矩阵微积分及其应用

矩阵梯度、线性回归与范数理论

Jingbo Xia

2026 秋——人工智能 2401/2402

目录

1. 矩阵梯度及其应用 3
2. 线性回归 39
3. 范数理论及其应用 127

1. Outline

1. 矩阵梯度及其应用 3
   - 梯度的定义与记号 6
   - 向量变量的梯度公式 11
   - 矩阵变量的梯度公式 24
2. 线性回归 39
   - 线性回归与一般意义下的回归问题 40
   - 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
   - 回归问题中的记号准备 52
   - 线性回归与最小二乘法 66
   - 正则化线性回归（Ridge 与 LASSO） 100
   - LASSO 回归与 Python 代码示例 117
3. 范数理论及其应用 127
   - 向量范数 128
   - 矩阵诱导范数与加权线性回归 168
   - 向量范数的等价性* 187
   - 矩阵范数 200
   - 矩阵的诱导范数 210
   - 三种相容的矩阵范数 232
   - Frobenius 范数的迹运算* 258

内容结构

- 向量梯度与矩阵梯度的定义；
- 线性函数、二次型和复合函数的梯度；
- 迹、逆矩阵、行列式和对数行列式的梯度；
- 对称矩阵、对角矩阵等结构约束下的梯度。

记号原则

向量统一使用粗体希腊字母，如 $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$；矩阵统一使用粗体大写字母，如 $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X}$；分量 α_i, β_i 以及标量 k, ℓ, s, t, λ 不加粗。

1. 矩阵梯度及其应用



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

【定义 **1-1**】向量梯度

设 $f: \mathbb{R}^n \to \mathbb{R}$ 是标量函数，输入为列向量 $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T \in \mathbb{R}^n$。定义

$$\nabla_{\boldsymbol{\alpha}} f = \frac{\partial f}{\partial \boldsymbol{\alpha}} = \begin{bmatrix} \frac{\partial f}{\partial \alpha_1} \\ \frac{\partial f}{\partial \alpha_2} \\ \vdots \\ \frac{\partial f}{\partial \alpha_n} \end{bmatrix} \in \mathbb{R}^{n \times 1}.$$

梯度的维度与方向

- $\nabla_{\boldsymbol{\alpha}} f$ 是列向量，与 $\boldsymbol{\alpha}$ 同形；
- 梯度给出函数值增长最快的方向；
- 梯度为零是无约束驻点的必要条件。

【定义 **1-2**】矩阵梯度

设 $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ 是标量函数，输入为矩阵 $\boldsymbol{X} = [X_{ij}] \in \mathbb{R}^{m\times n}$。定义矩阵梯度为

$$\nabla_{\boldsymbol{X}} f = \frac{\partial f}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{m1}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

【定义 1-3】**Frobenius** 范数

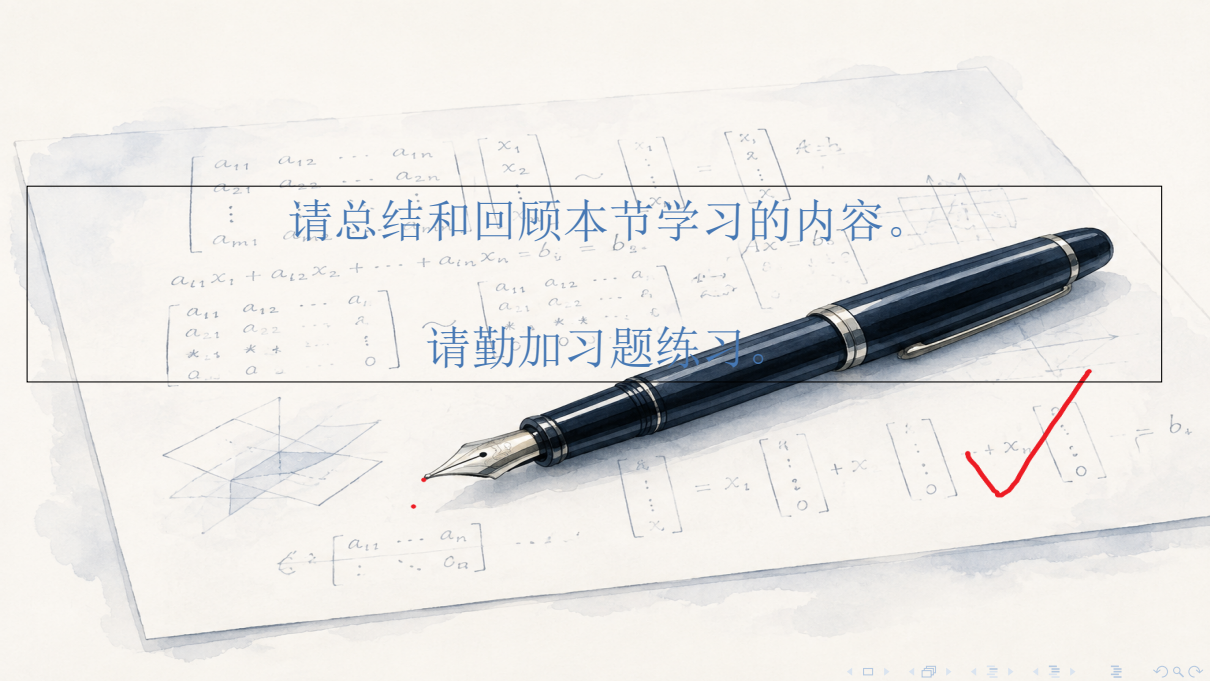
对 $\boldsymbol{X} \in \mathbb{R}^{m \times n}$，定义

$$\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2}.$$

它是矩阵空间中常用的欧氏型范数。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

定理 1-1：线性函数与二次型的梯度

设 $\boldsymbol{\beta}, \boldsymbol{\alpha} \in \mathbb{R}^n$，$\boldsymbol{A} \in \mathbb{R}^{n \times n}$，则

$$\nabla_{\boldsymbol{\alpha}} \left(\boldsymbol{\beta}^T \boldsymbol{\alpha} \right) = \boldsymbol{\beta}, \qquad \nabla_{\boldsymbol{\alpha}} \left(\boldsymbol{\alpha}^T \boldsymbol{A} \boldsymbol{\alpha} \right) = (\boldsymbol{A} + \boldsymbol{A}^T) \boldsymbol{\alpha}.$$

当 $\boldsymbol{A}$ 对称时，

$$\nabla_{\boldsymbol{\alpha}} \left(\boldsymbol{\alpha}^T \boldsymbol{A} \boldsymbol{\alpha} \right) = 2 \boldsymbol{A} \boldsymbol{\alpha}.$$

定理 1-2：仿射复合函数的梯度链式法则

设

$$\boldsymbol{\gamma} = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{\beta}, \qquad f(\boldsymbol{\alpha}) = h(\boldsymbol{\gamma}),$$

其中 $\boldsymbol{A}$ 为常矩阵，则

$$\nabla_{\boldsymbol{\alpha}} f = \boldsymbol{A}^T \nabla_{\boldsymbol{\gamma}} h.$$

例 1-1：线性函数的梯度

设

$$f(\boldsymbol{\alpha}) = \boldsymbol{\beta}^T\boldsymbol{\alpha} = \sum_{i=1}^{n} \beta_i \alpha_i.$$

求 $\nabla_{\boldsymbol{\alpha}} f$。

例 1-1：线性函数的梯度

设

$$f(\boldsymbol{\alpha}) = \boldsymbol{\beta}^T\boldsymbol{\alpha} = \sum_{i=1}^{n} \beta_i \alpha_i.$$

求 $\nabla_{\boldsymbol{\alpha}} f$。

解：对每个分量求偏导，

$$\frac{\partial f}{\partial \alpha_j} = \beta_j, \qquad j = 1, \ldots, n.$$

因此

$$\nabla_{\boldsymbol{\alpha}} f = \boldsymbol{\beta}.$$

例 1-2：一般二次型的梯度

设

$$f(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \boldsymbol{A} \boldsymbol{\alpha},$$

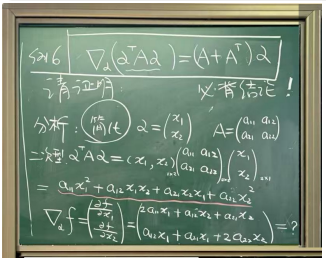
其中 $\boldsymbol{\alpha} \in \mathbb{R}^n$，$\boldsymbol{A} \in \mathbb{R}^{n \times n}$ 不一定对称。求 $\nabla_{\boldsymbol{\alpha}} f$。

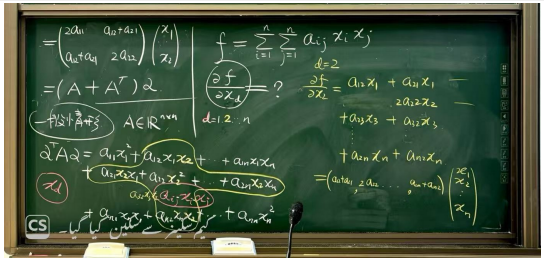
例 1-2：一般二次型的梯度

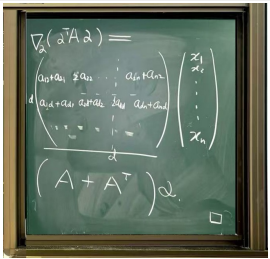
设

$$f(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \boldsymbol{A} \boldsymbol{\alpha},$$

其中 $\boldsymbol{\alpha} \in \mathbb{R}^n$，$\boldsymbol{A} \in \mathbb{R}^{n \times n}$ 不一定对称。求 $\nabla_{\boldsymbol{\alpha}} f$。







解：令

$$\boldsymbol{\alpha} = (x_1, x_2, \ldots, x_n)^T, \qquad \boldsymbol{A} = (a_{ij})_{n\times n}.$$

按分量展开，

$$f(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j.$$

对 x_k 求偏导：

$$\frac{\partial f}{\partial x_k} = \sum_{j=1}^{n} a_{kj}x_j + \sum_{i=1}^{n} a_{ik}x_i = (\boldsymbol{A}\boldsymbol{\alpha})_k + (\boldsymbol{A}^T\boldsymbol{\alpha})_k.$$

因此

$$\nabla_{\boldsymbol{\alpha}} f = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{\alpha}.$$

例 1-3：加权二次型的梯度

设

$$f(\boldsymbol{\alpha}) = (\boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{\beta})^T \boldsymbol{C}(\boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{\gamma}),$$

其中 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，$\boldsymbol{C} \in \mathbb{R}^{m\times p}$，$\boldsymbol{D} \in \mathbb{R}^{p\times n}$。求 $\nabla_{\boldsymbol{\alpha}} f$。

例 1-3：加权二次型的梯度

设

$$f(\boldsymbol{\alpha}) = (\boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{\beta})^T \boldsymbol{C}(\boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{\gamma}),$$

其中 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，$\boldsymbol{C} \in \mathbb{R}^{m\times p}$，$\boldsymbol{D} \in \mathbb{R}^{p\times n}$。求 $\nabla_{\boldsymbol{\alpha}} f$。

解：由定理 4-2，分别对两个仿射表达式使用链式法则，得到

$$\nabla_{\boldsymbol{\alpha}} f = \boldsymbol{A}^T \boldsymbol{C}(\boldsymbol{D}\boldsymbol{\alpha} + \boldsymbol{\gamma}) + \boldsymbol{D}^T \boldsymbol{C}^T (\boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{\beta}).$$

例 1-4：仿射变换后的标量函数

设

$$\boldsymbol{\gamma} = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{\beta}, \qquad f(\boldsymbol{\alpha}) = h(\boldsymbol{\gamma}).$$

求 $\nabla_{\boldsymbol{\alpha}} f$。

例 1-4：仿射变换后的标量函数

设

$$\boldsymbol{\gamma} = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{\beta}, \qquad f(\boldsymbol{\alpha}) = h(\boldsymbol{\gamma}).$$

求 $\nabla_{\boldsymbol{\alpha}} f$。

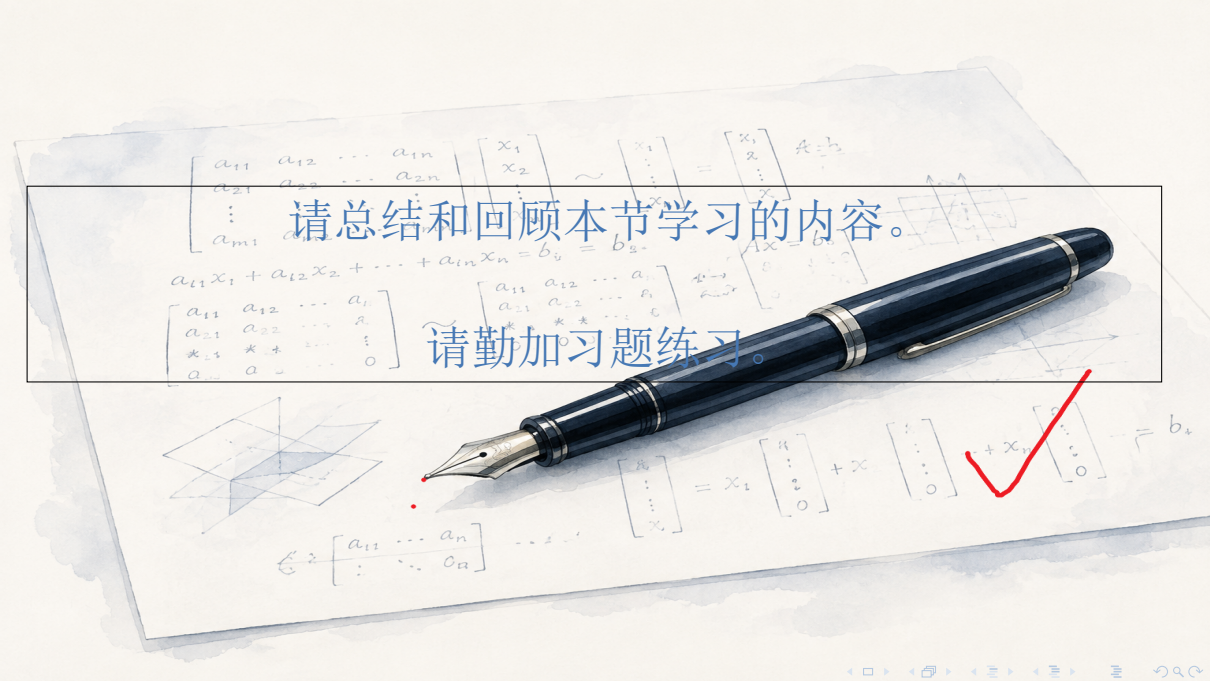
解：应用定理 4-2，

$$\nabla_{\boldsymbol{\alpha}} f = \boldsymbol{A}^T \nabla_{\boldsymbol{\gamma}} h.$$

这说明，先在中间变量 $\boldsymbol{\gamma}$ 下计算梯度，再乘以 $\boldsymbol{A}^T$，即可得到原变量 $\boldsymbol{\alpha}$ 下的梯度。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

定理 1-3：迹表达式的梯度公式

在维度相容的条件下，有

$$\nabla_{\boldsymbol{X}} \operatorname{tr}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}^T,$$

$$\nabla_{\boldsymbol{X}} \operatorname{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}^T\boldsymbol{B}^T,$$

$$\nabla_{\boldsymbol{X}} \operatorname{tr}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{B} + \boldsymbol{A}^T\boldsymbol{X}\boldsymbol{B}^T.$$

矩阵距离的梯度

若 $\boldsymbol{C}$ 为常矩阵，则

$$\nabla_{\boldsymbol{X}} \left(\frac{1}{2} \|\boldsymbol{X} - \boldsymbol{C}\|_F^2 \right) = \boldsymbol{X} - \boldsymbol{C}.$$

例 1-5：矩阵距离的梯度

设 $\boldsymbol{C}$ 为常矩阵，求

$$f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{C}\|_F^2$$

的矩阵梯度。

例 1-5：矩阵距离的梯度

设 $\boldsymbol{C}$ 为常矩阵，求

$$f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{C}\|_F^2$$

的矩阵梯度。

解：由 Frobenius 范数的定义，

$$f(\boldsymbol{X}) = \frac{1}{2}\sum_{i,j}(X_{ij} - C_{ij})^2.$$

逐个元素求偏导，得到

$$\frac{\partial f}{\partial X_{ij}} = X_{ij} - C_{ij}.$$

因此

$$\nabla_{\boldsymbol{X}} f = \boldsymbol{X} - \boldsymbol{C}.$$

例 1-6：迹的二次形式

设

$$f(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}),$$

其中 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$，$\boldsymbol{X} \in \mathbb{R}^{n\times m}$，$\boldsymbol{B} \in \mathbb{R}^{m\times m}$。求 $\nabla_{\boldsymbol{X}} f$。

例 1-6：迹的二次形式

设

$$f(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}),$$

其中 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$，$\boldsymbol{X} \in \mathbb{R}^{n\times m}$，$\boldsymbol{B} \in \mathbb{R}^{m\times m}$。求 $\nabla_{\boldsymbol{X}} f$。

解：直接应用定理 4-3：

$$\nabla_{\boldsymbol{X}} f = \boldsymbol{A}\boldsymbol{X}\boldsymbol{B} + \boldsymbol{A}^T\boldsymbol{X}\boldsymbol{B}^T.$$

若 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 均对称，则

$$\nabla_{\boldsymbol{X}} f = 2\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}.$$

定理 1-4：逆矩阵、行列式与对数行列式

在维度相容且 $\boldsymbol{X}$ 可逆时，

$$\nabla_{\boldsymbol{X}} \operatorname{tr}(\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{B}) = -\boldsymbol{X}^{-T}\boldsymbol{A}^{T}\boldsymbol{B}^{T}\boldsymbol{X}^{-T}.$$

此外，

$$\nabla_{\boldsymbol{X}} \det(\boldsymbol{X}) = \det(\boldsymbol{X})\boldsymbol{X}^{-T}.$$

当 $\boldsymbol{X}$ 正定时，

$$\nabla_{\boldsymbol{X}} \log\det(\boldsymbol{X}) = \boldsymbol{X}^{-T}.$$

使用条件

- 逆矩阵公式要求 $\boldsymbol{X}$ 可逆；
- 对数行列式通常要求 $\boldsymbol{X}$ 为正定矩阵；
- 所有结果的梯度均与 $\boldsymbol{X}$ 同形。

例 1-7：迹中包含逆矩阵

设

$$f(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{B}),$$

其中 $\boldsymbol{X}$ 可逆。求 $\nabla_{\boldsymbol{X}} f$。

例 1-7：迹中包含逆矩阵

设

$$f(\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{B}),$$

其中 $\boldsymbol{X}$ 可逆。求 $\nabla_{\boldsymbol{X}} f$。

解：由定理 4-4，

$$\nabla_{\boldsymbol{X}} f = -\boldsymbol{X}^{-T}\boldsymbol{A}^{T}\boldsymbol{B}^{T}\boldsymbol{X}^{-T}.$$

