个函数也是有要求的，就是端点处不变，也就是 $\eta(x_1) = \eta(x_2) = 0$ . 如下图所示



就是引入函数后，会在周边波动，但是端点出不能有任何波动，所以就是 $\eta(x_1) = \eta(x_2) = 0$ 。然后呢，一般我们会对这个 $\eta(x)$ 乘以一个 $\epsilon$ 。也就是有一个函数 $\delta(x) = \epsilon\eta(x)$ ,  $\epsilon \in \mathbb{R}$ 。这样表达起来更方便，我们要知道的是， $|\delta(x)| < |f(x)|$ 。引入函数后，我们的表达式就变为

$$A = \int L(f(x) + \epsilon\eta(x), f'(x) + \epsilon\eta'(x), x) dx$$

接下来我们就要看E-L方程，就是 $\frac{dA}{d\epsilon} = 0$ 。也就是

$$\begin{aligned} \frac{dA}{d\epsilon} &= \frac{d(\int L(f(x) + \epsilon\eta(x), f'(x) + \epsilon\eta'(x), x) dx)}{d\epsilon} \\ &= \int \left[ \frac{\partial L}{\partial(f + \epsilon\eta)} \frac{d(f + \epsilon\eta)}{d\epsilon} + \frac{\partial L}{\partial(f' + \epsilon\eta')} \frac{d(f' + \epsilon\eta')}{d\epsilon} \right] dx \\ &= \int \left[ \frac{\partial L}{\partial f(x)} \eta(x) + \frac{\partial L}{\partial f'(x)} \eta'(x) \right] dx = 0 \\ &\quad \text{Integration by parts} \\ &= \int \left[ \frac{\partial L}{\partial f(x)} \eta(x) - \frac{d}{dx} \left( \frac{\partial L}{\partial f'(x)} \right) \eta(x) \right] dx + \frac{\partial L}{\partial f'(x)} \eta(x) \Big|_{x_1}^{x_2} \\ &\quad \text{we know that } \forall \eta(x), \eta(x_1) = \eta(x_2) = 0 \\ &= \int \left[ \frac{\partial L}{\partial f(x)} - \frac{d}{dx} \left( \frac{\partial L}{\partial f'(x)} \right) \right] \eta(x) dx = 0 \end{aligned}$$

因为 $\eta(x)$ 是任意函数，所以只能 $\frac{\partial L}{\partial f(x)} - \frac{d}{dx} \left( \frac{\partial L}{\partial f'(x)} \right) = 0$ 。这也就是我们的Euler-Lagrange 方程。这个必须要记住。

这里要讲三个E-L equation的special case

这些需要记住

<https://farside.ph.utexas.edu/teaching/336L/Fluid/node266.html>

- $L(x, \dot{x}, t) = L(\dot{x}, t)$

也就是说E-L equation跟 $x$ 无关，这样就可以得到

$$\frac{\partial L}{\partial x} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) \rightarrow \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) = 0$$

也就是说 $\frac{\partial L}{\partial \dot{x}} = \text{constant}$ 。

- $L(x, \dot{x}, t) = L(x, t)$

也就是说E-L equation跟 $x$ 的导数无关，这样就可以得到

$$\frac{\partial L}{\partial x} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) \rightarrow \frac{\partial L}{\partial x} = 0.$$

但是这个结果会得出一个或多个curves。

- $L(x, \dot{x}, t) = L(x, \dot{x})$

也就是说E-L equation跟 $t$ 的无关，这样就可以得到

$$\frac{\partial L}{\partial x} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) \rightarrow L - \frac{\partial L}{\partial \dot{x}} \dot{x} = \text{constant}.$$

它得来的过程是这样的，就是我们有 $\frac{\partial L}{\partial x} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) = 0$ ，然后两边乘 $\dot{x}$ ，可以得到

$$\dot{x} \frac{\partial L}{\partial x} - \dot{x} \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) = 0$$

因为

$$\frac{d}{dt} \left( \dot{x} \frac{\partial L}{\partial \dot{x}} \right) = \dot{x} \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) + \ddot{x} \frac{\partial L}{\partial \dot{x}}$$

所以，我们可以得到

$$\frac{d}{dt} \left( \dot{x} \frac{\partial L}{\partial \dot{x}} \right) = \dot{x} \frac{\partial L}{\partial x} + \ddot{x} \frac{\partial L}{\partial \dot{x}}$$

因为 $L$ 跟 $t$ 无关, 所以会变成

$$\frac{dL}{dt} = \frac{d}{dx}(\dot{x} \frac{\partial L}{\partial \dot{x}})$$

所以就可以得到

$$L - \frac{\partial L}{\partial \dot{x}} \dot{x} = \text{constant}$$

我们在上课时老师还讲了**First variation**

First variation of a functional的意思就是the linear functional  $\delta J(x(t))$  mapping to the function  $\eta$  to

$$\delta J(x^*(t), \eta) = \lim_{\epsilon \rightarrow 0} \frac{J(x^*(t) + \epsilon \eta) - J(x^*(t))}{\epsilon} = \left. \frac{d}{d\epsilon} J(x^*(t) + \epsilon \eta) \right|_{\epsilon=0}$$

这里呢,  $x(t)$ ,  $\eta$ 都是functions,  $\epsilon$ 是scalar. 这个也叫做Gateaux derivative of the functional.

我们来看一个例子, **example**

Find the first variation of

$$f(x) = \int_0^1 g(x(t), t) dt$$

where  $x \in C'[0, 1]$ ,  $g \in C''$ . 这里稍微解释一下 $C'$ ,  $C''$ 是什么意思,  $C'$ 就是continuous differentiable.  $C''$ 的意思就是continuous twice differentiable.

那怎么解这题呢? 很简单, 照抄公式即可

就是

$$\begin{aligned} \delta f(x, \eta) &= \lim_{\epsilon \rightarrow 0} \frac{f(x + \epsilon \eta) - f(x)}{\epsilon} = \left. \frac{d}{d\epsilon} (f(x + \epsilon \eta)) \right|_{\epsilon=0} \\ &= \frac{d}{d\epsilon} \int_0^1 g(x(t) + \epsilon \eta(t), t) dt \\ &= \int_0^1 \left. \frac{d}{d\epsilon} g(x(t) + \epsilon \eta(t), t) \right|_{\epsilon=0} dt \\ &= \int_0^1 g_x(x(t) + \epsilon \eta(t), t) \eta(t) \Big|_{\epsilon=0} dt \\ &= \int_0^1 g_x(x(t), t) \eta(t) dt \end{aligned}$$

就是这么简单解。

我们再来看一个例子 **example**

Find the first variation of

$$J(y) = \int_a^b y y' dx$$

也是很简单, 照抄公式即可。

$$\begin{aligned} \delta J(y, \eta) &= \lim_{\epsilon \rightarrow 0} \frac{J(y + \epsilon \eta) - J(y)}{\epsilon} = \left. \frac{d}{d\epsilon} J(y + \epsilon \eta) \right|_{\epsilon=0} \\ &= \frac{d}{d\epsilon} \int_a^b (y + \epsilon \eta)(y' + \epsilon \eta') \Big|_{\epsilon=0} dx \\ &= \int_a^b \frac{d}{d\epsilon} (yy' + y\epsilon\eta' + y'\epsilon\eta + \epsilon^2\eta\eta') \Big|_{\epsilon=0} dx \\ &= \int_a^b (y\eta' + y'\eta + 2\epsilon\eta\eta') \Big|_{\epsilon=0} dx \\ &= \int_a^b (y\eta' + y'\eta) dx \end{aligned}$$

求出来的结果就是first variation of the functional。

接下来讲讲**Fundamental lemma of calculus of variations**

我们会看到, 之前我们的公式, 还有Euler-Lagrange equation中, 总是有arbitrary function, 于是就需要这些fundamental lemma的证明。



If a continuous function  $f : [a, b] \rightarrow \mathbb{R}$  satisfies the equality

$$\int_a^b f(x)\eta(x)dx = 0$$

for  $\forall \eta(x)$  on  $[a, b] \in \mathbb{R}$ , and  $\eta(a) = \eta(b) = 0$  then  $f$  is identically zero on  $[a, b]$ .

因此呢, 我们有一个 **necessary condition**

If  $x(t) \in C'$  is a (local) min or max (extremum), then **E-L** equation holds.

我们来看一个例子, **example**

Find the stationary curve of

$$\begin{aligned} J(x(t)) &= \int_0^1 x^2 + \dot{x}^2 dt \\ \text{s.t. } x(0) &= 1 \\ x(1) &= 1 \end{aligned}$$

我们知道E-L equation holds here. 所以, 我们让  $G(x, \dot{x}, t) = x^2 + \dot{x}^2$ . 就跟上面是一样的, 我们也得知道  $\dot{x}$  就是  $x'$ . 都是一样的. 然后我们可以按着E-L equation来, 我们必须记住这个公式

$$\frac{\partial L}{\partial f(x)} - \frac{d}{dx} \left( \frac{\partial L}{\partial f'(x)} \right) = 0$$

于是我们可以得到

$$\frac{\partial G}{\partial x} = 2x, \frac{\partial G}{\partial \dot{x}} = 2\dot{x}, \frac{d}{dt} \left( \frac{\partial G}{\partial \dot{x}} \right) = 2\ddot{x}.$$

根据E-L equation, 我们可以得到

$$\frac{\partial G}{\partial x} - \frac{d}{dt} \left( \frac{\partial G}{\partial \dot{x}} \right) = 2x - 2\ddot{x} = 0. \text{ subject to } x(0) = 1, x(1) = 1.$$

因为  $\ddot{x}$  是二阶导, 我们让其等于  $S^2$ , 因为  $x$  是原函数, 于是  $\ddot{x} - x = (S^2 - 1) = 0$ , 可以得到  $S = \pm 1$ .

于是, 我们stationary curve  $x^*(t) = Ae^t + Be^{-t}$ . 之所以有  $t$  和  $-t$ , 就是因为  $S = \pm 1$ . 然后我们又还有两个constraints. 代进去, 可以得到

$$x(0) = 1 = A + B, x(1) = 1 = Ae + Be^{-1}.$$

从而可以求出

$$A = \frac{1}{e+1}, B = \frac{e}{e+1}.$$

然后再代进  $J(x(t))$  中, 从而求出一个实数

$$J^* = \int_0^1 (x^*)^2 + ((\dot{x}^*))^2 dt. \text{ 代进去之后, 可以得到 } J^* = \frac{2(e-1)}{e+1}.$$

接下来要讲 **Linear homogeneous differential equation**.

这也就是刚刚的example中突然出现的  $x^*(t) = Ae^t + B^{-t}$  是怎么来的, 这里会解释。

<https://brilliant.org/wiki/homogeneous-linear-differential-equations/>

这个Linear homogeneous differential equation看起来是下面这样的, 这里需要注意一下,  $y^n$  表示求了几次导的意思

$$y^n(x) + a_1 y^{(n-1)}(x) + \dots + a_{n-1} y^1(x) + a_n y(x) = 0$$

然后这里呢,  $a_1, \dots, a_n$  都是constant.