例 1-8：对数行列式

设 $\boldsymbol{X}$ 为正定矩阵，求

$$f(\boldsymbol{X}) = \log\det(\boldsymbol{X})$$

的矩阵梯度。

例 1-8：对数行列式

设 $\boldsymbol{X}$ 为正定矩阵，求

$$f(\boldsymbol{X}) = \log\det(\boldsymbol{X})$$

的矩阵梯度。

解：由定理 4-4，

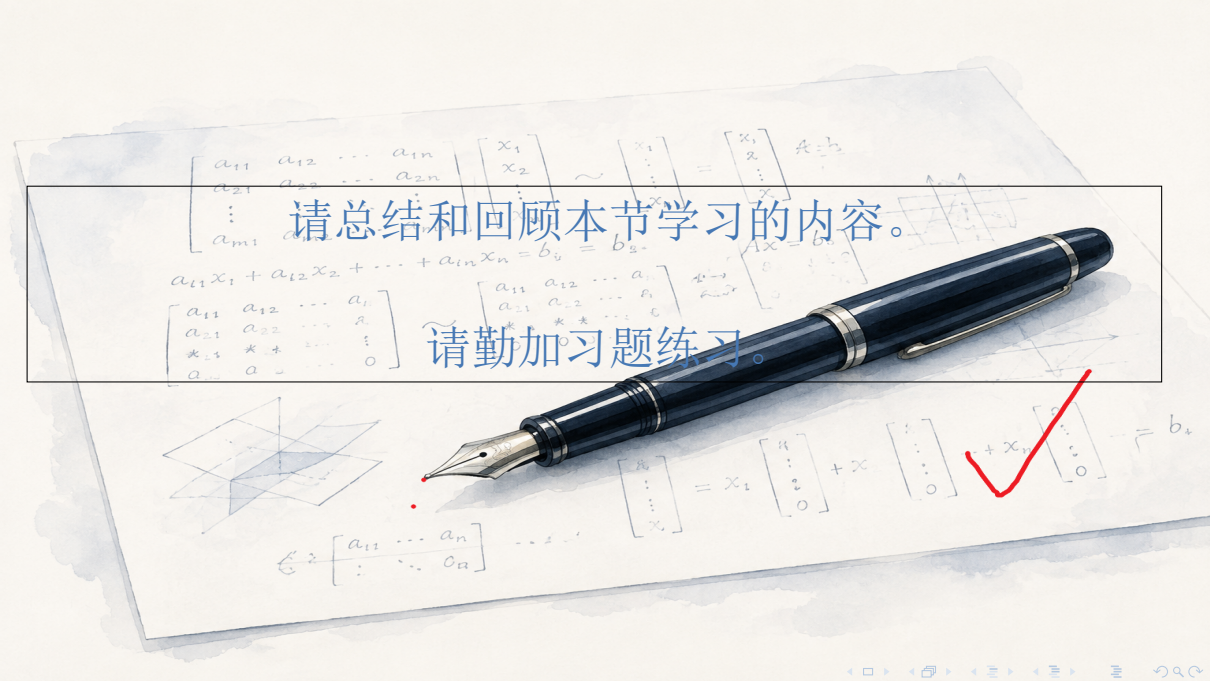
$$\nabla_{\boldsymbol{X}} f = \boldsymbol{X}^{-T}.$$

由于 $\boldsymbol{X}$ 对称时 $\boldsymbol{X}^{-T} = \boldsymbol{X}^{-1}$，正定矩阵情形下也常写成

$$\nabla_{\boldsymbol{X}} f = \boldsymbol{X}^{-1}.$$

请总结和回顾本节学习的内容。

请勤加习题练习。



2. Outline

1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

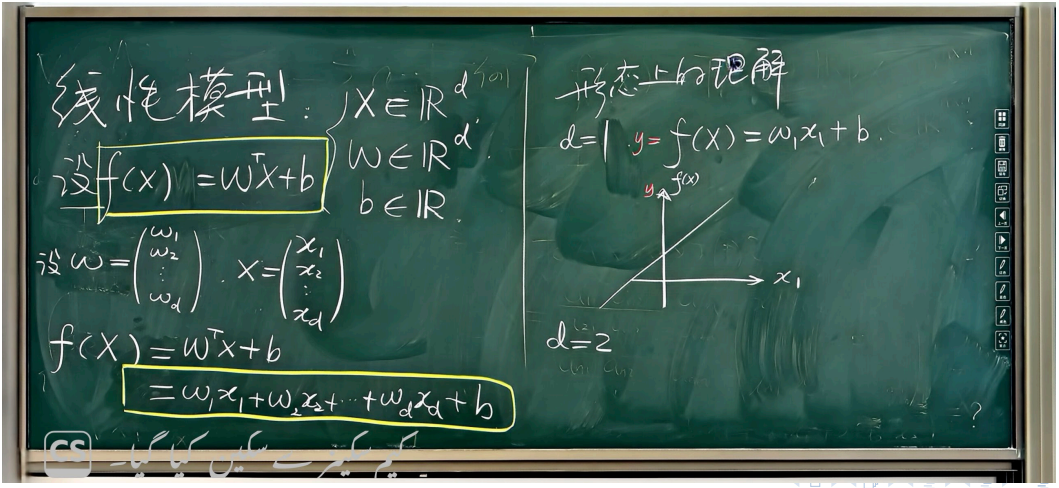
2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

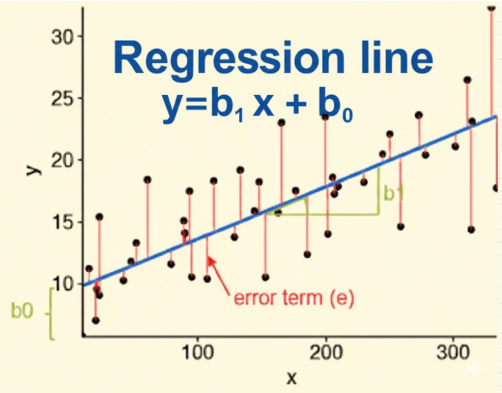
线性回归是一个经典的回归学习算法，
有利于理解 Loss 函数和正则项的作用，
这其中需要适量的矩阵微分的基础。

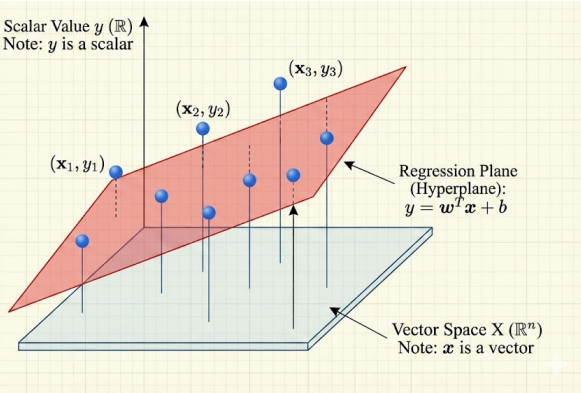
同时，也需要针对 $y = \boldsymbol{w}^T \boldsymbol{x} + b$获得一个高维空间的理解。

如何理解 $y = \boldsymbol{w}^T\boldsymbol{x} + b$ 这个记号？



如何理解 $y = \boldsymbol{w}^T\boldsymbol{x} + b$ 这个记号？





【定义 **2-1**】回归

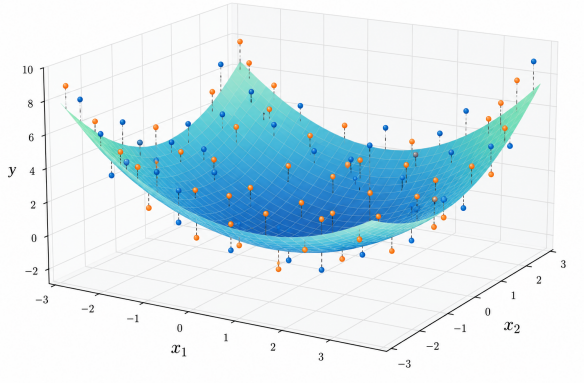
回归是一类监督学习任务：给定样本

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}, \qquad \boldsymbol{x}_i \in \mathbb{R}^p, \quad y_i \in \mathbb{R},$$

学习一个映射 $f : \mathbb{R}^p \to \mathbb{R}$，使其能够根据输入特征 $\boldsymbol{x}_i$ 预测相应的连续响应值 y_i。也就是说，希望满足

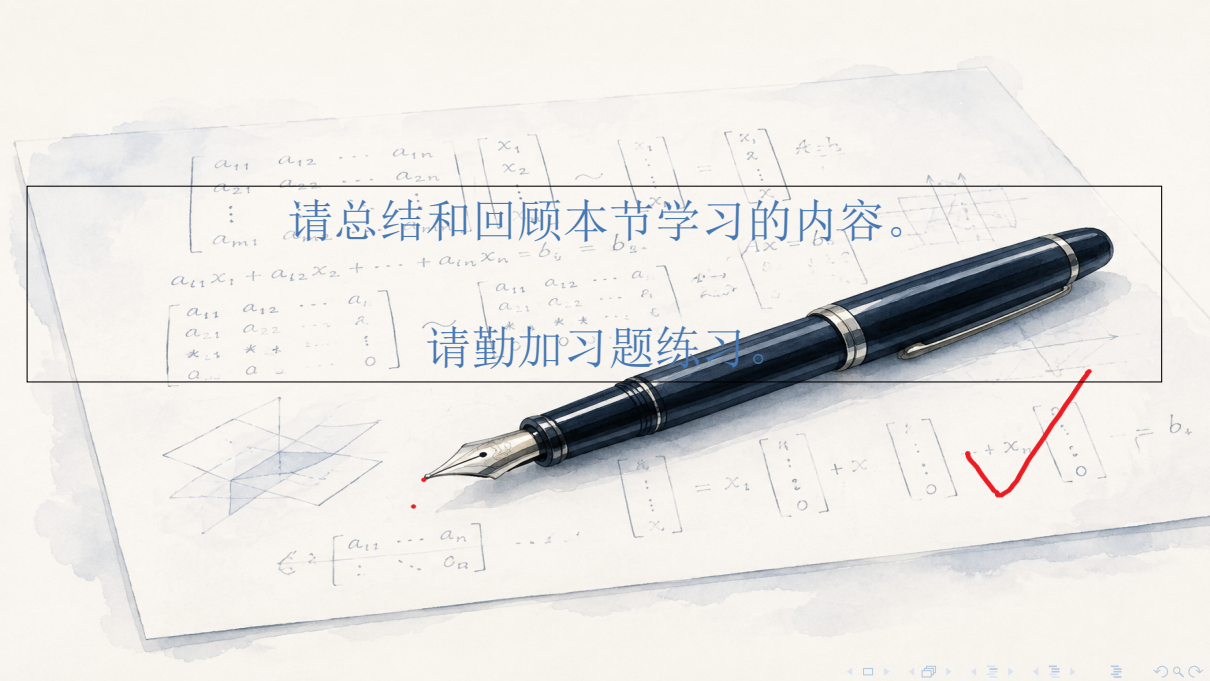
$$y_i \approx f(\boldsymbol{x}_i), \qquad i = 1, \ldots, n.$$

观测值集合 Obs: $\{\boldsymbol{x}_i, y_i\}$.
在回归问题中，给定 n 个样本 $\{\boldsymbol{x}_i, y_i\}$。其中，$\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T \in \mathbb{R}^p$ 是第 i 个样本的特征向量，$y_i \in \mathbb{R}$ 是相应的真实响应值。模型根据 $\boldsymbol{x}_i$ 给出预测值 $\hat{y}_i = f(\boldsymbol{x}_i, \boldsymbol{w})$，其中 $\boldsymbol{w}$ 表示模型参数。



请总结和回顾本节学习的内容。

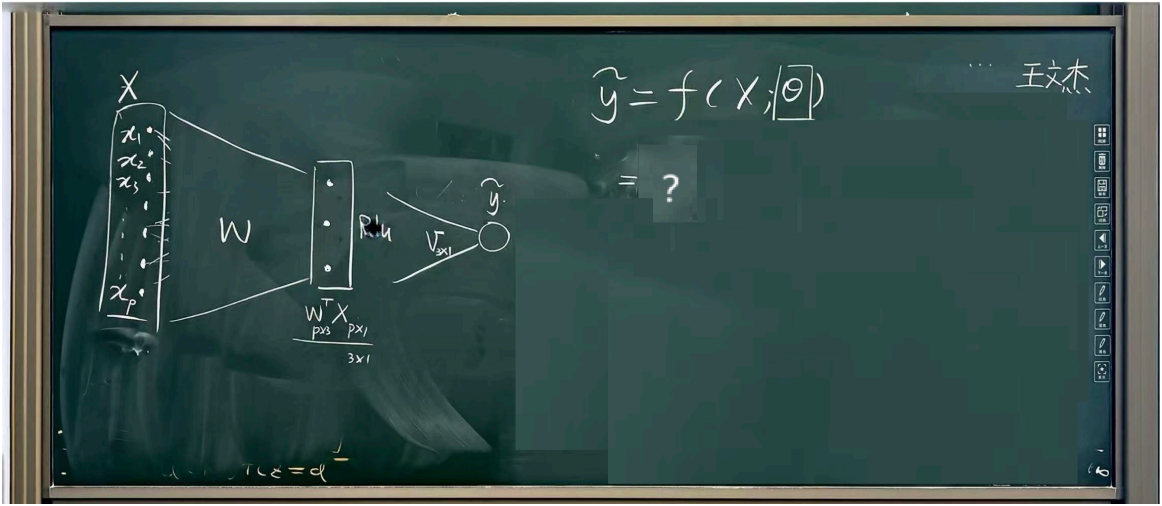
请勤加习题练习。

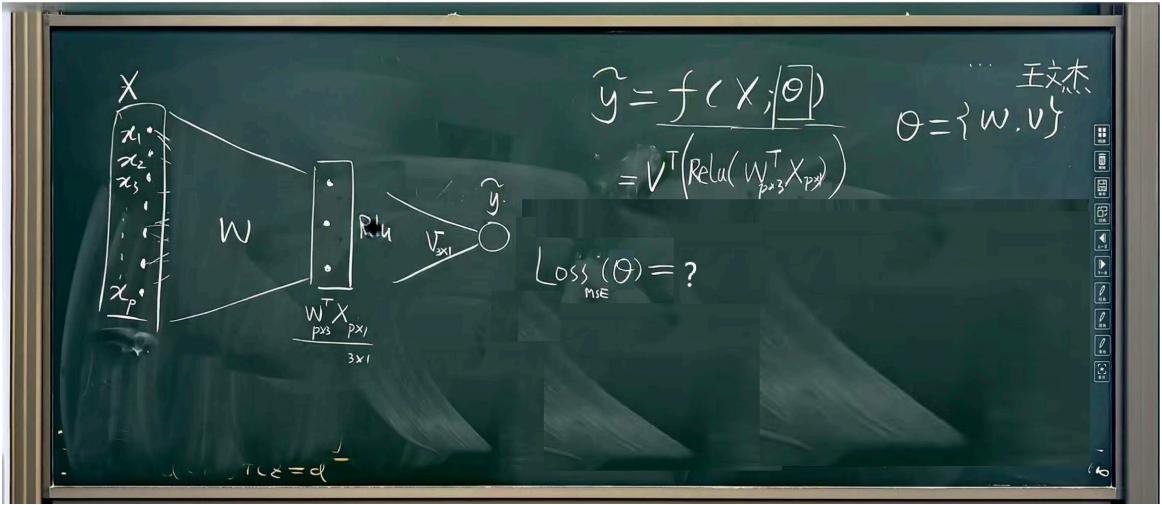


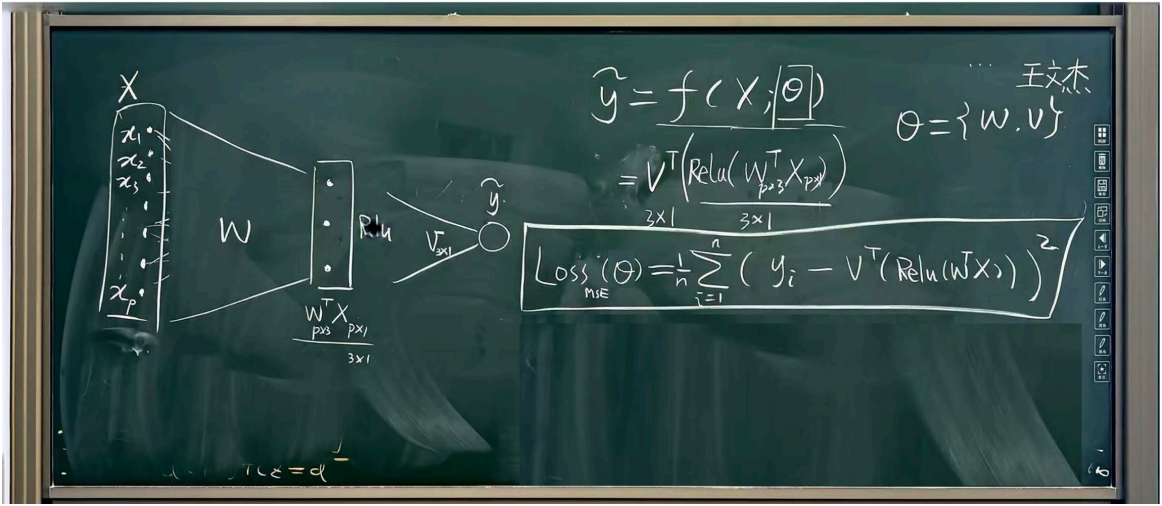
1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

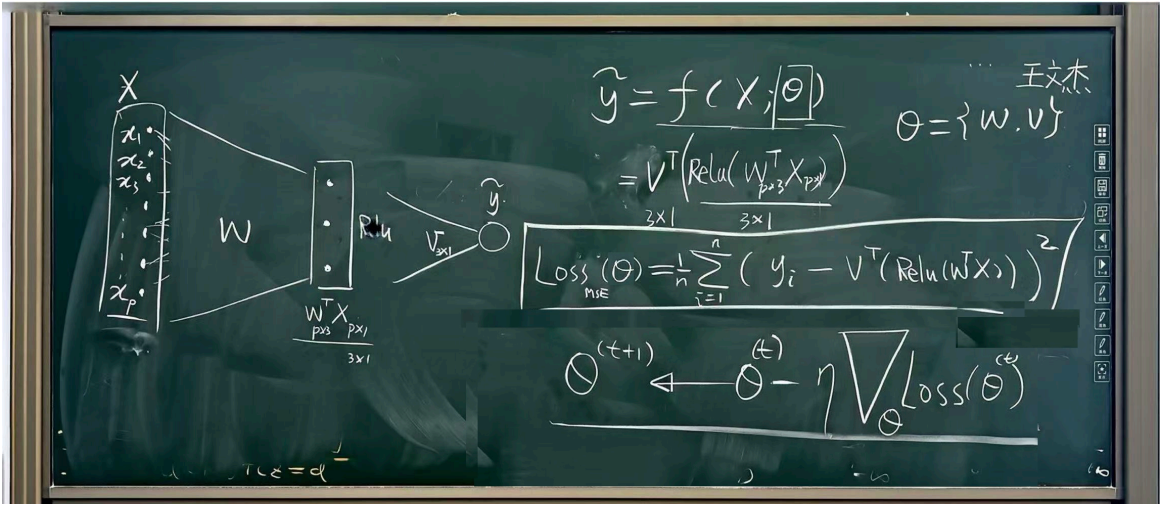
2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258


X
W
ReLU
$\hat{y}$
$W^T X$
王文杰
$\hat{y} = f(X;\theta)$
= ?