我们要解决的就是这个方程的solution是什么, 我们可能会猜测, 这个solution就是  $y = e^{rx}$ . 也就是  $\frac{d^k y}{dx^k} e^{rx} = r^k e^{rx}$ . 如果  $y = e^{rx}$  是solution的话, 那么就会满足

$$r^n e^{rx} + a_1 r^{n-1} e^{rx} + \dots + a_{n-1} r e^{rx} + a_n e^{rx} = 0$$

因为  $e^{rx} \neq 0$ , 所以一定会有

$$r^n + a_1 r^{n-1} + \dots + a_{n-1} r + a_n = 0$$

于是, 就有了这么一个definition, 就是对于一个differential equation

$$y^n(x) + a_1 y^{(n-1)}(x) + \dots + a_{n-1} y^1(x) + a_n y(x) = 0$$

它所对应的characteristic equation就是

$$r^n + a_1 r^{n-1} + \dots + a_{n-1} r + a_n = 0$$

举个简单的例子来看**example**

就是下面这个式子的solution是什么

$$y''' + 2y'' - y' - 2y = 0$$

很简单，我们就照着公式写，我们可以写出

$$S^3 + 2S^2 - S - 2 = 0$$

↓

$$S_1 = -2, S_2 = 1, S_3 = -1$$

于是，这个式子的solution就是

$$y(x) = Ae^{-2x} + Be^x + Ce^{-x}$$

根据characteristic polynomial 的root的情况不同，这个differential equation会有不太一样的solutions

- case of distinct real root

当roots是real而且是distinct的时候，那么solution就是不同的root的linear combination of  $e^{rx}$ . 也就是

$$y(x) = c_1 e^{r_1 x} + c_2 e^{r_2 x} + \dots + c_n e^{r_n x}$$

我们举个例子来看， **example**

Solve

$$y'' + 2y' - 8y = 0, y(0) = 5, y'(0) = -2$$

那么，我们就可以得到

$$S^2 + 2S - 8 = 0$$

↓

$$S_1 = -4, S_2 = 2$$

↓

$$y(x) = Ae^{-4x} + Be^{2x}$$

我们也知道 $y(0) = 5, y'(0) = -2$ , 所以，可以求出 $A = 2, B = 3$ , 于是solution就是 $y(x) = 2e^{-4x} + 3e^{2x}$ .

- real but multiple roots (repeated roots)

这个就是有重复root的情况，那么solution就会变成

$$y(x) = (c_1 + c_2 x + c_3 x^2 + \dots + c_k x^{k-1}) e^{rx}$$

我们来看一个例子， **example**, 就是

$$y''' - 7y'' + 11y' - 5y = 0$$

我们也是照旧，写公式，求root，我们可以写出

$$S^3 - 7S^2 + 11S - 5 = 0$$

↓

$$(S - 1)^2 (S - 5) = 0$$

↓

$$S_1 = S_2 = 1, S_3 = 5$$

于是，这道题的solution就变成了

$$y(x) = (c_1 + c_2 x) e^x + c_3 e^{5x}$$

我们再来看一道例题， **example**, 就是

Find the general solution of

$$y^{(4)} + 3y^{(3)} + 3y'' + y' = 0$$

同样，也是先求root，可以得到

$$S^4 + 3S^3 + 3S^2 + S = 0$$

↓

$$S(S + 1)^3 = 0$$

↓

$$S_1 = 0, S_2 = S_3 = S_4 = -1$$

于是这道题的solution就是

$$y(x) = c_1 + (c_2 + c_3 x + c_4 x^2) e^{-x}$$

- complex roots

如果是complex root, 那么root的形式就是 $x = a \pm bi$ .  $i = \sqrt{-1}$ , 就是虚数.

假设现在一个function有两队complex roots, 分别是 $\lambda_{1,2} = \alpha \pm \beta i$ ,  $\lambda_{3,4} = \gamma \pm \delta i$ , 那么其solution就是

$$y(x) = e^{\alpha x}(c_1 \cos \beta x + c_2 \sin \beta x) + e^{\gamma x}(c_3 \cos \delta x + c_4 \sin \delta x)$$

我们来看一个例子, **example**

Solve

$$y'' - 4y' + 5y = 0$$

那么我们可以写出其characteristic equation, 并可以得到root, 就是

$$S^2 - 4S + 5 = 0$$

↓

$$S_{1,2} = 2 \pm i$$

于是, 这个solution就是

$$y(x) = e^{2x}(c_1 \cos x + c_2 \sin x)$$

- multiplicity of complex roots (repeated root)

就是complex root也有重复的时候, 那么也很简单, 跟repeated real root做法是一样的。我们来看一个例子就懂了

Find the general solution of

$$y^{(4)} + 4y^{(3)} + 12y'' + 16y' + 16y = 0$$

其characteristic equation就是 $(S^2 + 2r + 4)^2 = 0$ , 这个root就是 $\lambda_{1,2,3,4} = -1 \pm \sqrt{3}i$ . 也就是重复了

所以, 它的general solution就是

$$y(x) = e^{-x}(c_1 \cos \sqrt{3}x + c_2 \sin \sqrt{3}x) + xe^{-x}(c_3 \cos \sqrt{3}x + c_4 \sin \sqrt{3}x)$$

### Shortest path problem

这其实也是上课讲的一个**example**, 就是

Find the general solution of

$$J(y(x)) = \int_a^b \sqrt{1 + y'(x)^2} dx$$

很简单, 我们仍旧是照套公式, 公式就是E-L equation,  $\frac{\partial L}{\partial y} - \frac{d}{dx} \left( \frac{\partial L}{\partial y'} \right) = 0$ .

这里呢,  $L = \sqrt{1 + y'(x)^2}$ , 那么, 可以得到 $\frac{\partial L}{\partial y} = 0$ , 因为没有与 $y$ 的部分。然后就是 $\frac{\partial L}{\partial y'} = \frac{y'(x)}{\sqrt{1 + y'(x)^2}}$ . 然后根据E-L equation, 我们可以得到 $0 - \frac{d}{dx} \left( \frac{\partial L}{\partial y'} \right) = 0$ . 所以可以推测出 $\frac{\partial L}{\partial y'} = c$ , 是一个常数, 因为这样对 $x$ 求导才会等于0嘛。

因为 $\frac{\partial L}{\partial y'} = \frac{y'(x)}{\sqrt{1 + y'(x)^2}}$ , 可以得到 $y'(x) = c\sqrt{1 + y'(x)^2}$ . 因为这样才会得到常数 $c$ . 然后我们再往回推, 可以得到

$$y'(x) = c\sqrt{1 + y'(x)^2}$$

↓

$$y'(x)^2 = c^2(1 + y'(x)^2)$$

↓

$$(1 - c^2)y'(x)^2 = c^2$$

↓

$$y'(x)^2 = c_1$$

↓

$$y'(x) = c_2$$

↓

$$y(x) = c_2 x + d$$

这里的 $c_1, c_2, d$ 都是常数。

### Shortest time problem

<https://www.zhihu.com/question/318376175>

<https://www.jianshu.com/p/961e890e88b2>

这也是上课讲的一个例子, **example**, 这也是最开头提到的最速降线的问题, 我们先不管公式怎么来的, 最终要solve的问题就是

$$t_{AB} = \frac{1}{\sqrt{2g}} \int_A^B \frac{\sqrt{1+y'^2}}{\sqrt{y}} dx$$

我们一看这个式子，就知道这与 $x$ 无关，只与 $y, y'$ 有关，于是，我们想到special case那里，可以得到

$$L - \frac{\partial L}{\partial y'} y' = \text{constant}$$

↓

$$\frac{\sqrt{1+y'^2}}{\sqrt{2gy}} - y' \left( \frac{y'}{\sqrt{2gy(1+y'^2)}} \right) = \frac{1}{\sqrt{2gy(1+y'^2)}} = C$$

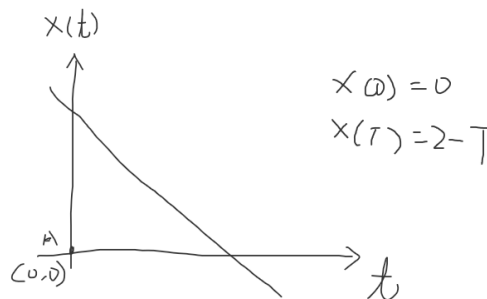
此时，我们可以得出  $\frac{1}{\sqrt{y(1+y'^2)}} = C_1$ ，所以  $y(1+y'^2) = C_2$ 。然后呢，就只能记住了，可以得到的曲线是

$$\begin{cases} x = A(\theta - \sin\theta) \\ y = A - A\cos\theta \end{cases}$$

这个就是最速曲线的方程

### Variable end-time problem

这个问题在于不是固定端，之前的问题都是两个端点固定，现在不是，有一个端点没有固定，是一个函数。举个例子来说，如下图所示，A点是固定的，但是B点不定，在于 $x(T) = 2 - T$ 那条函数上。



于是呢，这里，就有一个**定理, proposition**

Consider a scalar function  $z(t)$  which extremizer the function  $J(x) = \int_a^b L(x(t), \dot{x}(t), t) dt$  and satisfy  $x(a) = a, x(T) = f(T)$ , where  $T$  has to be determined and  $f(T)$  is given. Then, the necessary conditions for  $x$  to be extremum of the above problem are

- E-L equation:  $\frac{\partial L}{\partial x} = \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right)$ .
- Transversality condition:  $\left( L - \frac{\partial L}{\partial \dot{x}} (\dot{x} - \dot{f}) \right) \Big|_{t=T} = 0$ .

我们里来看一个**example**,

假设现在的functional是

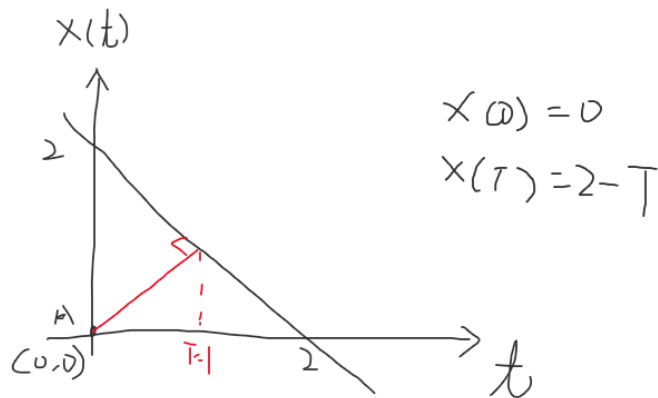
$$J = \int_0^T \sqrt{1 + \dot{x}(t)^2} dt$$

constraints 是  $x(0) = 0, x(T) = 2 - T$ . 我们仍旧是要求 $x(t)$ 的。怎么求呢？很简单，照着公式走，两个conditions