X
W
Relu
$\hat{y}$
$W^T X$
$\hat{y} = f(X;\theta)$
$= V^T(\mathrm{Relu}(W^T X))$
王文杰
$\theta = \{W, V\}$
$\mathrm{Loss}_{\mathrm{MSE}}(\theta) = ?$

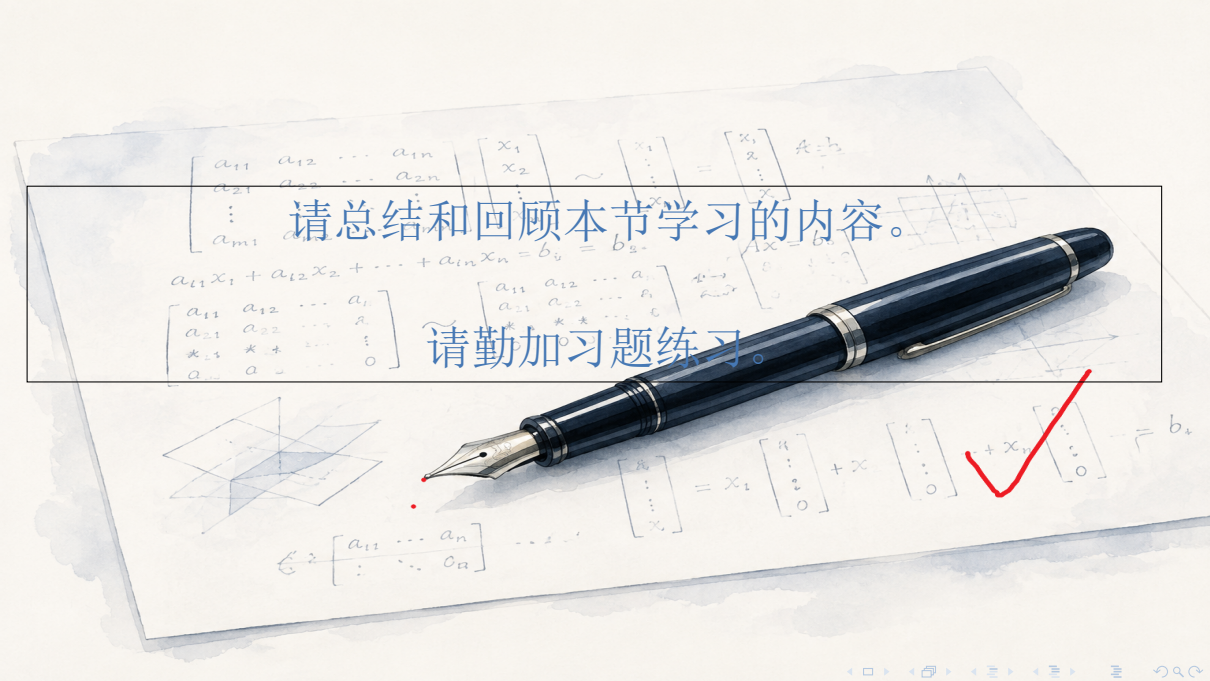

王文杰
$\hat{y} = f(X;\Theta)$
$= V^T(\mathrm{Relu}(W^T_{p\times 3} X_{p\times 1}))$
3×1
3×1
$\Theta = \{W, V\}$
$\mathrm{Loss}_{\mathrm{MSE}}(\Theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - V^T(\mathrm{Relu}(W^T X_i)))^2$
X
x_1
x_2
x_3
x_p
W
Relu
$V_{3\times 1}$
$\hat{y}$
$W^T X$
$p\times 3$ $p\times 1$
3×1


X
W
Relu
$\hat{y}$
$W^T X$
$\hat{y} = f(X;\theta)$
$= V^T(\mathrm{Relu}(W^T X))$
王文杰
$\theta = \{W, V\}$
$\mathrm{Loss}_{\mathrm{MSE}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - V^T(\mathrm{Relu}(W^T X_i)))^2$
$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \nabla_\theta \mathrm{Loss}(\theta^{(t)})$

在线性回归中，复杂模型 $f(\boldsymbol{x};\theta)$ 将简化为 $f(\boldsymbol{x};\boldsymbol{w},b)=\boldsymbol{w}^T\boldsymbol{x}+b$；相应地，模型参数也由 θ 简化为 $\boldsymbol{w}$（以及截距 b）。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

【定义 2-2】均方误差（**MSE**）

均方误差（Mean Squared Error, MSE）是回归任务中最常用的损失函数之一，其标准定义为：

$$\mathcal{J}_{MSE}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\boldsymbol{x}_i, \boldsymbol{w})\right)^2$$

其中，n 为样本数量，y_i 为第 i 个样本的真实值，$f(\boldsymbol{x}_i, \boldsymbol{w})$ 为模型在参数 $\boldsymbol{w}$ 下对输入 $\boldsymbol{x}_i$ 的预测值。

MSE 的优点

MSE 具有以下显著优点，使其在众多场景中成为首选：

- 处处可导性：MSE 函数在其定义域内处处连续可微，这一光滑特性使其能够与基于梯度的优化算法（如随机梯度下降）完美协同，便于求解最优模型参数。
- 对大误差的强惩罚：其损失计算中的平方项 $(y_i - \hat{y}_i)^2$ 使得预测值与真实值之间的较大偏差会被指数级放大。这一特性驱使模型在训练过程中优先修正严重的预测错误，有利于快速降低整体误差水平。
- 凸函数性：当模型 $f(\boldsymbol{x}, \boldsymbol{w})$ 为参数的线性函数时，MSE 损失函数是关于参数 $\boldsymbol{w}$ 的凸函数。凸性保证了优化过程能够收敛至全局最优解，避免了陷入局部极值点的问题，为参数估计提供了理论上的可靠性。

MSE 的局限与缺点

尽管优势突出，MSE 也存在固有的局限性，在使用时需审慎考虑：

- 对异常值高度敏感：平方项的放大效应是一把“双刃剑”。当数据中存在异常值（Outliers）时，其所产生的巨大误差会被平方操作进一步放大，导致损失函数值被少数异常样本所主导。这可能会使模型参数为拟合这些异常点而发生严重偏离，从而损害模型在正常数据上的泛化性能。
- 损失量纲问题：由于对误差进行了平方运算，MSE 损失值的量纲变为原目标变量量纲的平方。这为损失值的直观理解和不同量纲数据集间的横向比较带来了不便。有时会采用其平方根形式（RMSE）以恢复量纲的一致性。

与其他回归损失函数的对比

表 1: 常见回归损失函数对比

损失函数名称	数学表达式（简写）	核心特性与适用场景
均方误差（**MSE**）	$\frac{1}{n}\sum(y-\hat{y})^2$	对大误差惩罚严厉，收敛快，但对异常值敏感。适用于噪声较小、异常值少的回归问题。
平均绝对误差（**MAE**）	$\frac{1}{n}\sum\lvert y-\hat{y}\rvert$	对误差进行线性惩罚，对异常值的鲁棒性更强，但在零点不可导，优化相对平缓。
Huber Loss	$\begin{cases}\frac{1}{2}(y-\hat{y})^2, & 若\lvert y-\hat{y}\rvert\le\delta\\ \delta\lvert y-\hat{y}\rvert-\frac{1}{2}\delta^2, & 其他\end{cases}$	MSE 与 MAE 的折衷，在误差较小时为二次，较大时为线性，兼具光滑性与鲁棒性。

【定义 **2-3**】欧几里得范数（L_2 范数）

对于一个 p 维向量 $\boldsymbol{w} = (w_1, w_2, \ldots, w_p)^T \in \mathbb{R}^p$，其欧几里得范数定义为：

$$\|\boldsymbol{w}\|_2 = \sqrt{\sum_{j=1}^{p} w_j^2} = \sqrt{w_1^2 + w_2^2 + \cdots + w_p^2}$$

使用矩阵表示法，它可以简洁地写作：

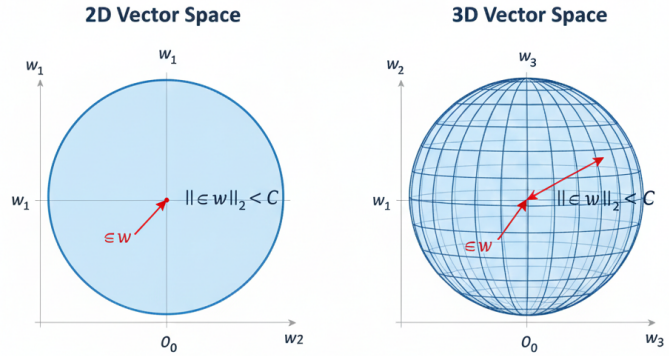
$$\|\boldsymbol{w}\|_2 = \sqrt{\boldsymbol{w}^T \boldsymbol{w}}$$

核心特性与几何意义

几何长度：

在二维空间（$p = 2$）中，这正是勾股定理：如果向量是 $\boldsymbol{w} = (x, y)^T$，那么其长度就是 $\sqrt{x^2 + y^2}$。

范数球 (Norm Ball)：$\|\boldsymbol{w}\|_2 \leq C$：
在二维情况下，所有满足条件的点构成一个实心圆盘。
在三维情况下，所有满足条件的点构成一个实心球体。



例 2-1：L_2 范数的计算

给定 $\boldsymbol{w} = (3, 4)^T$，计算 $\|\boldsymbol{w}\|_2$，并判断它是否属于范数球 $\{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \leq 5\}$。

解：

$$\|\boldsymbol{w}\|_2 = \sqrt{3^2 + 4^2} = 5.$$

因此 $\boldsymbol{w}$ 位于半径为 5 的 L_2 范数球边界上。

【定义 **2-4**】一范数（L_1 范数）

对于一个 p 维向量 $\boldsymbol{w} = (w_1, w_2, \ldots, w_p)^T \in \mathbb{R}^p$，其一范数定义为各分量绝对值之和：

$$\|\boldsymbol{w}\|_1 = \sum_{j=1}^{p} |w_j| = |w_1| + |w_2| + \cdots + |w_p|.$$

例 2-2：L_1 范数的计算

给定 $\boldsymbol{w} = (3, 4)^T$，计算 $\|\boldsymbol{w}\|_1$，并判断它是否属于范数球 $\{\boldsymbol{w} : \|\boldsymbol{w}\|_1 \leq 5\}$。

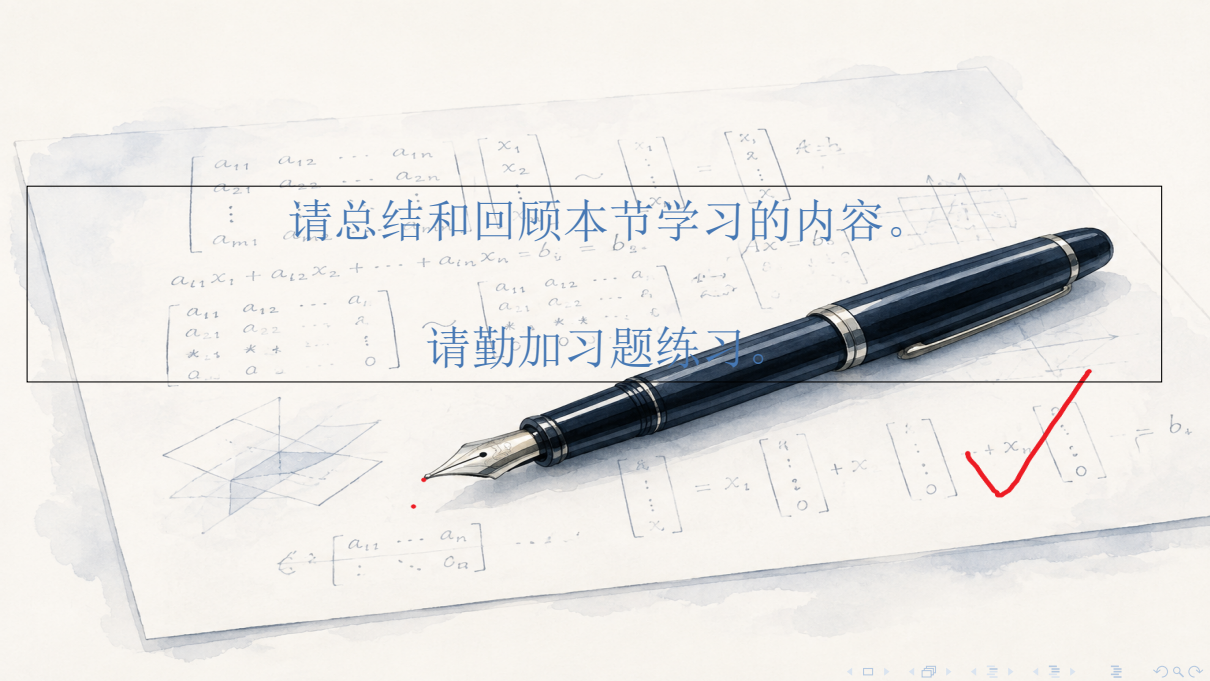
解：

$$\|\boldsymbol{w}\|_1 = |3| + |4| = 7.$$

由于 $7 > 5$，因此 $\boldsymbol{w}$ 不属于半径为 5 的 L_1 范数球。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

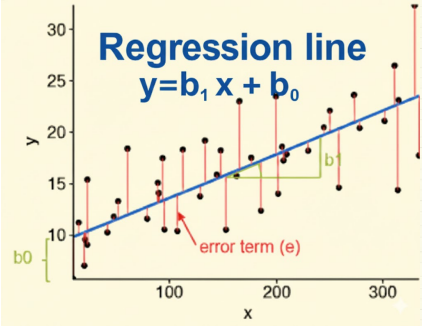
设原始特征 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T \in \mathbb{R}^d$，斜率参数 $\boldsymbol{w}_0 = (w_1, \ldots, w_d)^T$，截距为标量 b。完整的线性模型是

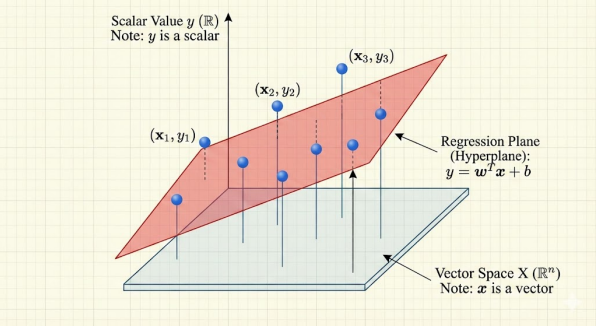
$$\hat{y}_i = \boldsymbol{w}_0^T \boldsymbol{x}_i + b.$$

把常数 1 加在每个特征向量末尾，同时把 b 加在参数向量末尾：

$$\hat{\boldsymbol{x}}_i = \begin{pmatrix} \boldsymbol{x}_i \\ 1 \end{pmatrix}, \qquad \hat{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{w}_0 \\ b \end{pmatrix}, \qquad \hat{y}_i = \hat{\boldsymbol{w}}^T \hat{\boldsymbol{x}}_i.$$

因此增广设计矩阵 $\boldsymbol{X} = (\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_n)^T$ 的最后一列全为 1。后文将增广后的 $\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{w}}$ 简记为 $\boldsymbol{x}_i, \boldsymbol{w}$，并以 $p = d + 1$ 表示其维数。


Regression line
y=b1 x + b0
error term (e)
b1
b0
y
x
30
25
20
15
10
100
200
300


Scalar Value y ($\mathbb{R}$)
Note: y is a scalar
$(\mathbf{x}_3, y_3)$
$(\mathbf{x}_2, y_2)$
$(\mathbf{x}_1, y_1)$
Regression Plane
(Hyperplane):
$y = \boldsymbol{w}^T \boldsymbol{x} + b$
Vector Space X ($\mathbb{R}^n$)
Note: $\boldsymbol{x}$ is a vector

Mean Squared Error (**MSE**) is the minimum goal.

$$\mathcal{J}_{MSE}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \boldsymbol{w}))^2 \tag{1}$$

那么线性回归的优化目标就是寻找最佳的参数 $\boldsymbol{w}^*$, 使得

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^p} \mathcal{J}_{MSE}(\boldsymbol{w})$$

Mean Squared Error (**MSE**) is the minimum goal.

$$\mathcal{J}_{MSE}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 \tag{2}$$

命题 2-1：均方误差的矩阵表示

令

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n\times p}, \qquad \boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n.$$

则

$$\mathcal{J}_{MSE}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}).$$

命题 2-2：MSE 关于参数的凸性

设线性模型为 $f(\boldsymbol{x}_i, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{x}_i$，则均方误差

$$\mathcal{J}_{MSE}(\boldsymbol{w}) = \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

关于参数 $\boldsymbol{w}$ 是凸函数。当 $\boldsymbol{X}^T\boldsymbol{X}$ 正定时，$\mathcal{J}_{MSE}$ 是严格凸函数。

证明：对 $0 \leq t \leq 1$，有

$$\mathcal{J}_{MSE}\big((1-t)\boldsymbol{w}_1 + t\boldsymbol{w}_2\big) = (1-t)\mathcal{J}_{MSE}(\boldsymbol{w}_1) + t\mathcal{J}_{MSE}(\boldsymbol{w}_2) - \frac{t(1-t)}{n}\big\|\boldsymbol{X}(\boldsymbol{w}_1 - \boldsymbol{w}_2)\big\|_2^2.$$

最后一项非负，因此 $\mathcal{J}_{MSE}$ 凸；当 $\boldsymbol{X}^T\boldsymbol{X}$ 正定且 $\boldsymbol{w}_1 \neq \boldsymbol{w}_2$ 时，最后一项严格为正，故 $\mathcal{J}_{MSE}$ 严格凸。

实际上，MSE 可以另写成：

$$\mathcal{J}_{MSE}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) := \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

这里 $\|\cdot\|_2$ 表示 L_2 范数。[a]

[a]此处请留意向量范数 l_2被用以表示目标优化函数。当然，这仅仅只是一个记号上的使用而已。

命题 2-3：最小二乘解的闭式表达

若 $\boldsymbol{X}^T\boldsymbol{X}$ 可逆，则均方误差的唯一最小点为

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{3}$$

证明见下一页。

证明：
令目标函数关于 $\boldsymbol{w}$ 的梯度为零：

$$
\begin{aligned}
0 &= \nabla_{\boldsymbol{w}} \mathcal{J}_{MSE}(\boldsymbol{w}) \\
&= \nabla_{\boldsymbol{w}} \left(\frac{1}{n} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \right) \\
&= \frac{1}{n} \nabla_{\boldsymbol{w}} \left((\boldsymbol{y}^T - \boldsymbol{w}^T \boldsymbol{X}^T)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \right) \\
&= \frac{1}{n} \nabla_{\boldsymbol{w}} \left(\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} \right) \\
&= ?
\end{aligned}
\tag{4}
$$

继续证明：
继续计算目标函数关于 $\boldsymbol{w}$ 的梯度：

$$
\begin{aligned}
0 &= \nabla_{\boldsymbol{w}} \mathcal{J}_{MSE}(\boldsymbol{w}) \\
&= \nabla_{\boldsymbol{w}} \left(\frac{1}{n} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \right) \\
&= \frac{1}{n} \nabla_{\boldsymbol{w}} \left((\boldsymbol{y}^T - \boldsymbol{w}^T \boldsymbol{X}^T)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \right) \\
&= \frac{1}{n} \nabla_{\boldsymbol{w}} \left(\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} \right) \\
&= \frac{1}{n} (0 - \boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w}) \\
&= \frac{2}{n} (-\boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w})
\end{aligned}
\tag{5}
$$

Let the gradient equal to zero, we have

$$
\hat{\boldsymbol{w}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{6}
$$

例 2-3：四点数据的普通最小二乘拟合

给定四个观测点 $(-1,-1),(0,0),(1,1),(2,6)$。用直线 $y=wx+b$ 拟合这些数据。请构造含截距的设计矩阵 $\boldsymbol{X}$ 和观测向量 $\boldsymbol{y}$，求普通 MSE 的最小点 $\boldsymbol{w}^*=(w^*,b^*)^T$ 与拟合直线。

例 2-3：四点数据的普通最小二乘拟合

对 $(-1,-1),(0,0),(1,1),(2,6)$ 用 $y = wx + b$ 拟合，构造 $\boldsymbol{X}, \boldsymbol{y}$，求普通 MSE 的最小点和拟合直线。

解：把每个输入 x_i 与常数 1 排在同一行，得到

$$\boldsymbol{X} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 6 \end{pmatrix}, \qquad \boldsymbol{w} = \begin{pmatrix} w \\ b \end{pmatrix}.$$

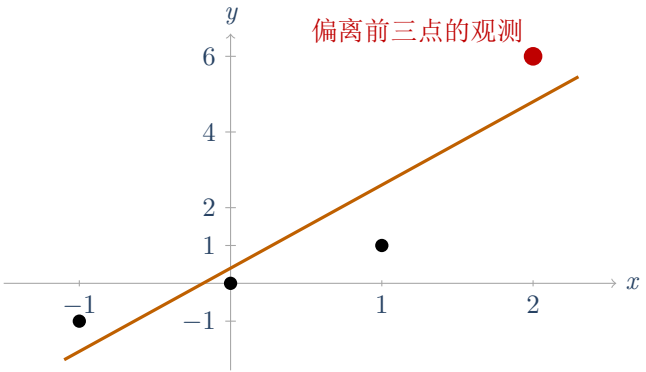
于是 $\boldsymbol{X}\boldsymbol{w}$ 的第 i 个分量正是 $wx_i + b$。

由最小二乘正规方程，

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 6 & 2 \\ 2 & 4 \end{pmatrix}, \qquad \boldsymbol{X}^T\boldsymbol{y} = \begin{pmatrix} 14 \\ 6 \end{pmatrix}.$$

解 $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{X}^T\boldsymbol{y}$，得

$$\boldsymbol{w}^* = \begin{pmatrix} 11/5 \\ 2/5 \end{pmatrix}, \qquad y = \frac{11}{5}x + \frac{2}{5}.$$


y
偏离前三点的观测
6
4
2
1
−1
−1
1
2
x
普通 MSE：$y = \frac{11}{5}x + \frac{2}{5}$

例 2-4：房价预测中的最小二乘回归

给定 $n = 4$ 套房屋的数据，特征为面积和到地铁站的距离。采用含截距的模型 $y = w_{\text{area}} x_{\text{area}} + w_{\text{dist}} x_{\text{dist}} + b$。请构造设计矩阵 $\boldsymbol{X}$ 与观测向量 $\boldsymbol{y}$，求最小二乘参数 $\boldsymbol{w}^*$，并解释各参数的含义。

原始数据 (Obs)
假设我们收集了 4 套公寓的数据：$\{ \boldsymbol{x}_i, y_i \}$, $i = 1, 2, 3, 4$
样本 1: 面积 50m^2, 距离 1km $\rightarrow$ 售价 200 万
样本 2: 面积 60m^2, 距离 2km $\rightarrow$ 售价 230 万
样本 3: 面积 80m^2, 距离 1km $\rightarrow$ 售价 310 万
样本 4: 面积 90m^2, 距离 5km $\rightarrow$ 售价 300 万

例 2-4：房价预测中的最小二乘回归

仍使用上一页的四套房屋数据。下面计算 $\boldsymbol{X}^T\boldsymbol{X}$、$\boldsymbol{X}^T\boldsymbol{y}$ 和 $\boldsymbol{w}^*$，其中模型包含截距 b；解释所得回归系数。

解：每行依次放面积、距离与常数 1，得到

$$\boldsymbol{X} = \begin{pmatrix} 50 & 1 & 1 \\ 60 & 2 & 1 \\ 80 & 1 & 1 \\ 90 & 5 & 1 \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} 200 \\ 230 \\ 310 \\ 300 \end{pmatrix}, \qquad \boldsymbol{w} = \begin{pmatrix} w_{\text{area}} \\ w_{\text{dist}} \\ b \end{pmatrix}.$$

因此 $\boldsymbol{X}\boldsymbol{w}$ 的每个分量都包含同一个截距 b。

计算正规方程两侧：

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 20600 & 700 & 280 \\ 700 & 31 & 9 \\ 280 & 9 & 4 \end{pmatrix}, \qquad \boldsymbol{X}^T\boldsymbol{y} = \begin{pmatrix} 75600 \\ 2470 \\ 1040 \end{pmatrix}.$$

例如，常数列对应的乘积包括

$$(\boldsymbol{X}^T\boldsymbol{X})_{13} = \sum_{i=1}^{4} x_{i,\text{area}} = 280, \qquad (\boldsymbol{X}^T\boldsymbol{y})_3 = \sum_{i=1}^{4} y_i = 1040.$$

解 $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{X}^T\boldsymbol{y}$，得到

$$\boldsymbol{w}^* = \frac{1}{39}\begin{pmatrix} 140 \\ -440 \\ 1330 \end{pmatrix} \approx \begin{pmatrix} 3.59 \\ -11.28 \\ 34.10 \end{pmatrix}.$$

所以拟合函数为

$$\hat{y} = \frac{140}{39}x_{\text{area}} - \frac{440}{39}x_{\text{dist}} + \frac{1330}{39}.$$

面积系数 $w^*_{\text{area}} \approx 3.59$：保持距离不变，面积每增加 $1\,\text{m}^2$，预测房价增加约 3.59 万元。
距离系数 $w^*_{\text{dist}} \approx -11.28$：保持面积不变，距离每增加 $1\,\text{km}$，预测房价减少约 11.28 万元。
截距 $b^* \approx 34.10$ 万元是模型在面积与距离均为零时的预测值。
这个输入超出样本范围，不能把截距直接解释成实际的基础房价。

残差向量为

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}^* = \frac{1}{13}\begin{pmatrix} -30 \\ 40 \\ 0 \\ -10 \end{pmatrix} \neq \mathbf{0}.$$

模型没有穿过全部四个样本点，但它使普通 MSE 取得最小值。

关于这个结果的几个总结和思考：

- 首先，$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$是一个显式解，该模型的求解过程中适度使用了一些矩阵微分的技巧。

关于这个结果的几个注解：

- 首先，$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$是一个显式解，该模型的求解过程中适度使用了一些矩阵微分的技巧。
- 其次，这个模型及其求解所对应的“机器学习”思想在哪里？

关于这个结果的几个注解：

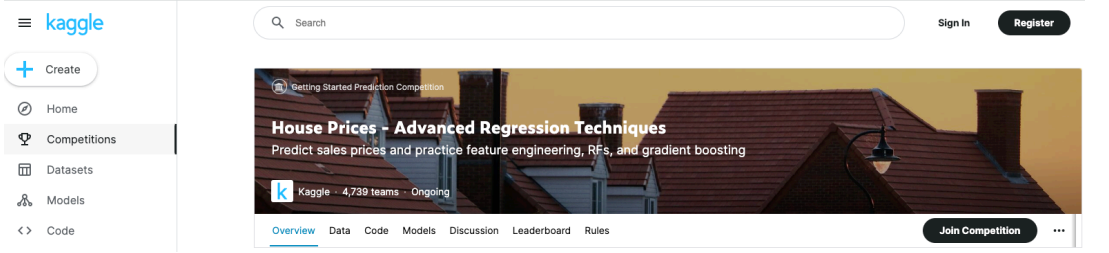
- 首先，$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$是一个显式解，该模型的求解过程中适度使用了一些矩阵微分的技巧。
- 其次，这个模型及其求解所对应的“机器学习”思想在哪里？
- 但是，这个解并不完美。为什么？tips: if $p > n$...

关于这个结果的几个注解：

- 首先，$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$是一个显式解，该模型的求解过程中适度使用了一些矩阵微分的技巧。
- 其次，这个模型及其求解所对应的“机器学习”思想在哪里？
- 但是，这个解并不完美。为什么？tips: if $p > n$...
- 一个细节：针对 $\mathcal{J}_{MSE}(\boldsymbol{w})$ 的梯度下降是如何进行的，公式是？

算法竞赛和更多的思考

Kaggle 竞赛，Boston 房价预测 [a] [b].



[a] https://www.kaggle.com/c/house-prices-advanced-regression-techniques
[b] 百度搜索。

算法比拼，考虑：

- 运算结果在测试集上的准确性以及可迁移性（鲁棒性），
- 时间复杂度和算法效率 (寻优效率)，
- 结果的深层分析（可解释性）.

如果上述梯度等于 0，无法求出最佳解。有哪些办法？

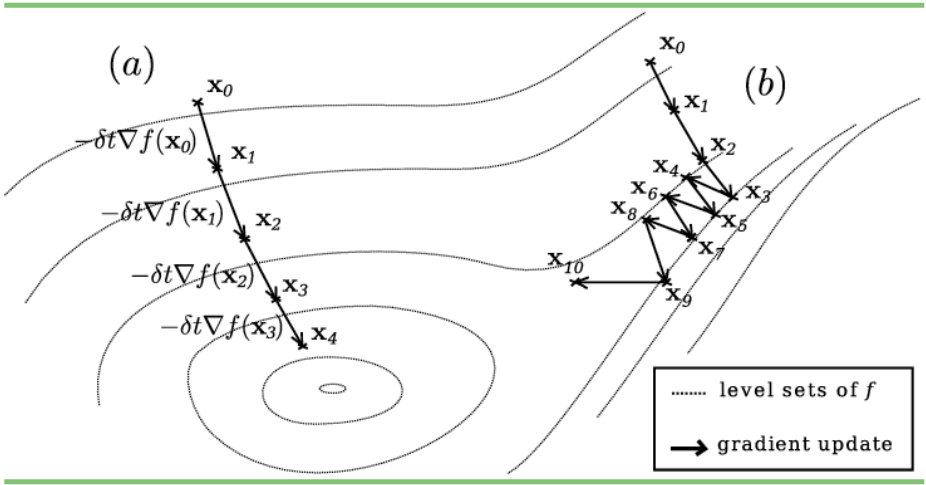
Gradient descent (aka., steepest descent) is for minimizing multidimensional smooth convex objective functions of the form $\mathcal{J}: \mathbb{R}^p \to \mathbb{R}$.

算法 2-4：梯度下降

1: **Input:** Initial point $\boldsymbol{w}^{(0)}$, gradient norm tolerance ε
2: Set $t = 0$
3: **while** $\|\nabla \mathcal{J}_{MSE}(\mathbf{w}^{(t)})\| \geq \varepsilon$ **do**
4: $\quad \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta^{(t)} \nabla \mathcal{J}_{MSE}(\mathbf{w}^{(t)})$
5: $\quad t = t + 1$
6: **end while**
7: **Return:** $\boldsymbol{w}^{(t)}$

这里最关键的迭代步骤是：

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta^{(t)} \nabla \mathcal{J}_{MSE}(\mathbf{w}^{(t)}) \tag{7}$$


(a)
$\mathbf{x}_0$
$-\delta t \nabla f(\mathbf{x}_0)$
$\mathbf{x}_1$
$-\delta t \nabla f(\mathbf{x}_1)$
$\mathbf{x}_2$
$-\delta t \nabla f(\mathbf{x}_2)$
$\mathbf{x}_3$
$-\delta t \nabla f(\mathbf{x}_3)$
$\mathbf{x}_4$
(b)
$\mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_4$
$\mathbf{x}_5$
$\mathbf{x}_6$
$\mathbf{x}_7$
$\mathbf{x}_8$
$\mathbf{x}_9$
$\mathbf{x}_{10}$
level sets of f
gradient update

问：当 $\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ 无法计算，具体该如何使用梯度下降？

问：在使用 Pytorch 编写神经网络的时候，哪些部分是与梯度下降有关？

问：在使用 Pytorch 编写神经网络的时候，哪些部分是与梯度下降有关？

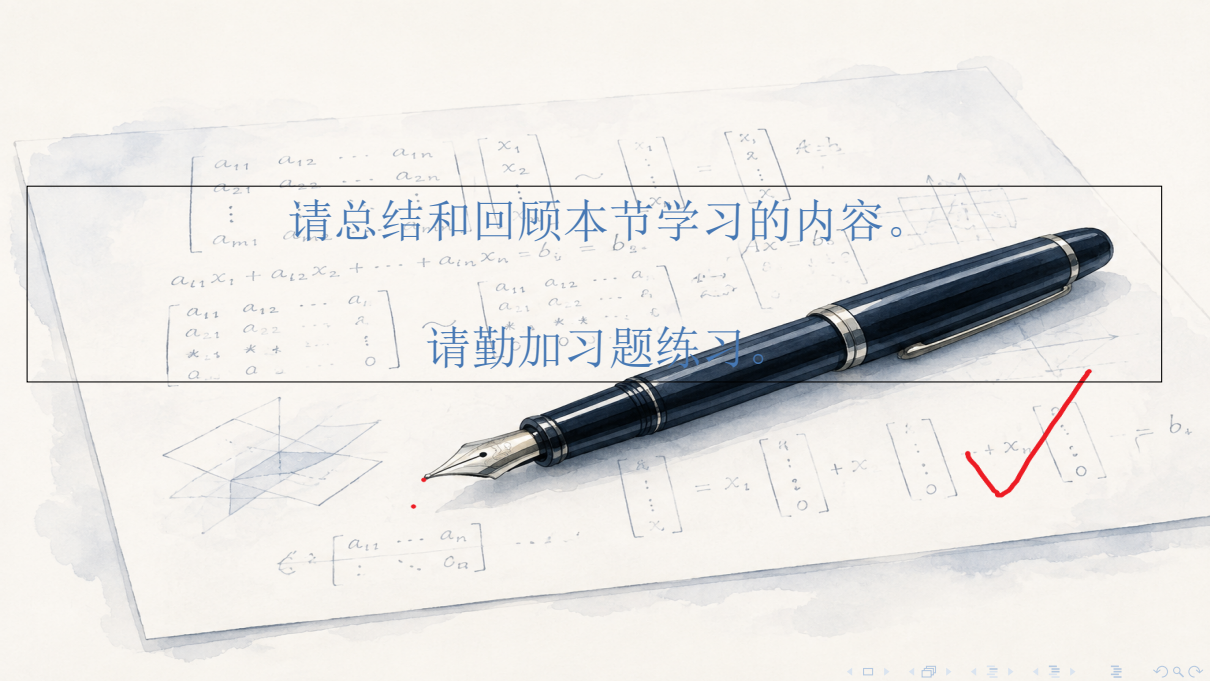
```
import torch
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim
from torch.utils.data import DataLoader, TensorDataset
import matplotlib.pyplot as plt
```

问：在使用 Pytorch 编写神经网络的时候，哪些部分是与梯度下降有关？

```
optimizer = optim.SGD(
    model.parameters(),
    lr=0.01,
    momentum=0.9,
    weight_decay=1e-4
)
```

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

回忆最小二乘线性回归的目标函数

The square sum of error terms is the minimum goal.

$$\mathcal{J}_{MSE}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 \tag{8}$$

回忆最小二乘线性回归的目标函数

The square sum of error terms is the minimum goal.

$$\mathcal{J}_{MSE}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 \tag{9}$$

考虑最小二乘线性回归的变种，**ridge regression/**岭回归, 它引入一个$L2$ regularizer (正则项):

$$\mathcal{J}_{Ridge}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 + \lambda\|\boldsymbol{w}\|_2^2. \tag{10}$$

几个问题：

增添了正则项后，对比最小二乘下的线性回归 $\boldsymbol{w}$ 估计结论，对这个结论的影响在哪里？

对 $\mathcal{J}_{MSE}(\boldsymbol{w})$ 的梯度下降训练影响在哪里？
这是可以通过列式求解获知的。请尝试。

考虑最小二乘线性回归的变种，**ridge regression**/岭回归, 它引入一个$L2$ regularizer (正则项):

$$\mathcal{J}_{Ridge}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2 + \lambda\|\boldsymbol{w}\|_2^2. \text{ —Ridge regression.} \tag{11}$$

a b

[a]凸规划的结论告诉我们，增加了正则项后，以上 Loss 函数的最优化问题等同于

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{w}))^2, \qquad \text{s.t.} \quad \|\boldsymbol{w}\|_2^2 \le C. \tag{12}$$

此处 C与 Lagrange 常数 λ有关.

[b]增加的正则项使得解搜索需要在 $\|\boldsymbol{w}\|_2^2 \le C$ 的范围内进行。这减小了可行解的搜索空间。

命题 2-5：Ridge 惩罚形式的约束解释

设 $\lambda \geq 0$，并令 $\boldsymbol{w}_\lambda$ 是

$$\min_{\boldsymbol{w}} \left\{ \mathcal{J}_{MSE}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2 \right\}$$

的一个最优解。取 $C_\lambda = \|\boldsymbol{w}_\lambda\|_2^2$，则 $\boldsymbol{w}_\lambda$ 也是问题

$$\min_{\boldsymbol{w}} \mathcal{J}_{MSE}(\boldsymbol{w}), \qquad \text{s.t. } \|\boldsymbol{w}\|_2^2 \leq C_\lambda$$

的一个最优解；在满足凸性和适当约束资格条件时，反向结论也成立。

证明： Ridge 目标函数是约束问题的拉格朗日函数。由最优性条件，任何满足 $\|\boldsymbol{w}\|_2^2 \leq C_\lambda$ 的参数都不会得到比 $\boldsymbol{w}_\lambda$ 更小的原始目标函数值；反向结论由凸优化的 KKT 条件给出。

正则项的 **min-max** 算法实质
构造 Lagrange 函数引入拉格朗日乘子 $\lambda \geq 0$，将约束项挂载到目标函数之后：

$$L(\boldsymbol{w}, \lambda) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda(\|\boldsymbol{w}\|_1 - C)$$

这里 λ 的物理意义是：为了满足约束，我们愿意在原始残差目标上付出的“代价权重”。
整理为 Min-Max 形式按照我们在最大分隔面中讨论的逻辑，Lasso 问题可以写成：

$$\min_{\boldsymbol{w}} \left[\max_{\lambda \geq 0} L(\boldsymbol{w}, \lambda) \right]$$

原因：如果违背约束 ($\|\boldsymbol{w}\|_1 > t$)，即括号内 $(\|\boldsymbol{w}\|_1 - t) > 0$。此时 $\max_{\lambda \geq 0}$ 操作会通过让 $\lambda \to \infty$，使整个函数值变为无穷大。如果满足约束 ($\|\boldsymbol{w}\|_1 \leq t$)，即括号内 ≤ 0。
为了达到最大值，最优的 λ 必须被挤压到 0（或者在边界上取某个值），此时函数回归到原始的 $\frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$。结论：这个 Min-Max 结构迫使优化过程在 $\boldsymbol{w}$ 空间中寻找那个既能让残差最小，又不会让 λ 逃向无穷大的解。

实际上，，我们通常不直接解这个 Min-Max 问题，而是利用 Lagrange 对偶性。当我们固定一个 λ 值（将其视为超参数）时，目标函数变为：

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2+\lambda\|\boldsymbol{w}\|_1-\text{const.}$$

这样我们就得到了刚才的目标函数。

例 2-5：LASSO 稀疏性的几何解释



在很多问题求解的情况下，我们希望线性回归能获得一个包含很多 0 分量的 $\boldsymbol{w}$，以达到稀疏性。为什么会这样？

怎么理解一个具有稀疏性特征的 $\boldsymbol{w}$？例如从图示中可以观察到什么？

——服务于可解释性。

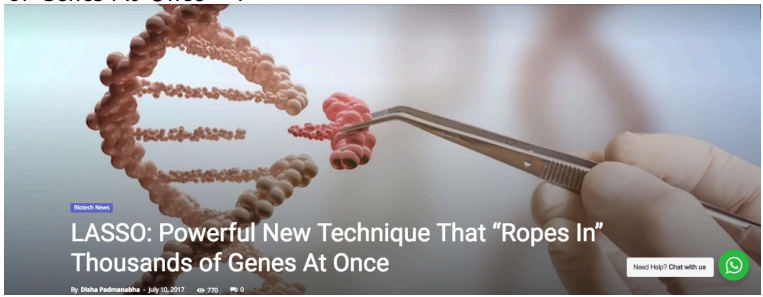
例 2-5：LASSO 稀疏性的几何解释

请结合 L_1 范数球的几何形状，说明 LASSO 为什么容易产生零分量，以及这种稀疏性如何服务于特征选择与模型解释。

解：L_1 范数球的边界具有位于坐标轴上的尖角，而损失函数的等高线往往先与这些尖角相交，因此最优参数向量常出现零分量。这使 LASSO 能够在拟合数据的同时筛选重要特征。

Why do we like to have sparsity in $\boldsymbol{w}$? Here is an example:

Associate genotypes to a given phenotype. Ref "LASSO: Powerful New Technique That 'Ropes In' Thousands of Genes At Once" [a].