- E-L equation: 就是  $\frac{\partial L}{\partial x} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) = 0$ , 因为与 $x$ 无关，所以，可以得出  $\frac{\dot{x}}{\sqrt{1+\dot{x}^2}} = c$ , 这是常数。这之前我们也有写过，我们可以得到  $\dot{x} = c_1$ , 就是另一个常数，于是我们的  $x(t) = c_1 t + c_2$ . 因为  $x(0) = 0 \rightarrow c_2 = 0$ . 所以， $x(t) = c_1 t$ .
- 第二个condition就是Transversality condition, 也就是  $\left( L - \frac{\partial L}{\partial \dot{x}} (\dot{x} - \dot{f}) \right) \Big|_{t=T} = 0$ . 在这里，我们知道  $\dot{f} = -1$ , 我们可以得到

$$\begin{aligned}
& \left( L - \frac{\partial L}{\partial \dot{x}} (\dot{x} - \dot{f}) \right) \Big|_{t=T} = 0 \\
& = \left( \sqrt{1 + \dot{x}(t)^2} - \frac{\dot{x}}{\sqrt{1 + \dot{x}^2}} (\dot{x} + 1) \right) \Big|_{t=T} \\
& \quad \downarrow \\
& \sqrt{1 + \dot{x}(t)^2} = \frac{\dot{x}}{\sqrt{1 + \dot{x}^2}} (\dot{x} + 1) \Big|_{t=T} \\
& \quad \downarrow \\
& 1 + \dot{x}(t)^2 = \dot{x}(\dot{x} + 1) \Big|_{t=T} \\
& \quad \downarrow \\
& \dot{x}(t) = 1
\end{aligned}$$

于是, 我们可以总结出  $c_1 = 1$ . 可以得到  $x(t) = t$ , 于是可以求出,  $T = 1$ . 也就是如下图所示



### vector case

之前只是讲过scalar的情况, 现在来看看**vector**的情况

就是假设

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

然后functional是

$$\begin{aligned}
J(x(t)) &= \int_a^b L(x_1(t), \dots, x_n(t), \dot{x}_1(t), \dots, \dot{x}_n(t), t) dt \\
&= \int_a^b L(x(t), \dot{x}(t), t) dt
\end{aligned}$$

即使vector的情况, 也是可以照样用E-L equation的

$$\frac{\partial L}{\partial x} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) = 0$$

然后, if any component  $x_j(a)$  is not specified, then  $\frac{\partial L}{\partial \dot{x}_j} \Big|_{t=a} = 0$ ,  $x_j(b)$  is not specified, then  $\frac{\partial L}{\partial \dot{x}_j} \Big|_{t=b} = 0$ .

这里来看一个例子, **example**

Solve

$$x = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

and

$$L(x, \dot{x}, t) = \dot{x}_1^2 - \dot{x}_2^2 + 2x_1x_2 - 2x_2^2$$

and

$$x[0] = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and

$$J(x(t)) = \int_0^1 L(x, \dot{x}, t) dt$$

就是这么一个题目，得求 $x$ 。方法也是一样，先求E-L equation

$$\begin{aligned} \frac{\partial L}{\partial x_1} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}_1} \right) &= 2x_2 - \frac{d}{dt} 2\dot{x}_1 = 2x_2 - 2\ddot{x}_1 = 0 \\ \frac{\partial L}{\partial x_2} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}_2} \right) &= 2x_1 - 4x_2 - \frac{d}{dt} 2\dot{x}_2 = 2x_1 - 4x_2 - 2\ddot{x}_2 = 0 \end{aligned}$$

于是，我们可以得到

$$\begin{aligned} \ddot{x}_1 &= x_2 \rightarrow \ddot{x}_2 = \ddot{\ddot{x}}_1 \\ \ddot{x}_2 - 2x_2 &= -x_1 \\ \downarrow \\ \ddot{\ddot{x}}_1 &= x_2 = 2x_2 - x_1 = 2\ddot{x}_1 - x_1 \\ \downarrow \\ \ddot{\ddot{x}}_1 - 2\ddot{x}_1 + x_1 &= 0 \end{aligned}$$

于是，我们现在又可以转换成求根的情况了，就是变成 $S^4 - 2S^2 + 1 = 0$ ，我们可以得到 $S_{1,2} = 1, S_{3,4} = -1$ 。

于是，我们可以得出

$$\begin{aligned} x_1(t) &= Ae^t + Bte^t + Ce^{-t} + Dte^{-t} \\ x_1'(t) &= Ae^t + B(e^t + te^t) - Ce^{-t} + D(e^{-t} - te^{-t}) \\ x_2 = x_1''(t) &= (A + 2B + Bt)e^t + (C - 2D + Dt)e^{-t} \end{aligned}$$

我们又还有constraints,

$$x[0] = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

于是，我们可以得出

$$x[0] = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} A + C \\ A + 2B + C - 2D \end{bmatrix}$$

看回我们的题目，我们是 $J = \int_0^1$ 。所以，有一个natural boundary condition，也就是not specified的情况下，

$$\begin{aligned} \left. \frac{\partial L}{\partial \dot{x}_1} \right|_{t=1} &= 0 \\ \left. \frac{\partial L}{\partial \dot{x}_2} \right|_{t=1} &= 0 \end{aligned}$$

于是，我们就有了四个式子，就可以求出 $A, B, C, D$ 来啦。

接下来我们就来看一下有**constraint**的情况，有constraint的情况其实也就跟之前的Lagrangian function一样的做法

这里的constraint分为Integral constraint和non-integral constraint两种

- Integral constraint  
比如说，一般形式就是

$$\min J(x(t)) = \int_a^b G(x, \dot{x}, t) dt$$

$$s.t. \quad \int_a^b g(x, \dot{x}, t) dt = l$$

我们的做法就是先写成Lagrangian function, 就是

$$Q = G + \lambda g = \int_a^b G(x, \dot{x}, t) dt + \lambda (\int_a^b g(x, \dot{x}, t) dt - l).$$
 这里的 $\lambda$ 就是 Lagrangian multiplier.