[a] https://www.biotecnika.org/2017/07/lasso-powerful-new-technique-that-ropes-in-thousands-of-genes-at-once/

LASSO: Least absolute shrinkage and selection operator.
利用 L_1范数，我们获得 **Lasso Regression** 的目标函数。

$$\mathcal{J}_{Lasso}(\boldsymbol{w}) = \mathcal{J}_{MSE}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1. \tag{13}$$

a b c

[a] Equivalently, from a view of convex optimization, the minimization of the above loss function suffices to:

$$\min_{\boldsymbol{w}} \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2, \qquad \text{s.t.} \quad \|\boldsymbol{w}\|_1 < C. \tag{14}$$

Here, C is a constant, related to λ.

[b] 请对照 Ridge 回归，LASSO 的唯一区别在于正则项中的范数选择。不同的范数选择，为什么会带来 $\boldsymbol{w}$ 的稀疏性呢?

[c] 实际上，$L0$范数会带来更加直接的特征约减。选用 L_1的原因是为了计算的可能性。

L_1 范数（L_1 Norm），也称为曼哈顿距离（Manhattan Distance）。它与之前讨论的 L_2 范数在几何形状和物理特性上有显著区别。

前面已经给出了 L_1 范数的定义。下面重点讨论其几何形状、稀疏性与稳健性：

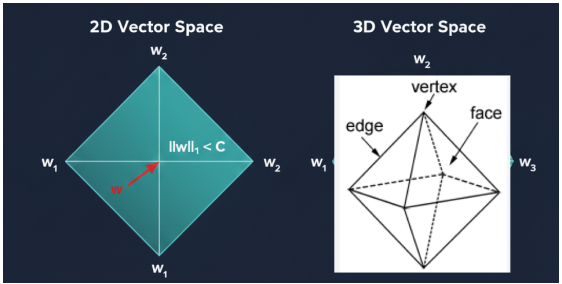
L_1 范数的关键特性:

1. 几何形状：菱形与正八面体

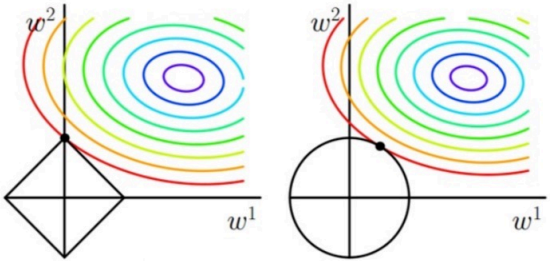
L_1 范数球 $\|\boldsymbol{w}\|_1 \leq C$ 的边界在不同维度下呈现出“尖锐”的特征：

二维 ($p = 2$)：是一个旋转了 45 度的正方形（菱形）。其顶点位于坐标轴上，例如 $(C, 0), (0, C), (-C, 0), (0, -C)$。

三维 ($p = 3$)：是一个正八面体（Regular Octahedron）。



2. 稀疏性 (Sparsity) ——这是 L_1 范数在机器学习中被广泛应用的核心原因。
直观：由于 L_1 范球的“角”正好位于坐标轴上，寻优过程中，等高线往往最先接触到这些“角”。
结果：最优解 $\boldsymbol{w}^*$ 中许多分量为 0。常被用作特征选择，从大量特征中筛选重要部分。



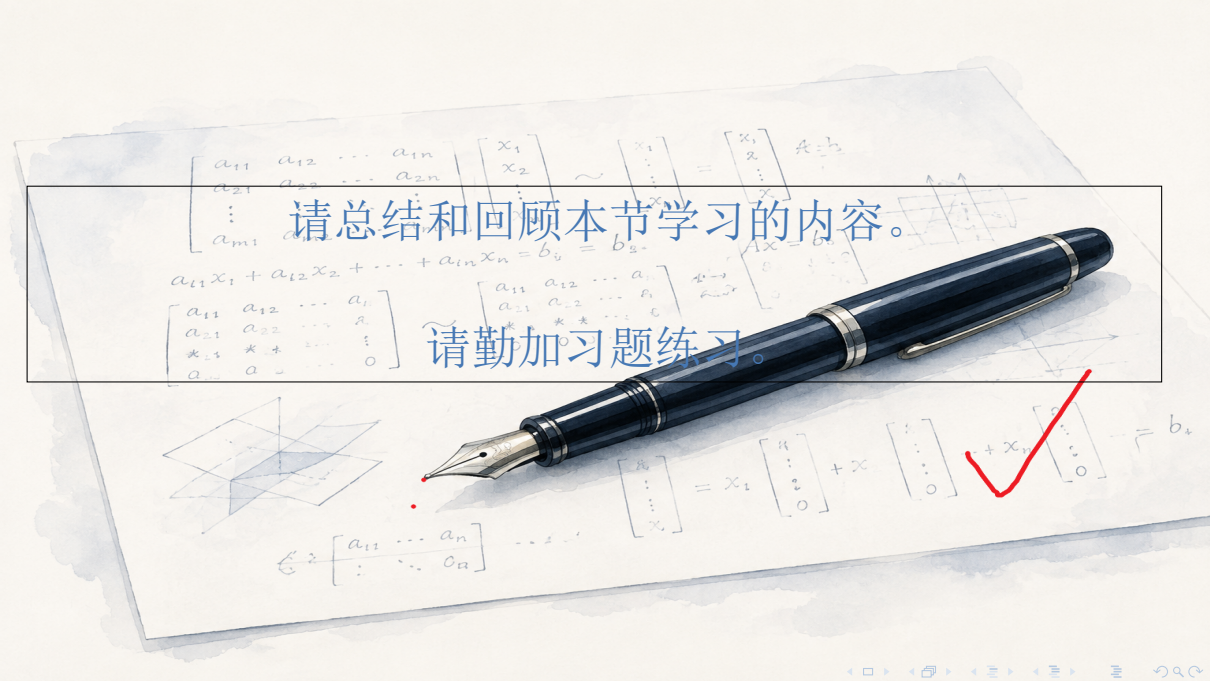
(a) ℓ_1-ball meets quadratic function. ℓ_1-ball has corners. It's very likely that the meet-point is at one of the corners.

(b) ℓ_2-ball meets quadratic function. ℓ_2-ball has no corner. It is very unlikely that the meet-point is on any of axes."

3. 稳健性 (Robustness)
相比于 L_2 范数对误差进行平方（这会极大地放大离群点/Outliers 的影响），L_1 范数对异常值不那么敏感，因此在处理含有噪声或离群点的数据时更加稳健。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

例 2-6：LASSO 回归的 Python 实现

给定下列样本数据，使用 Python 实现 LASSO 回归，并观察正则化参数 λ 变化时回归系数的变化。

解：示例代码来自知乎[a]。**Command lines**

```
>git clone https://github.com/PytLab/MLBox.git
```

The raw data are:

Raw data for regression

```
1       0.455   0.365   0.095   0.514   0.2245  0.101   0.15    15
1       0.35    0.265   0.09    0.2255  0.0995  0.0485  0.07    7
-1      0.53    0.42    0.135   0.677   0.2565  0.1415  0.21    9
1       0.44    0.365   0.125   0.516   0.2155  0.114   0.155   10
0       0.33    0.255   0.08    0.205   0.0895  0.0395  0.055   7
0       0.425   0.3     0.095   0.3515  0.141   0.0775  0.12    8
-1      0.53    0.415   0.15    0.7775  0.237   0.1415  0.33    20
...
```

[a] https://zhuanlan.zhihu.com/p/30535220

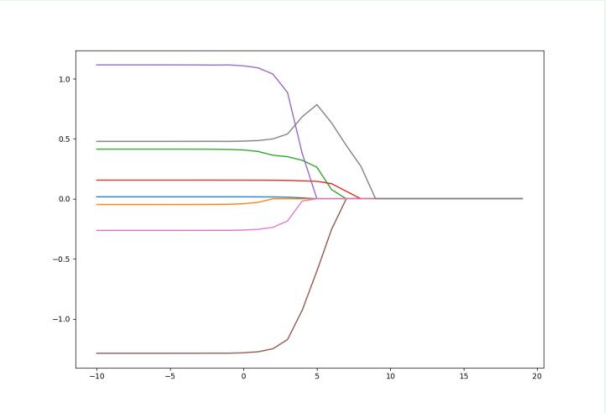
例 2-6：LASSO 回归的 Python 实现

比较不同 λ 下的权重输出，说明 LASSO 如何逐步将部分系数压缩为 0。

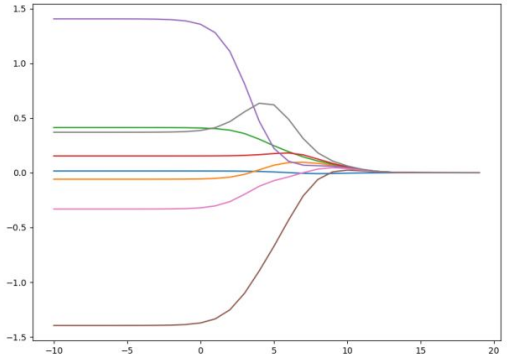
解： w approximate 0, while λ increases. **Command lines**

```
lambda = e^(0), w = [[ 0.0164  -0.0412  0.4066   0.1553  1.1076 -1.2825  -0.2605   0.4802]]
lambda = e^(1), w = [[ 0.0161 -0.0295  0.3941  0.1550   1.0905 -1.2741  -0.2547  0.4847]]
lambda = e^(2), w = [[ 0.0153   0.     0.3626  0.1542  1.0391 -1.2494  -0.2378  0.4987]]
lambda = e^(3), w = [[ 0.01325  0.     0.3505  0.1528  0.8850 -1.1720  -0.1846  0.5397]]
lambda = e^(4), w = [[ 0.0076  0.          0.3209  0.1497  0.3782 -0.9284  -0.0188  0.6836]]
lambda = e^(5), w = [[ 0.   0.     0.2627  0.1453  0.         -0.6018   0.      0.7843]]
lambda = e^(6), w = [[ 0.    0.     0.0766  0.1260  0.          -0.2552   0.    0.6322]]
lambda = e^(7), w = [[ 0.     0.    0.  0.0628  0.  0. 0.   0.4449]]
lambda = e^(8), w = [[ 0.   0.     0.   0.    0.  0. 0.   0.2707]]
lambda = e^(9), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(10), w = [[ 0.  0.  0.  0.  0.  0.  0. 0.]]
lambda = e^(11), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(12), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(13), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(14), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(15), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(16), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(17), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(18), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
lambda = e^(19), w = [[ 0.  0.  0.  0.  0.  0.  0.  0.]]
```

解： LASSO regression: w_i approximates zero when λ increases.

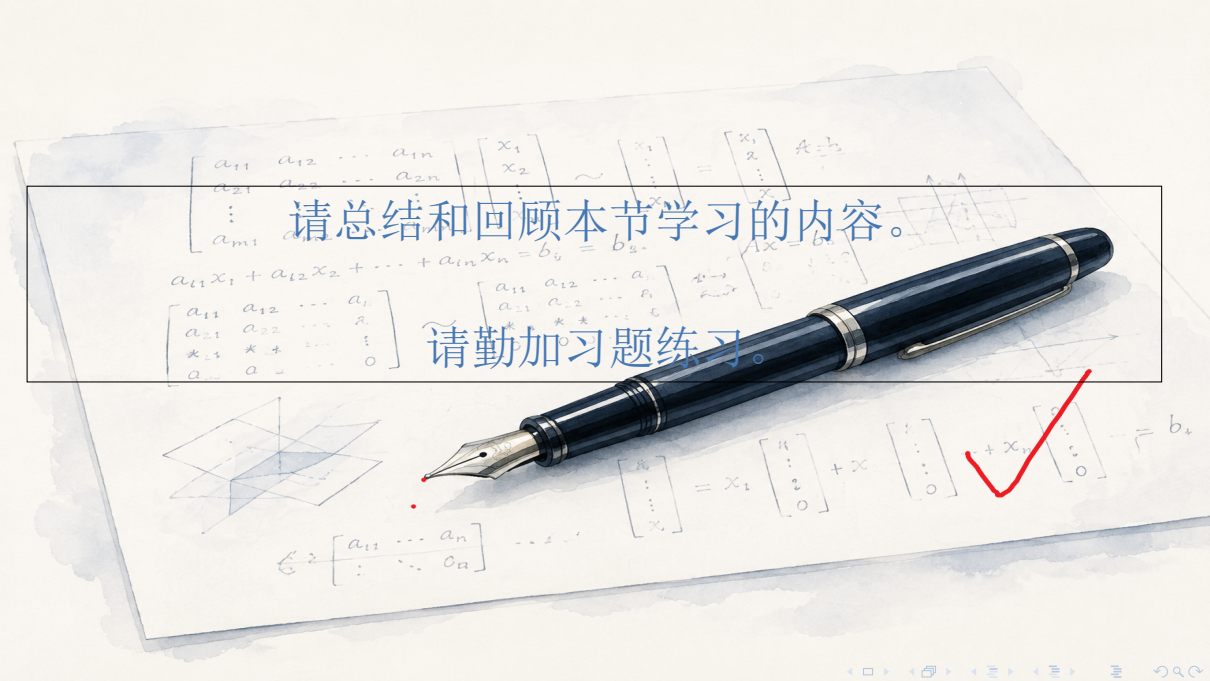


解： Ridge regression: w_i doesn't approximates zero very quickly when λ increases.



请总结和回顾本节学习的内容。

请勤加习题练习。



通过线性回归的模型的学习，我们对模型的定义、求解等诸多细节有了条理清晰的理解。有同学评论中问道一有趣的问题：

“为什么要介绍 ridge 回归？如果我们需引入 LASSO 回归的时候用处时，ridge 回归的意义何在？”

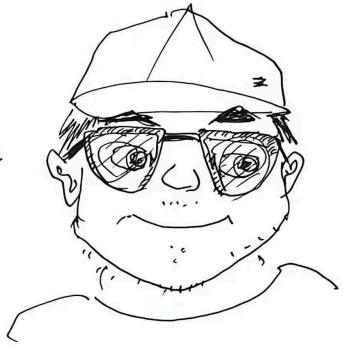


这是一个有趣味的问题。有这么一些解读：

① 首先，我们已明白，在最小二乘情况下，
特征为零的时候下降得的解并不真能舍弃。

② 动机：如果 Cost 函数换成 ridge 版本，
特征为零的分析结果是否产生了变化？

③ 换言，不论 ridge 还是 lasso，他都给予了
带噪的优化问题 $\tilde{L}$ 把权衡明白事实情况
下的 min-max 解说。这个思路是普适的。



3. Outline

1. 矩阵梯度及其应用 3
   - 梯度的定义与记号 6
   - 向量变量的梯度公式 11
   - 矩阵变量的梯度公式 24
2. 线性回归 39
   - 线性回归与一般意义下的回归问题 40
   - 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
   - 回归问题中的记号准备 52
   - 线性回归与最小二乘法 66
   - 正则化线性回归（Ridge 与 LASSO） 100
   - LASSO 回归与 Python 代码示例 117
3. 范数理论及其应用 127
   - 向量范数 128
   - 矩阵诱导范数与加权线性回归 168
   - 向量范数的等价性* 187
   - 矩阵范数 200
   - 矩阵的诱导范数 210
   - 三种相容的矩阵范数 232
   - Frobenius 范数的迹运算* 258

1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

本章围绕三个问题展开：

- 如何在抽象线性空间中衡量向量的大小？
- 改变误差的衡量方式，会怎样影响线性回归的拟合结果？
- 如何用矩阵范数描述矩阵对向量的作用？

我们先从坐标空间中的收敛与常用向量范数出发，再推广到抽象空间；随后用加权回归展示范数的建模作用，最后讨论范数等价与矩阵范数。

带 * 的章节为略讲内容，并以浅蓝底色标示。

【定义 **3-1**】$\mathbb{C}^n$ 中向量序列的收敛

设

$$\boldsymbol{w}^{(k)} = \left(w_1^{(k)}, \ldots, w_n^{(k)}\right)^T, \qquad \boldsymbol{w} = \left(w_1, \ldots, w_n\right)^T.$$

若

$$\lim_{k\to\infty} w_i^{(k)} = w_i \qquad (i = 1, 2, \ldots, n),$$

则称 $\{\boldsymbol{w}^{(k)}\}$ 收敛于 $\boldsymbol{w}$，记作

$$\lim_{k\to\infty} \boldsymbol{w}^{(k)} = \boldsymbol{w} \qquad \text{或} \qquad \boldsymbol{w}^{(k)} \longrightarrow \boldsymbol{w} \quad (k \to \infty).$$

【定义 **3-2**】$\mathbb{C}^n$ 中向量的模

对 $\boldsymbol{w}=(w_1,\ldots,w_n)^T$，定义

$$|\boldsymbol{w}|=\left(|w_1|^2+\cdots+|w_n|^2\right)^{1/2}.$$

命题 3-1：向量序列收敛的坐标与模判据

$\boldsymbol{w}^{(k)} \to \boldsymbol{w}$ 当且仅当每个坐标数列都收敛于相应坐标；等价地，$\left|\boldsymbol{w}^{(k)} - \boldsymbol{w}\right| \to 0$。

证明由有限项非负数之和趋于零当且仅当每一项都趋于零，

$$\begin{aligned}
\lim_{k\to\infty} \boldsymbol{w}^{(k)} = \boldsymbol{w} &\iff \lim_{k\to\infty} w_i^{(k)} = w_i \quad (i = 1, \dots, n) \\
&\iff \lim_{k\to\infty} \left| w_i^{(k)} - w_i \right|^2 = 0 \quad (i = 1, \dots, n) \\
&\iff \lim_{k\to\infty} \left| \boldsymbol{w}^{(k)} - \boldsymbol{w} \right| = 0.
\end{aligned}$$

【定义 **3-3**】向量范数

设 V 是数域 $\mathbb{K}$ 上的线性空间。对任意 $\boldsymbol{\alpha}, \boldsymbol{\beta} \in V$，定义实数 $\|\boldsymbol{\alpha}\|$，并要求

$$
\begin{aligned}
&(1) \quad \|\boldsymbol{\alpha}\| \geq 0, \qquad \|\boldsymbol{\alpha}\| = 0 \Longleftrightarrow \boldsymbol{\alpha} = \mathcal{O};\\
&(2) \quad \|k \odot \boldsymbol{\alpha}\| = |k| \, \|\boldsymbol{\alpha}\|, \qquad \forall k \in \mathbb{K};\\
&(3) \quad \|\boldsymbol{\alpha} \oplus \boldsymbol{\beta}\| \leq \|\boldsymbol{\alpha}\| + \|\boldsymbol{\beta}\|.
\end{aligned}
$$

满足上述三个条件的 $\|\boldsymbol{\alpha}\|$ 称为向量 $\boldsymbol{\alpha}$ 的范数。

例 3-1：判断一个函数是否为范数

在 $\mathbb{R}^2$ 上令 $g(\boldsymbol{w}) = |w_1|$，其中 $\boldsymbol{w} = (w_1, w_2)^T$。$g$ 是向量范数吗？

例 3-1：判断一个函数是否为范数

判断 $g(\boldsymbol{w}) = |w_1|$ 是否为 $\mathbb{R}^2$ 上的范数。

解不是。取 $\boldsymbol{w} = (0,1)^T \neq \mathbf{0}$，却有 $g(\boldsymbol{w}) = 0$，不满足范数的正定性。

例 3-2：欧氏范数

在欧氏空间中，

$$\|\boldsymbol{w}\|_2 = |\boldsymbol{w}| = \sqrt{(\boldsymbol{w}, \boldsymbol{w})}$$

是一种向量范数。请验证三个范数条件。

例 3-2：欧氏范数

验证欧氏范数 $\|\boldsymbol{w}\|_2 = \sqrt{(\boldsymbol{w}, \boldsymbol{w})}$ 满足范数条件。

解内积的正定性给出 $\|\boldsymbol{w}\|_2 \geq 0$，且等号仅在 $\boldsymbol{w} = \boldsymbol{0}$ 时成立；内积的线性与共轭对称性给出 $\|k\boldsymbol{w}\|_2 = |k| \, \|\boldsymbol{w}\|_2$。再由 Cauchy–Schwarz 不等式，

$$\begin{aligned} \|\boldsymbol{w} + \boldsymbol{v}\|_2^2 &= \|\boldsymbol{w}\|_2^2 + 2\,\mathrm{Re}\,(\boldsymbol{w}, \boldsymbol{v}) + \|\boldsymbol{v}\|_2^2 \\ &\leq \left(\|\boldsymbol{w}\|_2 + \|\boldsymbol{v}\|_2\right)^2. \end{aligned}$$

两边开方即得三角不等式，故欧氏模是向量范数。

例 3-3：计算向量的模与二范数

设 $\boldsymbol{w} = (1, -2, 2)^T$。求 $|\boldsymbol{w}|$ 与 $\|\boldsymbol{w}\|_2$。

例 3-3：计算向量的模与二范数

对 $\boldsymbol{w} = (1, -2, 2)^T$，计算模与二范数。

解

$$|\boldsymbol{w}| = \|\boldsymbol{w}\|_2 = \sqrt{1^2 + (-2)^2 + 2^2} = 3.$$

例 3-4：$\mathbb{C}^n$ 中的 p-范数

在线性空间 $\mathbb{C}^n$ 中，对 $\boldsymbol{w} = (w_1, \ldots, w_n)^T$，定义

$$\|\boldsymbol{w}\|_p = \left(\sum_{i=1}^{n} |w_i|^p \right)^{1/p}, \qquad 1 \le p < \infty,$$

验证 $\|\boldsymbol{w}\|_p$ 是向量范数。

例 3-4：$\mathbb{C}^n$ 中的 p-范数

证明 $\|\cdot\|_p$ 满足范数的三个条件（$1 \le p < \infty$）。

解正定性和齐次性直接成立。对三角不等式，令 $\boldsymbol{w} = (w_1, \ldots, w_n)^T$、$\boldsymbol{v} = (v_1, \ldots, v_n)^T$。当 $p = 1$ 时，

$$\|\boldsymbol{w} + \boldsymbol{v}\|_1 = \sum_i |w_i + v_i| \le \sum_i (|w_i| + |v_i|) = \|\boldsymbol{w}\|_1 + \|\boldsymbol{v}\|_1 .$$

当 $p > 1$ 时，关键是证明三角不等式。取共轭指数 $q > 1$，满足 $\frac{1}{p} + \frac{1}{q} = 1$。Hölder 不等式给出

$$\sum_i |a_i b_i| \le \left(\sum_i |a_i|^p \right)^{1/p} \left(\sum_i |b_i|^q \right)^{1/q}.$$

因此

$$\begin{aligned}\|\boldsymbol{w}+\boldsymbol{v}\|_p^p &= \sum_i |w_i+v_i|^p \\ &= \sum_i |w_i+v_i|\,|w_i+v_i|^{p-1} \\ &\le \sum_i |w_i|\,|w_i+v_i|^{p-1} + \sum_i |v_i|\,|w_i+v_i|^{p-1} \\ &\le \left(\|\boldsymbol{w}\|_p + \|\boldsymbol{v}\|_p\right)\|\boldsymbol{w}+\boldsymbol{v}\|_p^{p-1}.\end{aligned}$$

若 $\boldsymbol{w}+\boldsymbol{v}\neq\boldsymbol{0}$，约去右侧的 $\|\boldsymbol{w}+\boldsymbol{v}\|_p^{p-1}$，即得

$$\|\boldsymbol{w}+\boldsymbol{v}\|_p \le \|\boldsymbol{w}\|_p + \|\boldsymbol{v}\|_p.$$

当 $\boldsymbol{w}+\boldsymbol{v}=\boldsymbol{0}$ 时结论显然成立。

例 3-5：计算向量三范数

设 $\boldsymbol{w} = (1, -1)^T$。求 $\|\boldsymbol{w}\|_3$。

例 3-5：计算向量三范数

对 $\boldsymbol{w} = (1, -1)^T$，计算三范数。

解

$$\|\boldsymbol{w}\|_3 = \left(|1|^3 + |-1|^3\right)^{1/3} = \sqrt[3]{2}.$$

【定义 **3-4**】三种典型坐标范数

$$\text{1-范数} \quad \|\boldsymbol{w}\|_1 = \sum_i |w_i|\,,$$

$$\text{2-范数} \quad \|\boldsymbol{w}\|_2 = \left(\sum_i |w_i|^2\right)^{1/2},$$

$$\infty\text{-范数} \quad \|\boldsymbol{w}\|_\infty = \lim_{p\to+\infty} \|\boldsymbol{w}\|_p = \max_i |w_i|\,.$$

最后一个等式可由

$$|w_{i_0}| \le \|\boldsymbol{w}\|_p \le n^{1/p}|w_{i_0}|\,, \qquad |w_{i_0}| = \max_i |w_i|\,,$$

令 $p \to +\infty$ 得到。易验证 $\|\boldsymbol{w}\|_\infty$ 也满足范数的三个条件。

例 3-6：计算向量一范数

设 $\boldsymbol{w} = (1, -2, 3)^T$。求 $\|\boldsymbol{w}\|_1$。

例 3-6：计算向量一范数

对 $\boldsymbol{w} = (1, -2, 3)^T$，计算一范数。

解

$$\|\boldsymbol{w}\|_1 = |1| + |-2| + |3| = 6.$$

例 3-7：计算向量无穷范数

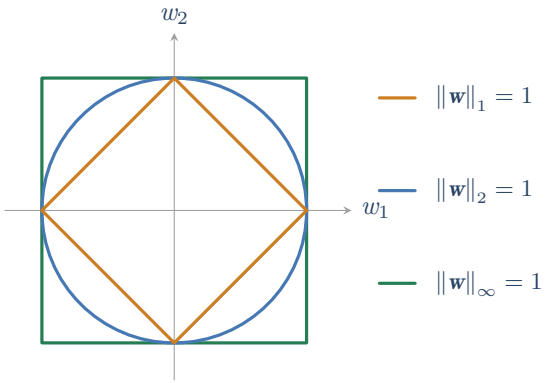
设 $\boldsymbol{w} = (1, -2, 3)^T$。求 $\|\boldsymbol{w}\|_\infty$。

例 3-7：计算向量无穷范数

对 $\boldsymbol{w} = (1, -2, 3)^T$，计算无穷范数。

解

$$\|\boldsymbol{w}\|_\infty = \max\{|1|, |-2|, |3|\} = 3.$$



在 $\mathbb{R}^2$ 中，菱形、圆和正方形对应不同范数；$\|\boldsymbol{w}\|_1 \geq \|\boldsymbol{w}\|_2 \geq \|\boldsymbol{w}\|_\infty$。

例 3-8：多项式空间中的系数范数

设

$$P_2[x] = \{p(x) = w_0 + w_1 x + w_2 x^2 : \ w_0, w_1, w_2 \in \mathbb{R}\},$$

按通常的多项式加法与数乘构成线性空间。取一组基 $\mathcal{B} = (1, x, x^2)$。对 $p(x) = w_0 + w_1 x + w_2 x^2$，定义

$$\|p\|_{\mathcal{B},1} := |w_0| + |w_1| + |w_2|,$$
$$\|p\|_{\mathcal{B},2} := \left(|w_0|^2 + |w_1|^2 + |w_2|^2\right)^{1/2}.$$

请分别验证它们满足范数的正定性、齐次性和三角不等式。

例 3-8：多项式空间中的系数范数

设

$$P_2[x] = \{p(x) = w_0 + w_1 x + w_2 x^2 : \ w_0, w_1, w_2 \in \mathbb{R}\},$$

按通常的多项式加法与数乘构成线性空间。取一组基 $\mathcal{B} = (1, x, x^2)$。对 $p(x) = w_0 + w_1 x + w_2 x^2$，定义

$$\|p\|_{\mathcal{B},1} := |w_0| + |w_1| + |w_2|,$$
$$\|p\|_{\mathcal{B},2} := \left(|w_0|^2 + |w_1|^2 + |w_2|^2\right)^{1/2}.$$

请分别验证它们满足范数的正定性、齐次性和三角不等式。

解：对 $p(x) = w_0 + w_1 x + w_2 x^2$，令 $T_{\mathcal{B}}(p) = \boldsymbol{w} = (w_0, w_1, w_2)^T$。题设的两种函数就是 $\|p\|_{\mathcal{B},r} = \|T_{\mathcal{B}}(p)\|_r$（$r = 1, 2$）。

坐标映射 $T_{\mathcal{B}}: P_2[x] \to \mathbb{R}^3$ 是线性双射，故 $\|p\|_{\mathcal{B},r} \geq 0$，且 $\|p\|_{\mathcal{B},r} = 0$ 当且仅当 $T_{\mathcal{B}}(p) = \mathbf{0}$，即 $p = 0$。对 $r = 1, 2$、$k \in \mathbb{R}$ 及 $p, q \in P_2[x]$，

$$\|kp\|_{\mathcal{B},r} = \|kT_{\mathcal{B}}(p)\|_r = |k| \, \|p\|_{\mathcal{B},r},$$
$$\|p+q\|_{\mathcal{B},r} = \|T_{\mathcal{B}}(p) + T_{\mathcal{B}}(q)\|_r \leq \|p\|_{\mathcal{B},r} + \|q\|_{\mathcal{B},r}.$$

所以它们都是 $P_2[x]$ 上的范数。

例 3-9：换基后的系数范数

在 $P_2[x]$ 中取两组基 $\mathcal{B} = (1, x, x^2)$ 与 $\mathcal{B}' = (1, x, x + x^2)$。按各自基下的系数绝对值之和、平方和的平方根，分别定义系数 1-范数和 2-范数。对 $p(x) = x^2$，求它在两组基下的范数值并比较。

例 3-9：换基后的系数范数

在 $P_2[x]$ 中取两组基 $\mathcal{B}=(1,x,x^2)$ 与 $\mathcal{B}'=(1,x,x+x^2)$。按各自基下的系数绝对值之和、平方和的平方根，分别定义系数 1-范数和 2-范数。对 $p(x)=x^2$，求它在两组基下的范数值并比较。

解：因为 $x^2=(x+x^2)-x$，所以

$$T_{\mathcal{B}}(p)=(0,0,1)^T,\qquad T_{\mathcal{B}'}(p)=(0,-1,1)^T,$$

进而

$$\|p\|_{\mathcal{B},1}=\|p\|_{\mathcal{B},2}=1,\qquad \|p\|_{\mathcal{B}',1}=2,\quad \|p\|_{\mathcal{B}',2}=\sqrt{2}.$$

系数范数依赖所选基；$P_2[x]$ 的系数 1-范数、2-范数须注明基。

【定义 **3-5**】基下坐标诱导的范数

设 V^n 为 $\mathbb{C}$ 上的 n 维线性空间，$\mathcal{B}=(\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_n)$ 是它的一组基。任意 $\boldsymbol{\alpha}\in V^n$ 可唯一表示为

$$\boldsymbol{\alpha}=a_1\odot\boldsymbol{\varepsilon}_1\oplus\cdots\oplus a_n\odot\boldsymbol{\varepsilon}_n,\qquad \mathbf{a}=(a_1,\ldots,a_n)^T\in\mathbb{C}^n.$$

对 $1\le p\le\infty$，定义

$$\|\boldsymbol{\alpha}\|_{\mathcal{B},p}\triangleq\|\mathbf{a}\|_p.$$

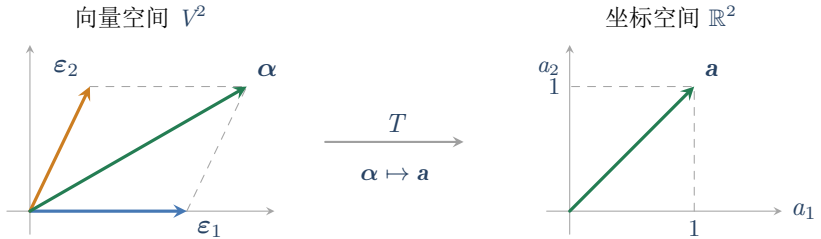
称 $\|\boldsymbol{\alpha}\|_{\mathcal{B},p}$ 为相对于基 $\mathcal{B}$ 的坐标诱导 p-范数。

例 3-10：验证基下坐标诱导的范数

设 V^n 为 $\mathbb{C}$ 上的 n 维线性空间，$\mathcal{B}$ 是它的一组基，$\mathbf{a}$ 是 $\boldsymbol{\alpha}$ 在该基下的坐标向量。对每个 $1 \leq p \leq \infty$，逐一验证

$$\|\boldsymbol{\alpha}\|_{\mathcal{B},p} = \|\mathbf{a}\|_p$$

满足范数的正定性、齐次性与三角不等式。



取 $\mathcal{B} = (\varepsilon_1, \varepsilon_2)$。对 $\boldsymbol{\alpha} = \varepsilon_1 \oplus \varepsilon_2$，坐标向量 $\mathbf{a} = (1,1)^T$，因此 $\|\boldsymbol{\alpha}\|_{\mathcal{B},1} = \|\mathbf{a}\|_1 = 2$。

例 3-11：计算基下坐标诱导的范数

在 $V=\{c_0+c_1t:c_0,c_1\in\mathbb{R}\}$ 中取基 $\varepsilon_1=1,\ \varepsilon_2=t$，并用这组基下坐标的一范数定义 $\|\cdot\|_V$。求 $\boldsymbol{\alpha}=2-t$ 的 $\|\boldsymbol{\alpha}\|_V$。

例 3-11：计算基下坐标诱导的范数

基为 $\varepsilon_1 = 1,\ \varepsilon_2 = t$，计算 $\boldsymbol{\alpha} = 2 - t$ 的坐标一范数。

解　因为 $\boldsymbol{\alpha} = 2\varepsilon_1 - \varepsilon_2$，所以其坐标向量为 $\boldsymbol{a} = (2, -1)^T$。因此

$$\|\boldsymbol{\alpha}\|_V = \|\boldsymbol{a}\|_1 = |2| + |-1| = 3.$$

【定义 **3-6**】正定矩阵诱导的范数

设 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 是实对称正定矩阵。对任意 $\boldsymbol{w} \in \mathbb{R}^n$，定义

$$\|\boldsymbol{w}\|_{\boldsymbol{A}} \triangleq \left(\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}\right)^{1/2}$$

为由正定矩阵 $\boldsymbol{A}$ 诱导的向量范数。

例 3-12：验证正定矩阵诱导的范数

设 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 是实对称正定矩阵。证明

$$\|\boldsymbol{w}\|_{\boldsymbol{A}} = \left(\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}\right)^{1/2}$$

满足正定性、齐次性和三角不等式，因而是 $\mathbb{R}^n$ 上的向量范数。

例 3-12：验证正定矩阵诱导的范数

设 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 是实对称正定矩阵。证明 $\|\boldsymbol{w}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}$ 满足正定性、齐次性和三角不等式。

解：由 $\boldsymbol{A}$ 正定，$\|\boldsymbol{w}\|_{\boldsymbol{A}} \geq 0$，且等号当且仅当 $\boldsymbol{w} = \boldsymbol{0}$。对任意 $k \in \mathbb{R}$，

$$\|k\boldsymbol{w}\|_{\boldsymbol{A}} = \sqrt{k^2\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}} = |k|\,\|\boldsymbol{w}\|_{\boldsymbol{A}}.$$

由于 $\langle \boldsymbol{w}, \boldsymbol{v}\rangle_{\boldsymbol{A}} = \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{v}$ 是内积，Cauchy–Schwarz 不等式给出

$$\begin{aligned}\|\boldsymbol{w}+\boldsymbol{v}\|_{\boldsymbol{A}}^2 &= \|\boldsymbol{w}\|_{\boldsymbol{A}}^2 + 2\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{v} + \|\boldsymbol{v}\|_{\boldsymbol{A}}^2 \\ &\leq \left(\|\boldsymbol{w}\|_{\boldsymbol{A}} + \|\boldsymbol{v}\|_{\boldsymbol{A}}\right)^2.\end{aligned}$$

两边开方得三角不等式，故 $\|\cdot\|_{\boldsymbol{A}}$ 是向量范数。

例 3-13：计算正定矩阵诱导的范数

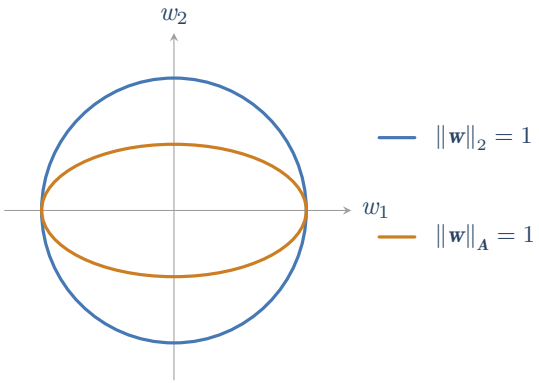
取 $\boldsymbol{A} = \mathrm{diag}(1, 4)$、$\boldsymbol{w} = (1, 1)^T$。求 $\|\boldsymbol{w}\|_{\boldsymbol{A}}$。