于是, 我们就可以写E-L equation了, 那么就是

$$\frac{d}{dx}(G + \lambda g) = \frac{d}{dt} \frac{d}{dx}(G + \lambda g).$$

我们来看一个例子, **example**

$$J(x) = \int_a^b x(t) dt$$

$$s.t. \quad \int_a^b \sqrt{1 + \dot{x}^2} dt = 0$$

$$\text{and } x(a) = x_a, x(b) = x_b.$$

于是, 我们就要先写Lagrangian function, 就是

$$Q = G + \lambda g = x + \lambda \sqrt{1 + \dot{x}^2}$$

$$\text{然后我们再写E-L equation, 就是 } \frac{\partial L}{\partial x} - \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}} \right) = 0.$$

$$\text{我们可以得到 } \frac{\partial Q}{\partial x} = 1, \frac{\partial Q}{\partial \dot{x}} = \frac{\lambda \dot{x}}{\sqrt{1 + \dot{x}^2}}. \text{ 于是我们可以得到}$$

$$1 - \frac{d}{dt} \left( \frac{\lambda \dot{x}}{\sqrt{1 + \dot{x}^2}} \right) = 0$$

$$\downarrow$$

$$\frac{d}{dt} \left( \frac{\lambda \dot{x}}{\sqrt{1 + \dot{x}^2}} \right) = 1$$

$$\downarrow$$

$$d \left( \frac{\lambda \dot{x}}{\sqrt{1 + \dot{x}^2}} \right) = dt$$

$$\downarrow$$

$$\frac{\lambda \dot{x}}{\sqrt{1 + \dot{x}^2}} = t + C$$

$$\downarrow$$

$$\dot{x}^2 = \frac{(t + C)^2}{\lambda^2 - (t + C)^2}$$

于是, 我们可以得到  $x = \pm \sqrt{\lambda^2 - (t + C)^2} + d$ , 这里,  $d$ 也是constant. 接着写, 我们可以得到

$$(x - d)^2 + (t + C)^2 = \lambda^2$$

这里呢, 我们会有三个constant, 也就是三个参数, 分别是 $C, d, \lambda$ . 然后我们代进constraint中, 也就是

$$x(a) = x_a, x(b) = x - b, \text{ 还有就是 } \int_a^b \sqrt{1 + \dot{x}^2} dt = 0. \text{ 就可以得到这个参数了}$$

- non-integral constraint

这里就是non-integral constraint的情况了, 我们来看看

$$J(x) = \int_a^b G(x, \dot{x}, t) dt$$

$$s.t. \quad g_i(x, \dot{x}, t) dt = 0, i = 1, \dots, k$$

$$\text{and } x(a) = x_a, x(b) = x_b.$$

同样, 我们也是要先写Lagrangian function, 也就是

$$Q = G + \sum_{i=1}^k \lambda_i g_i$$

然后, 我们再写E-L equation

现在, 我们来看一个例子, **example**

$$\min J(x) = \int_0^1 (x_1^2 + \dot{x}_1^2 + x_2^2 + \dot{x}_2^2) dt$$

$$s.t. \quad \dot{x}_1 = -x_1 + x_2$$

$$\text{and } x_1(0) = 1, x_1(1) = x_2(0) = x_2(1) = 0.$$

我们先写Lagrangian function, 就是

$$Q = (x_1^2 + \dot{x}_1^2 + x_2^2 + \dot{x}_2^2) + \lambda(\dot{x}_1 + x_1 - x_2)$$

然后, 我们求E-L equation, 就是

$$\begin{aligned} \frac{\partial Q}{\partial x} - \frac{d}{dt} \left( \frac{\partial Q}{\partial \dot{x}} \right) &= 0 \\ \downarrow \\ \begin{bmatrix} 2x_1 + \lambda \\ 2x_2 - \lambda \end{bmatrix} &= \frac{d}{dt} \begin{bmatrix} 2\dot{x}_1 + \lambda \\ 2\dot{x}_2 \end{bmatrix} = \begin{bmatrix} 2\ddot{x}_1 + \dot{\lambda} \\ 2\ddot{x}_2 \end{bmatrix} \end{aligned}$$

于是, 我们可以得到

$$\begin{aligned} \ddot{x}_1 - x_1 + \frac{1}{2}(\dot{\lambda} - \lambda) &= 0 \\ \ddot{x}_2 - x_2 + \frac{1}{2}\lambda &= 0 \\ \dot{x}_1 + x_1 - x_2 &= 0 \end{aligned}$$

## Unit 8 Dynamic Programming

DP这一章节呢, 就主要是讲例子为主

### Stochastic optimal control problem

这个问题就是find the control law来最小化cost。这是用DP来解决的问题。我们来看一个具体的例子, 就是

#### example

consider the linear stochastic system described by

$$x_{k+1} = x_k + u_k + w_k$$

with  $\mathbb{E}[x_0] = 0, \mathbb{E}[x_0^2] = 1, \mathbb{E}[w_k] = 0, \mathbb{E}[w_k^2] = 1$ . suppose  $N = 2$ , and the cost criterion is  $J = \mathbb{E}(x_0^2 + x_1^2 + x_2^2)$ . And  $u_k$ , the control policy  $\Phi = \{\phi_0, \phi_1\}$  with  $\phi_0(x) = -2x, \phi_1(x) = -3x$ . What is  $\min J$  when  $N = 2$ .