例 3-13：计算正定矩阵诱导的范数

对 $\boldsymbol{A} = \mathrm{diag}(1, 4)$、$\boldsymbol{w} = (1, 1)^T$，计算诱导范数。

解

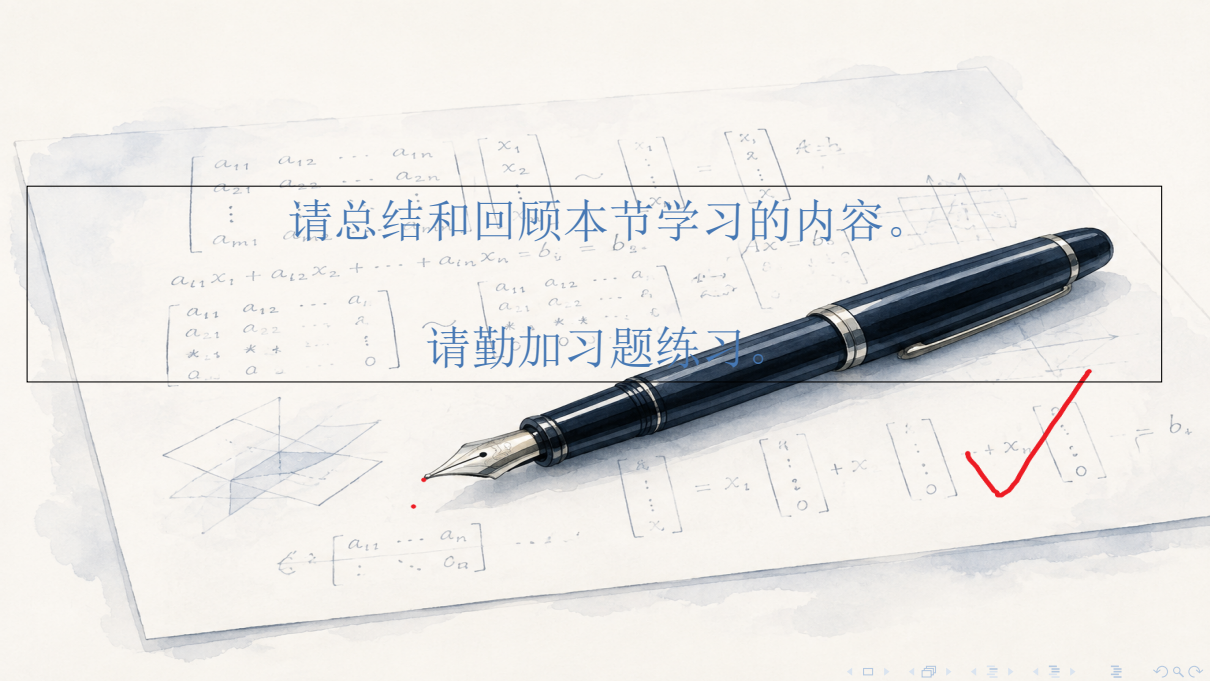
$$\|\boldsymbol{w}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}} = \sqrt{1 \cdot 1^2 + 4 \cdot 1^2} = \sqrt{5}.$$



当 $\boldsymbol{A} = \mathrm{diag}(1, 4)$ 时，$\|\boldsymbol{w}\|_{\boldsymbol{A}}^2 = w_1^2 + 4w_2^2$；橙色单位球沿 w_2 方向收窄。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

上一节要点：向量范数

- 范数是满足正定性、齐次性与三角不等式的实值函数，用来度量向量的“大小”。
- 常用 1、2、∞、p 范数的单位球形状不同，但刻画的是同一个收敛性。
- 取定一组基的坐标、或取一个正定矩阵，都可以诱导出新的范数。

本节：矩阵诱导范数与加权线性回归

- 把回归误差写成范数：从均方误差到加权二次型。
- 权重矩阵的含义：各样本的可信度如何进入目标函数。
- 加权最小二乘的求解与比较，以及 GLS、马氏距离。

沿用线性回归中的记号：$\boldsymbol{X} \in \mathbb{R}^{n\times p}$、$\boldsymbol{y} \in \mathbb{R}^{n}$、$\boldsymbol{w} \in \mathbb{R}^{p}$。原来的均方误差可写为

$$\begin{aligned}\mathcal{J}_{MSE}(\boldsymbol{w}) &= \frac{1}{n}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\right\|_2^2 = \frac{1}{n}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\right\|_{\boldsymbol{I}}^2 \\ &= \frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{I}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}).\end{aligned}$$

现在选取实对称正定矩阵 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$，将残差的欧氏范数换为由 $\boldsymbol{A}$ 诱导的范数：

$$\mathcal{J}_{\boldsymbol{A}}(\boldsymbol{w}) := \frac{1}{n}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\right\|_{\boldsymbol{A}}^2 = \frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}).$$

回归函数 $\boldsymbol{X}\boldsymbol{w}$ 不变，改变的是衡量残差的损失；$\boldsymbol{A}=\boldsymbol{I}$ 时恢复普通 MSE。

若 $\boldsymbol{A} = \mathrm{diag}(a_1, \ldots, a_n)$，其中 $a_i > 0$，则

$$\mathcal{J}_{\boldsymbol{A}}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} a_i \big(y_i - \boldsymbol{x}_i^T \boldsymbol{w}\big)^2.$$

这是加权均方误差：较小的 a_i 降低第 i 个观测对拟合的影响。权重应表达观测可靠性或误差模型，而不应仅因某个点看起来偏离就自动调低。有时分母会取 $\sum_i a_i$；对固定的正权重，这与本讲的 $1/n$ 只差一个正常数，不改变最优参数。本讲保留 $1/n$ 以便与 MSE 对照。

由于 $\boldsymbol{A}$ 对称，展开目标函数可得

$$\begin{aligned}\mathcal{J}_{\boldsymbol{A}}(\boldsymbol{w}) &= \frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}) \\ &= \frac{1}{n}\big(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}-\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}-\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}+\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}\big) \\ &= \frac{1}{n}\big(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}-2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}+\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}\big).\end{aligned}$$

因此，

$$\nabla_{\boldsymbol{w}}\mathcal{J}_{\boldsymbol{A}}(\boldsymbol{w}) = \frac{1}{n}\left(-2\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}+2\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}\right) = \frac{2}{n}\boldsymbol{X}^T\boldsymbol{A}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}).$$

令梯度为零。若 $\boldsymbol{X}$ 列满秩，则 $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$ 可逆，故

$$\boldsymbol{w}_{\boldsymbol{A}}^* = (\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}.$$

这里 $\boldsymbol{A}$ 作用于 $\mathbb{R}^n$ 中的误差，而不是 $\mathbb{R}^p$ 中的参数。

例 3-14：普通最小二乘与加权最小二乘的比较

沿用线性回归章的四个观测点 $(x_i, y_i) = (-1,-1), (0,0), (1,1), (2,6)$，用直线 $y = wx + b$ 拟合。记 $\boldsymbol{w} = (w, b)^T$，并设

$$\boldsymbol{X} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 6 \end{pmatrix}, \qquad \boldsymbol{A} = \operatorname{diag}\left(1, 1, 1, \frac{1}{7}\right).$$

分别求普通 MSE 与 $\boldsymbol{A}$-加权损失的最小点，并比较两条拟合直线。

例 3-14：普通最小二乘与加权最小二乘的比较

用直线 $y = wx + b$ 拟合四个点 $(-1,-1),(0,0),(1,1),(2,6)$。第四点的权重为 $1/7$。分别求普通 MSE 与加权损失的最小点，并比较两条直线。

解：令 $\boldsymbol{w} = (w,b)^T$，则

$$\boldsymbol{X} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}, \qquad \boldsymbol{X}^T = \begin{pmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 6 \end{pmatrix}.$$

普通 MSE 对应 $\boldsymbol{A} = \boldsymbol{I}$。先计算正规方程两侧：

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 2 \\ 2 & 4 \end{pmatrix},$$

$$\boldsymbol{X}^T\boldsymbol{y} = \begin{pmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 1 \\ 6 \end{pmatrix} = \begin{pmatrix} 14 \\ 6 \end{pmatrix}.$$

因此，

$$\boldsymbol{w}_{\boldsymbol{I}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{20} \begin{pmatrix} 4 & -2 \\ -2 & 6 \end{pmatrix} \begin{pmatrix} 14 \\ 6 \end{pmatrix} = \begin{pmatrix} 11/5 \\ 2/5 \end{pmatrix}.$$

对应的拟合直线为 $y = \frac{11}{5}x + \frac{2}{5}$。

加入权重后，第四个观测的权重为 $1/7$。先分别左乘 $\boldsymbol{A}$：

$$\boldsymbol{AX} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1/7 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2/7 & 1/7 \end{pmatrix},$$

$$\boldsymbol{Ay} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1/7 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 1 \\ 6 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 6/7 \end{pmatrix}.$$

再左乘 $\boldsymbol{X}^T$，得到加权正规方程两侧：

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = \begin{pmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2/7 & 1/7 \end{pmatrix} = \frac{1}{7} \begin{pmatrix} 18 & 2 \\ 2 & 22 \end{pmatrix},$$

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} = \begin{pmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 1 \\ 6/7 \end{pmatrix} = \frac{1}{7} \begin{pmatrix} 26 \\ 6 \end{pmatrix}.$$

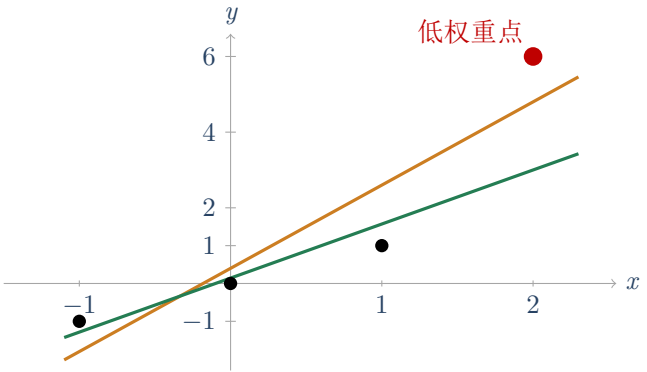
于是

$$\boldsymbol{w}_{\boldsymbol{A}}^{*}=(\boldsymbol{X}^{T}\boldsymbol{A}\boldsymbol{X})^{-1}\boldsymbol{X}^{T}\boldsymbol{A}\boldsymbol{y}=\frac{1}{392}\begin{pmatrix}22 & -2\\ -2 & 18\end{pmatrix}\begin{pmatrix}26\\ 6\end{pmatrix}=\begin{pmatrix}10/7\\ 1/7\end{pmatrix}.$$

因此，加权拟合直线为

$$y=\frac{10}{7}x+\frac{1}{7}.$$

它更接近前三个点所在的 $y=x$。



普通 MSE：$y = \frac{11}{5}x + \frac{2}{5}$　　加权损失：$y = \frac{10}{7}x + \frac{1}{7}$

权重体现对观测可靠性的建模假设；离群不等于错误，也不自动意味着应降权。

Prompt：

【场景】比较普通 MSE 与矩阵诱导的加权损失。四个样本点为 $(-1,-1),(0,0),(1,1),(2,6)$，模型为 $y=wx+b$。令 $\boldsymbol{w}=(w,b)^T$，$\boldsymbol{X}$ 的第 i 行为 $(x_i,1)$，并考虑

$$\mathcal{J}_{\boldsymbol{A}}(\boldsymbol{w})=\frac{1}{4}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}).$$

【问题】我们已取 $\boldsymbol{A}=\mathrm{diag}(1,1,1,1/7)$获得结果，现在考虑 $\boldsymbol{A}$ 为非对角的实对称正定矩阵。请问，这能带来哪些改变？包括（并不局限于）以下角度：

- Q1：如何理解 $\boldsymbol{A}$ 中非零的非对角元素的作用？
- Q2：针对当前的数据，如何设计一个有依据的 $\boldsymbol{A}$，回归函数会怎样变化？
- Q3：哪些数据情境可能因此受益？又可能有什么代价或风险？

请从二次型展开出发解释非对角项，不要预设它一定改善拟合；给出具体矩阵时，先验证其对称正定性，并说明结论依赖的建模假设。对无法由计算支持的判断，明确标为猜想。

本例每行输入为 $(x_i, 1)$，第二列的 1 表示截距 b；残差定义为 $r_i = y_i - (wx_i + b)$，即“观测值减预测值”。令 $\boldsymbol{w} = (w, b)^T$，于是

$$\boldsymbol{X} = \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{pmatrix}, \qquad \boldsymbol{y} = (-1, 0, 1, 6)^T, \qquad \boldsymbol{r}^T = (w - b - 1, -b, 1 - w - b, 6 - 2w - b).$$

对称矩阵 $\boldsymbol{A}$ 下，二次型展开为

$$\mathcal{J}_{\boldsymbol{A}}(\boldsymbol{w}) = \frac{1}{4}\boldsymbol{r}^T\boldsymbol{A}\boldsymbol{r} = \frac{1}{4}\sum_{i=1}^{4} a_{ii}r_i^2 + \frac{1}{2}\sum_{1 \le i < j \le 4} a_{ij}r_i r_j.$$

对角 $\boldsymbol{A}$：逐点加权

$a_{ii}r_i^2$ 控制第 i 点的代价；如 $a_{44} = 1/7$，第四点权重就是其他点的 $1/7$。

非对角 $\boldsymbol{A}$：成对考察

交叉项 $\frac{1}{2}a_{ij}r_i r_j$ 同时含有 r_i, r_j。

方差描述测量误差通常有多大：σ_i^2 越大，观测值的不确定性通常越大。若各观测误差相互独立，协方差矩阵没有非对角项：

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2).$$

取 $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$，便得到

$$\boldsymbol{A} = \operatorname{diag}\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2}\right), \qquad a_{ii} = \frac{1}{\sigma_i^2}.$$

也就是说，权重与误差方差成反比：方差越大，残差平方在损失中的系数越小。

回到本例：为什么第四点的权重是 $1/7$?

假设前三个观测的误差方差均为 1，第四个为 7：

$$\boldsymbol{\Sigma} = \operatorname{diag}(1, 1, 1, 7) \quad \Longrightarrow \quad \boldsymbol{A} = \boldsymbol{\Sigma}^{-1} = \operatorname{diag}(1, 1, 1, 1/7).$$

因此 $a_{44} = 1/7$。这是“第四个观测误差方差较大”这一假设的结果；不是因为它看起来离群，就自动应该降权。

看第 i,j 个残差对应的矩阵小块。由于 $\boldsymbol{A}$ 对称，两个位置上的系数相同：

$$\begin{pmatrix} r_i & r_j \end{pmatrix} \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix} \begin{pmatrix} r_i \\ r_j \end{pmatrix} = a_{ii} r_i^2 + 2a_{ij} r_i r_j + a_{jj} r_j^2.$$

交叉项出现两次，分别来自矩阵的第 (i,j) 与第 (j,i) 项。因此在本例损失中，这一对坐标的交叉项是 $\frac{1}{2} a_{ij} r_i r_j$。

	$r_i r_j > 0$（两个残差同号）	$r_i r_j < 0$（两个残差异号）
$a_{ij} > 0$	交叉项为正，抬高损失	交叉项为负，降低损失
$a_{ij} < 0$	交叉项为负，降低损失	交叉项为正，抬高损失

表格只说明这一项对损失的局部贡献；它不代表总损失一定升高或降低。要求 $\boldsymbol{A}$ 对称正定，才能保证完整二次型对非零残差为正。

设线性模型为 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{w}+\boldsymbol{\varepsilon}$，且 $\boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma}$ 正定。协方差矩阵元素 $\Sigma_{ij}=\mathrm{Cov}(\varepsilon_i,\varepsilon_j)$：对角线是误差方差，非对角线表示误差间的共同变化。
记残差 $\boldsymbol{r}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}$。正态分布的密度关于 $\boldsymbol{w}$ 的部分为

$$p(\boldsymbol{y}\mid\boldsymbol{w})\propto\exp\left(-\frac{1}{2}\boldsymbol{r}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{r}\right).$$

取负对数后，要最小化 $\frac{1}{2}\boldsymbol{r}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{r}+C=2\mathcal{J}_{\boldsymbol{\Sigma}^{-1}}(\boldsymbol{w})+C$，其中 C 与 $\boldsymbol{w}$ 无关。乘以正数 2 或加上常数都不改变最优参数，因此最大似然与此处的加权最小二乘有相同解。

广义最小二乘（GLS）

令 $\boldsymbol{A}=\boldsymbol{\Sigma}^{-1}$。若 $\boldsymbol{X}$ 满列秩，GLS 解为

$$\widehat{\boldsymbol{w}}=(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

因为 $\boldsymbol{\Sigma}^{-1}$ 正定且 $\boldsymbol{X}$ 满列秩，$\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ 正定，故可逆。权重矩阵由误差模型推出。

因为 $\boldsymbol{\Sigma}$ 对称正定，其逆矩阵 $\boldsymbol{\Sigma}^{-1}$ 也对称正定。具体地，对任意非零向量 $\boldsymbol{z}$，令 $\boldsymbol{u} = \boldsymbol{\Sigma}^{-1}\boldsymbol{z}$，则

$$\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z} = \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} > 0.$$

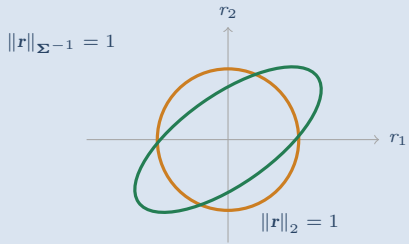
所以可应用本章【定义 3-5】：

$$\|\boldsymbol{r}\|_{\boldsymbol{\Sigma}^{-1}} = \sqrt{\boldsymbol{r}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{r}}.$$

它衡量残差 $\boldsymbol{r}$ 到零向量的加权距离；两个向量 $\boldsymbol{r}, \boldsymbol{s}$ 之间的对应距离，则是 $\|\boldsymbol{r} - \boldsymbol{s}\|_{\boldsymbol{\Sigma}^{-1}}$，称为马氏距离（Mahalanobis distance）。

范数与误差假设相连

GLS 损失为 $\frac{1}{4}\|\boldsymbol{r}\|^2_{\boldsymbol{\Sigma}^{-1}}$。协方差矩阵描述误差的尺度与相关结构；它的逆矩阵决定如何衡量残差。因此，统计模型也确定了这里使用的加权范数。



高方差方向上的单位球半径较长。

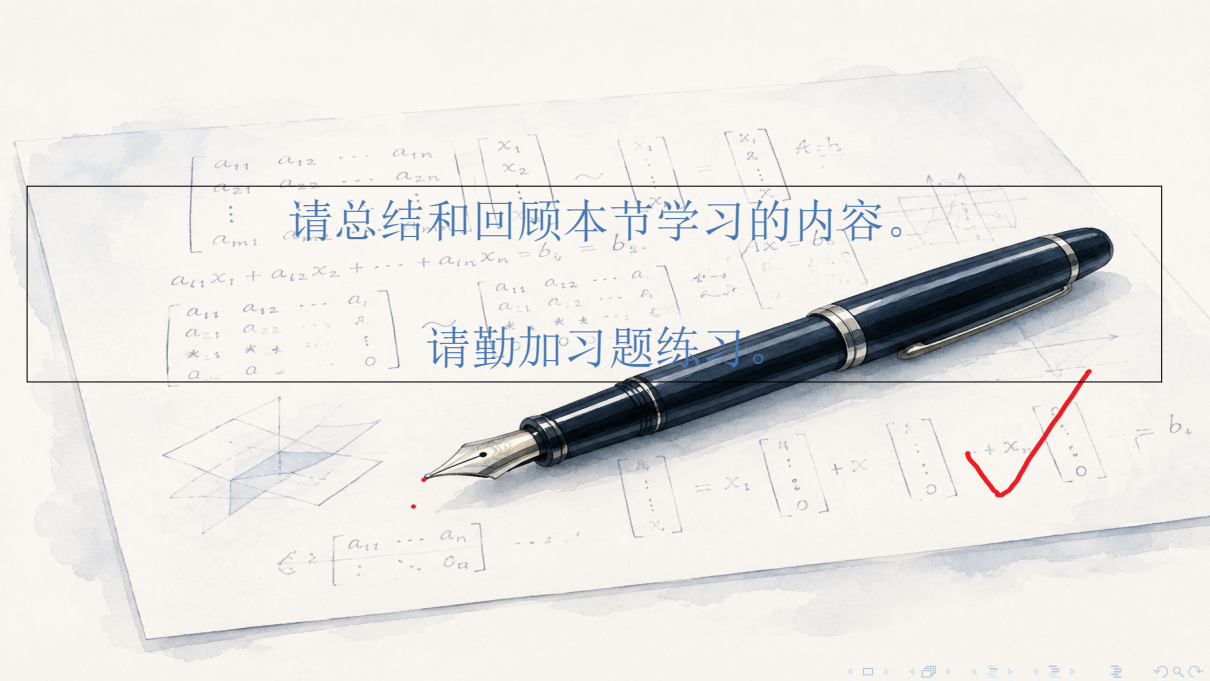
同一组四个数据点，采用不同的误差假设，就会得到不同的权重矩阵和拟合结果：

误差假设	$\boldsymbol{A}$ 的设定	拟合结果
独立且方差相同	$\boldsymbol{I}$	$y = \frac{11}{5}x + \frac{2}{5}$
独立，第四点方差为其他点的 7 倍	$\mathrm{diag}(1, 1, 1, 1/7)$	$y = \frac{10}{7}x + \frac{1}{7}$
误差相关	$\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$	指定完整的 $\boldsymbol{\Sigma}$ 后才能计算