所以呢, 我们先在就要来计算这个 $J$ 了。其实很简单, 就是照抄公式而已

$$\begin{aligned} x_1 &= x_0 + u_0 + w_0 \\ &= x_0 + \phi_0(x) + w_0 \\ &= x_0 - 2x_0 + w_0 \\ &= -x_0 + w_0 \end{aligned}$$

因为 $N = 2$ , 所以还要计算 $x_2$

$$\begin{aligned} x_2 &= x_1 + u_1 + w_1 \\ &= x_1 + \phi_1(x) + w_1 \\ &= x_1 - 3x_1 + w_1 \\ &= -2x_1 + w_1 \\ &= -2(-x_0 + w_0) + w_1 \\ &= 2x_0 - 2w_0 + w_1 \end{aligned}$$

于是, 我们就可以计算cost  $J$  了

$$\begin{aligned} J(\Phi) &= \mathbb{E}(x_0^2 + x_1^2 + x_2^2) \\ &= \mathbb{E}(x_0^2) + \mathbb{E}(x_1^2) + \mathbb{E}(x_2^2) \\ &= \mathbb{E}(x_0^2) + \mathbb{E}(-x_0 + w_0)^2 + \mathbb{E}(2x_0 - 2w_0 + w_1)^2 \\ &= 1 + 2 + 9 = 12 \end{aligned}$$

### Gambling problem

A Gambler enters a game, at time  $k$ , he may stake any amount  $u_k > 0$  that does not exceed his current fortune  $x_k$  (it is defined to be his initial capital plus his gain or minus his loss thus far). If he wins, he gets back his stake plus an additional amount equal to his stake so that his fortune will increase from  $x_k$  to  $x_k + u_k$ . If he losses, his fortune decreases to  $x_k - u_k$ . His probability of winning at each stake is  $p$  where  $\frac{1}{2} < p < 1$ , so that his probability of losing is  $1 - p$ . His objective is to maximize  $\mathbb{E} \log x_N$  where  $x_N$  is his fortune after  $N$  plays.

Then the stochastic control problem is characterized by the stake equation



$$x_{k+1} = x_k + u_k w_k$$

where  $p(w_k = 1) = p, p(w_k = -1) = 1 - p$ . Since there are no per stage costs, we can write down the DP equation

$$V_k(x) = \max_u \mathbb{E}[V_{k+1}(x + u_k w_k)]$$

with terminal condition  $V_N(x) = \log x$ .

Since it is not obvious what is the form of the function  $V_k(x)$ , we do one step of DP computation starting from the known terminal condition at time  $N$ .

$$\begin{aligned} V_{N-1}(x) &= \max_u \mathbb{E} \log(x + u w_{N-1}) \\ &= \max_u \{p \log(x + u) + (1 - p) \log(x - u)\} \end{aligned}$$

Differentiating for  $u$ , we get

$$\frac{p}{x + u} - \frac{1 - p}{x - u} = 0$$

When we simplify the function, we get

$$u_{N-1} = (2p - 1)x_{N-1}$$

and we can substitute the  $u$  to the DP equation, then we can get

$$\begin{aligned} V_{N-1}(x) &= p \log 2px + (1 - p) \log 2(1 - p)x \\ &= p \log 2p + p \log x + (1 - p) \log 2(1 - p) + (1 - p) \log x \\ &= \log x + p \log 2p + (1 - p) \log 2(1 - p) \end{aligned}$$

We see that the function  $\log x + \alpha x$  fits the form of  $V_{N-1}(x)$  as well as  $V_N(x)$ . This suggests that we can try the following guess for the optimal value function

$$V_k(x) = \log x + \alpha_k$$

Putting into the DP equation, we find that

$$\begin{aligned} \log x + \alpha_k &= \max_u \mathbb{E} \{ \log(x + u w_{k-1} + \alpha_{k+1}) \} \\ &= \max_u \mathbb{E} \{ p \log(x + u) + (1 - p) \log(x - u) + \alpha_{k+1} \} \end{aligned}$$

Noting that the maximization is the same as that for time  $N - 1$ , we have again the optimizing  $u_k$  given by

$$u_k = (2p - 1)x_k$$

Substituting, we get

$$\begin{aligned} \log x + \alpha_k &= p \log(2px) + (1 - p) \log 2(1 - p)x + \alpha_{k+1} \\ &= p \log 2p + p \log x + (1 - p) \log 2(1 - p) + (1 - p) \log x + \alpha_{k+1} \\ &= \log x + \alpha_{k+1} + p \log 2p + (1 - p) \log 2(1 - p) \end{aligned}$$

We see that the trial solution indeed solves the dynamic programming equation if we set the sequence  $\alpha_k$  to be given by the equation

$$\begin{aligned} \alpha_k &= \alpha_{k+1} + p \log 2p + (1 - p) \log 2(1 - p) \\ &= \alpha_{k+1} + \log 2 + p \log p + (1 - p) \log(1 - p) \end{aligned}$$

with terminal condition  $\alpha_N = 0$ . This completely determines the optimal policy for this gambling problem.

### Inventory control problem

A store needs to order inventory at the beginning of each day to fill the needs of customers. We assume that whatever stock ordered is delivered immediately. We assume, for simplicity, that the cost per unit stock order is 1 and the holding cost per unit item remaining unsold at the end of day is also 1. Furthermore, there is a shortage cost per unit demand unfilled of 3. The stochastic control problem is: given the probability distribution for the random demand during the day, find the optimal planning policy for 2 days to minimize the expected cost, subject to a storage constraint of 2 items.

To analyze this problem, let us introduce mathematical notation and make precise our assumptions.

Let  $x_k$  be the stock available at the beginning of the  $k^{th}$  day,  $u_k$  the stock ordered at the beginning of the  $k^{th}$  day.  $w_k$  the random demand during the  $k^{th}$  day. The storage constraint of 2 units translate to the inequality  $x_k + u_k \leq 2$ . Since stock is nonnegative and integer-valued, we must also have  $0 \leq x_k, 0 \leq u_k$ . The  $x_k$  process is then seen to satisfy the equation

$$x_{k+1} = \max(0, x_k + u_k - w_k)$$

Now, let us assume that the probability distribution of  $w_k$  is the same for all  $k$ , given by  $p(w_k = 0) = 0.1, p(w_k = 1) = 0.7, p(w_k = 2) = 0.2$ .

Assume also that the initial stock  $x_0 = 0$ . The cost function is given by (这里呢，第一项 $u_k$ 就是每日进货，cost是1，第二项就是一天中最后剩下下来要存储的，cost是1，最后一项是进货的量不够，订单太多，损失的cost，每损失一个的cost是3)。

$$L_k(x_k, u_k, w_k) = u_k + \max(0, x_k + u_k - w_k) + 3\max(0, w_k - x_k - u_k)$$

When  $N = 1$ , since we are planning for today and tomorrow. So the DP algorithm gives

$$V_k^* = \min_{0 \leq u_k \leq 2-x} \mathbb{E}\{u_k + \max(0, x + u_k - w_k) + 3\max(0, w_k - x - u_k) + V_{k+1}^*[\max(0, x + u_k - w_k)]\}$$

with  $V_2^*(x) = 0$  for all  $x$ .

We now proceed backwards

$$V_1^*(x) = \min_{0 \leq u_1 \leq 2-x} \mathbb{E}\{u_1 + \max(0, x + u_1 - w_1) + 3\max(0, w_1 - x - u_1)\}$$

Now the values that  $x$  can take on are 0, 1, 2, and so is  $u_1$ . Hence, using the probability distribution for  $w_1$ , we get

$$V_1^*(0) = \min_{0 \leq u_1 \leq 2-x} \mathbb{E}\{u_1 + 0.1\max(0, u_1) + 0.3\max(0, -u_1) + 0.7\max(0, u_1 - 1) + 2.1\max(0, 1 - u_1) + 0.2\max(0, u_1 - 2) + 0.6\max(0, 2 - u_1)\}$$

For  $u_1 = 0$ , the **R.H.S.** of the above function is  $= 2.1 + 1.2 = 3.3$ .

For  $u_1 = 1$ , the **R.H.S.** of the above function is  $= 1 + 0.1 + 0.6 = 1.7$ .

For  $u_1 = 2$ , the **R.H.S.** of the above function is  $= 2 + 0.2 + 0.7 = 2.9$ .

Hence the minimizing  $u_1$  for  $x_1 = 0$  is 1 so that  $\phi_1^*(0) = 1$ , and  $V_1^*(0) = 1.7$ .

Similarly, for  $x_1 = 1$ , we obtain

$$V_1^*(1) = \min_{0 \leq u_1 \leq 2-x} \mathbb{E}\{u_1 + \max(0, 1 + u_1 - w_1) + 3\max(0, w_1 - 1 - u_1)\} = 0.7 \text{ for the choice } u_1 = 0$$

Hence,  $\phi_1^*(1) = 0$ , and  $V_1^*(1) = 0.7$ .

Finally, for  $x_1 = 2$ , we have

$$V_1^*(2) = \min_{0 \leq u_1 \leq 2-x} \mathbb{E}\{u_1 + \max(0, 2 + u_1 - w_1) + 3\max(0, w_1 - 2 - u_1)\} = 0.9$$

In this case, no decision on  $u_1$  is necessary since it is constrained to be 0. Hence  $\phi_1^*(2) = 0$ . Now to go back to  $k = 0$ , we apply the  $V_k^*$  to get

$$V_0^* = \min_{0 \leq u_0 \leq 2-x} \mathbb{E}\{u_0 + \max(0, x + u_0 - w_0) + 3\max(0, w_0 - x - u_0) + V_1^*[\max(0, x + u_0 - w_0)]\}$$

Since the initial condition is taken to be  $x = 0$ , we need only compute  $V_0^*(0)$ . This gives

$$\begin{aligned} V_0^* &= \min_{0 \leq u_0 \leq 2} \mathbb{E}\{u_0 + \max(0, x + u_0 - w_0) + 3\max(0, w_0 - x - u_0) + V_1^*[\max(0, x + u_0 - w_0)]\} \\ &= \min_{0 \leq u_0 \leq 2} \{u_0 + 0.1\max(0, u_0) + 0.3\max(0, -u_0) \\ &\quad + 0.1V_1^*[\max(0, u_0)] + 0.7\max(0, u_0 - 1) + 2.1\max(0, 1 - u_0) \\ &\quad + 0.7V_1^*[\max(0, u_0 - 1)] + 0.2\max(0, u_0 - 2) + 0.6\max(0, 2 - u_0) \\ &\quad + 0.2V_1^*[\max(0, u_0 - 2)]\} \end{aligned}$$

Then we can use the values of  $V_1^*$  computed at the previous step, we find that for

For  $u_0 = 0$ , the **R.H.S.** of the above function is  $= 5.0$ .

For  $u_0 = 1$ , the **R.H.S.** of the above function is  $= 3.3$ .

For  $u_0 = 2$ , the **R.H.S.** of the above function is  $= 3.82$ .

Hence, the minimizing  $u_0$  is  $u_0 = 1$  and

$$V_0^*(0) = 3.3 \text{ with } \phi_0^*(0) = 1$$

Had the initial state been 1, we would have

$$V_0^*(1) = 2.3 \text{ with } \phi_0^*(1) = 0$$

and had  $x_0$  been 2, we would have

$$V_0^*(2) = 1.82 \text{ with } \phi_0^*(2) = 0$$

The above calculations completely characterize the optimal policy  $\Phi^*$ . Note that the optimal control policy is given as a look-up table, not as an analytical expression.