读表时按这个顺序想：先提出误差假设，再写出协方差矩阵 $\boldsymbol{\Sigma}$，最后取逆得到权重 $\boldsymbol{A}$。仅知道“误差相关”还不够；必须知道各误差之间相关的具体程度，才能确定拟合直线。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

上一节要点：矩阵诱导范数与加权线性回归

- 加权损失 $\left\|\boldsymbol{W}^{1/2}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y})\right\|_2^2$：权重越大，该样本对拟合的影响越大。
- 加权最小二乘的梯度与正规方程，及其与普通最小二乘的系数差异。
- 非对角权重对应广义最小二乘与马氏距离。

本节：向量范数的等价性（略讲）

- 换一种范数，收敛性与极值会不会改变？
- 等价常数的估计，以及常用范数之间的显式比较。
- 有限维空间中“任意两个范数等价”。

【定义 **3-7**】等价范数

设 $\|\cdot\|_\alpha$ 与 $\|\cdot\|_\beta$ 是线性空间 V^n 上的两个向量范数。若存在正常数 c_1, c_2，使得

$$c_1 \|\boldsymbol{w}\|_\alpha \leq \|\boldsymbol{w}\|_\beta \leq c_2 \|\boldsymbol{w}\|_\alpha, \qquad \forall \boldsymbol{w} \in V^n,$$

则称 $\|\cdot\|_\alpha$ 与 $\|\cdot\|_\beta$ 等价。

例 3-15：给出一组范数等价常数

对任意 $\boldsymbol{w} \in \mathbb{R}^2$，给出正常数 c_1, c_2，使

$$c_1 \|\boldsymbol{w}\|_1 \leq \|\boldsymbol{w}\|_\infty \leq c_2 \|\boldsymbol{w}\|_1 .$$

例 3-15：给出一组范数等价常数

给出 $\mathbb{R}^2$ 上一范数与无穷范数的等价常数。

解任意两个坐标的绝对值均不超过其最大值，故

$$\|\boldsymbol{w}\|_\infty \leq \|\boldsymbol{w}\|_1 \leq 2\|\boldsymbol{w}\|_\infty .$$

因此可取 $c_1 = \frac{1}{2}$、$c_2 = 1$。

命题 3-2：范数等价是等价关系

- 自反性：$1 \cdot \|\boldsymbol{w}\|_\alpha \leq \|\boldsymbol{w}\|_\alpha \leq 1 \cdot \|\boldsymbol{w}\|_\alpha$。
- 对称性：$\frac{1}{c_2} \|\boldsymbol{w}\|_\beta \leq \|\boldsymbol{w}\|_\alpha \leq \frac{1}{c_1} \|\boldsymbol{w}\|_\beta$。
- 传递性：若 $\|\cdot\|_\alpha \sim \|\cdot\|_\beta$，且 $c_1' \|\boldsymbol{w}\|_\beta \leq \|\boldsymbol{w}\|_\gamma \leq c_2' \|\boldsymbol{w}\|_\beta$，则 $c_1 c_1' \|\boldsymbol{w}\|_\alpha \leq \|\boldsymbol{w}\|_\gamma \leq c_2 c_2' \|\boldsymbol{w}\|_\alpha$。

例 3-16：$\mathbb{C}^n$ 中常用范数的比较

对任意 $\boldsymbol{w} = (w_1, \ldots, w_n)^T$，有

$$\|\boldsymbol{w}\|_\infty \le \|\boldsymbol{w}\|_1 \le n\|\boldsymbol{w}\|_\infty, \qquad \|\boldsymbol{w}\|_\infty \le \|\boldsymbol{w}\|_2 \le \sqrt{n}\|\boldsymbol{w}\|_\infty,$$

以及

$$\|\boldsymbol{w}\|_2 \le \|\boldsymbol{w}\|_1 \le \sqrt{n}\|\boldsymbol{w}\|_2.$$

请证明上述三组不等式。

例 3-16：$\mathbb{C}^n$ 中常用范数的比较

证明 1-范数、2-范数和 ∞-范数之间的三组比较不等式。

解记 $M = \max_i |w_i|$。逐项有 $|w_i| \le M$，而取到最大值的某一项等于 M，故

$$M \le \sum_i |w_i| \le nM.$$

同理，$M^2 \le \sum_i |w_i|^2 \le nM^2$，开方得 $\|\boldsymbol{w}\|_\infty \le \|\boldsymbol{w}\|_2 \le \sqrt{n}\,\|\boldsymbol{w}\|_\infty$。
最后，$\sum_i |w_i|^2 \le (\sum_i |w_i|)^2$，且由 Cauchy–Schwarz 不等式，

$$\sum_i |w_i| \le \sqrt{\sum_i 1^2}\sqrt{\sum_i |w_i|^2} = \sqrt{n}\,\|\boldsymbol{w}\|_2 .$$

于是第三组不等式成立。

定理 3-3：线性空间 V^n 中任意两个范数等价

有限维线性空间中的任意两个向量范数等价。

证明（用到有限维闭有界集的紧致性与极值定理）只需证明任意向量范数 $\|\cdot\|_\star$ 与基下坐标 2-范数 $\|\cdot\|_{\mathcal{B},2}$ 等价。设 $\mathcal{B}=(\boldsymbol{\varepsilon}_1,\dots,\boldsymbol{\varepsilon}_n)$ 为一组基，且

$$\boldsymbol{\alpha}=a_1\boldsymbol{\varepsilon}_1+\cdots+a_n\boldsymbol{\varepsilon}_n,\qquad \mathbf{a}=(a_1,\dots,a_n)^T.$$

记 $\|\boldsymbol{\alpha}\|_{\mathcal{B},2}:=\|\mathbf{a}\|_2$，并令

$$S=\{\mathbf{a}\in\mathbb{C}^n:\|\mathbf{a}\|_2=1\},$$

并定义

$$f(\mathbf{a})=\|a_1\boldsymbol{\varepsilon}_1+\cdots+a_n\boldsymbol{\varepsilon}_n\|_\star.$$

由反三角不等式及基向量的有限性，

$$|f(\mathbf{a})-f(\mathbf{b})|\le\sum_{i=1}^{n}|a_i-b_i|\,\|\boldsymbol{\varepsilon}_i\|_\star\quad(\mathbf{a},\mathbf{b}\in\mathbb{C}^n),$$

证明（续）由上式可知 f 连续。集合 S 是闭有界集，f 在 S 上连续，故可取得最小值 c_1 和最大值 c_2。因为 S 中的向量非零，所以 $c_1 > 0$。对任意非零 $\boldsymbol{\alpha}$，写成

$$\boldsymbol{\beta} = \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_{\mathcal{B},2}} = \sum_{i=1}^{n} b_i \boldsymbol{\varepsilon}_i, \qquad b_i = \frac{a_i}{\|\boldsymbol{\alpha}\|_{\mathcal{B},2}}.$$

则 $\boldsymbol{b} = (b_1, \ldots, b_n)^T \in S$，从而

$$c_1 \leq f(\boldsymbol{b}) \leq c_2.$$

乘以 $\|\boldsymbol{\alpha}\|_{\mathcal{B},2}$ 得

$$c_1 \|\boldsymbol{\alpha}\|_{\mathcal{B},2} \leq \|\boldsymbol{\alpha}\|_{\star} \leq c_2 \|\boldsymbol{\alpha}\|_{\mathcal{B},2}.$$

当 $\boldsymbol{\alpha} = \mathbf{0}$ 时上式显然成立。因此，有限维线性空间中的任意两个范数均等价。

定理 3-4：$\mathbb{C}^n$ 中的收敛性与范数

对任意向量范数 $\|\cdot\|$，有

$$\lim_{k\to\infty} \boldsymbol{w}^{(k)} = \boldsymbol{w} \quad \Longleftrightarrow \quad \lim_{k\to\infty} \left\| \boldsymbol{w}^{(k)} - \boldsymbol{w} \right\| = 0.$$

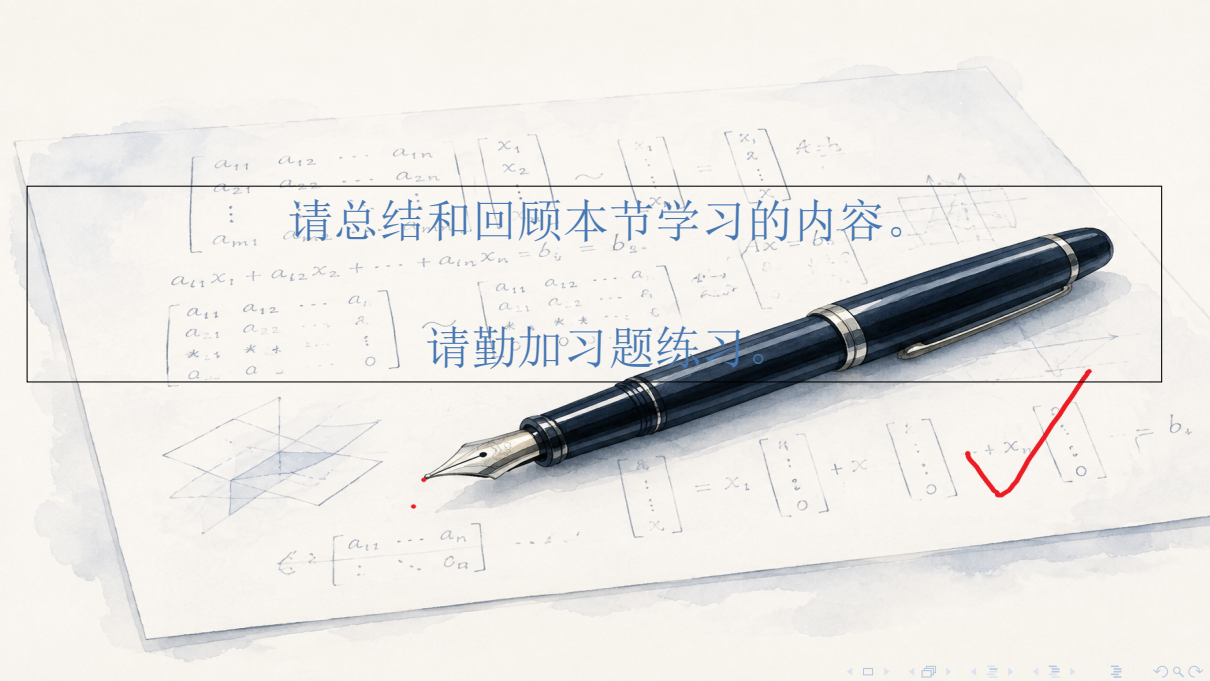
证明只需选取 1-范数。由坐标收敛的定义，

$$\begin{aligned}
\boldsymbol{w}^{(k)} \to \boldsymbol{w} &\Longleftrightarrow w_i^{(k)} \to w_i \quad (i = 1, 2, \ldots, n) \\
&\Longleftrightarrow \left| w_i^{(k)} - w_i \right| \to 0 \quad (i = 1, 2, \ldots, n) \\
&\Longleftrightarrow \sum_{i=1}^{n} \left| w_i^{(k)} - w_i \right| \to 0 \\
&\Longleftrightarrow \left\| \boldsymbol{w}^{(k)} - \boldsymbol{w} \right\|_1 \to 0.
\end{aligned}$$

又因为有限维空间中的任意两个范数等价，结论对任意向量范数都成立。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

上一节要点：向量范数的等价性

- 任意两个向量范数等价：存在与向量无关的常数 c_1, c_2，使 $c_1 \|\cdot\|_\alpha \le \|\cdot\|_\beta \le c_2 \|\cdot\|_\alpha$。
- 常用范数之间有显式的比较常数。
- 向量序列的收敛性与所选范数无关。

本节：矩阵范数

- 矩阵是线性变换在某组基下的坐标表。
- 矩阵范数在范数公理之外，还要满足次可乘性 $\|\boldsymbol{AB}\| \le \|\boldsymbol{A}\| \, \|\boldsymbol{B}\|$。
- Frobenius 范数，以及矩阵范数与向量范数的相容性。

谈矩阵的范数之前，先回到线性变换的图景。选定一组基后，每个向量 $\boldsymbol{w}$ 有一列坐标；线性变换 $\mathcal{A}$ 作用在 $\boldsymbol{w}$ 上，其像的坐标正是用矩阵 $\boldsymbol{A}$ 去乘：

$$\boldsymbol{w} \xrightarrow{\mathcal{A}} \mathcal{A}(\boldsymbol{w}) \qquad \Longleftrightarrow \qquad \text{坐标 } \boldsymbol{w} \xrightarrow{\times \boldsymbol{A}} \boldsymbol{A}\boldsymbol{w}.$$

矩阵 $\boldsymbol{A}$ 不是一堆孤立的数，而是变换 $\mathcal{A}$ 在这组基下的坐标表（默认标准基）。

于是" 矩阵范数" 要量的是变换

既然变换靠 $\boldsymbol{A}$ 实现，" 这个变换能把向量放大多少" 就成了自然的问题——矩阵范数正是为回答它而定义的。这也解释了为何我们额外要求 $\|\boldsymbol{A}\boldsymbol{B}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$：两个变换复合的放大率，不该超过各自放大率之积。

【定义 **3-8**】矩阵范数

在集合 $\mathbb{C}^{m\times n}$ 中，对任意 $\boldsymbol{A}\in\mathbb{C}^{m\times n}$ 定义实数 $\|\boldsymbol{A}\|$，并满足

$$
\begin{aligned}
&(1)\quad \|\boldsymbol{A}\|\geq 0, \qquad \|\boldsymbol{A}\|=0\Longleftrightarrow \boldsymbol{A}=\boldsymbol{O}_{m\times n};\\
&(2)\quad \|k\boldsymbol{A}\|=|k|\,\|\boldsymbol{A}\|, \qquad \forall k\in\mathbb{C};\\
&(3)\quad \|\boldsymbol{A}+\boldsymbol{B}\|\leq\|\boldsymbol{A}\|+\|\boldsymbol{B}\|;\\
&(4)\quad \|\boldsymbol{A}\boldsymbol{B}\|\leq\|\boldsymbol{A}\|\,\|\boldsymbol{B}\|, \qquad \forall \boldsymbol{B}\in\mathbb{C}^{n\times l}.
\end{aligned}
$$

满足前四条性质的 $\|\boldsymbol{A}\|$ 称为矩阵 $\boldsymbol{A}$ 的范数。

本讲采用包含次乘性条件 (4) 的矩阵范数约定；有些教材将 (1)–(3) 称为矩阵范数，再把 (4) 单独称为次乘性。阅读时应留意约定差异。

例 3-17：判断一个矩阵函数是否为范数

对二阶方阵令 $N(\boldsymbol{A}) = \max_{i,j} |a_{ij}|$。它满足矩阵范数的乘法不等式吗？

例 3-17：判断一个矩阵函数是否为范数

检查二阶方阵的最大元素绝对值是否满足乘法不等式。

解不满足。取 $\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$，则 $N(\boldsymbol{A}) = 1$，而

$$\boldsymbol{A}^2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}, \qquad N(\boldsymbol{A}^2) = 2 > 1 = N(\boldsymbol{A})^2.$$

所以 N 不是本讲定义下的矩阵范数。

例 3-18：Frobenius 范数是矩阵范数

设 $\boldsymbol{A} = (a_{ij}) \in \mathbb{C}^{m\times n}$，定义

$$\|\boldsymbol{A}\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2},$$

称为 $\boldsymbol{A}$ 的 **Frobenius** 范数（弗罗贝尼乌斯范数）。验证它是矩阵范数。

例 3-18：Frobenius 范数是矩阵范数

验证 $\|\boldsymbol{A}\|_F = \left(\sum_{i,j}|a_{ij}|^2\right)^{1/2}$ 满足矩阵范数四条件。

解把 $\boldsymbol{A}$ 拉直为向量 $\mathrm{vec}(\boldsymbol{A})$，则 $\|\boldsymbol{A}\|_F = \|\mathrm{vec}(\boldsymbol{A})\|_2$，故正定性、齐次性、三角不等式由向量 2-范数直接得到。只需验证次可乘性：

$$\begin{aligned}\|\boldsymbol{AB}\|_F^2 &= \sum_{i,j}\left|\sum_k a_{ik}b_{kj}\right|^2 \le \sum_{i,j}\left(\sum_k |a_{ik}|^2\right)\left(\sum_k |b_{kj}|^2\right)\\ &= \left(\sum_{i,k}|a_{ik}|^2\right)\left(\sum_{k,j}|b_{kj}|^2\right) = \|\boldsymbol{A}\|_F^2\,\|\boldsymbol{B}\|_F^2,\end{aligned}$$

其中不等号用了 Cauchy–Schwarz 不等式。开方即得 $\|\boldsymbol{AB}\|_F \le \|\boldsymbol{A}\|_F\,\|\boldsymbol{B}\|_F$。故 Frobenius 范数是矩阵范数。

【定义 **3-9**】相容的矩阵范数与向量范数

设 $\|\boldsymbol{A}\|_M$ 是 $\mathbb{C}^{m\times n}$ 上的矩阵范数，$\|\cdot\|_{V_n}$ 和 $\|\cdot\|_{V_m}$ 分别是 $\mathbb{C}^n$ 与 $\mathbb{C}^m$ 上的向量范数。若

$$\|\boldsymbol{A}\boldsymbol{w}\|_{V_m} \leq \|\boldsymbol{A}\|_M \|\boldsymbol{w}\|_{V_n}, \qquad \forall \boldsymbol{A} \in \mathbb{C}^{m\times n},\ \forall \boldsymbol{w} \in \mathbb{C}^n,$$

则称矩阵范数 $\|\cdot\|_M$ 与这组向量范数相容。

相容性给出矩阵对输入向量的放大倍数上界：$\|\boldsymbol{A}\boldsymbol{w}\|_{V_m} \le \|\boldsymbol{A}\|_M \, \|\boldsymbol{w}\|_{V_n}$.

- 误差传播：输入含误差 $\delta\boldsymbol{w}$，则输出误差满足

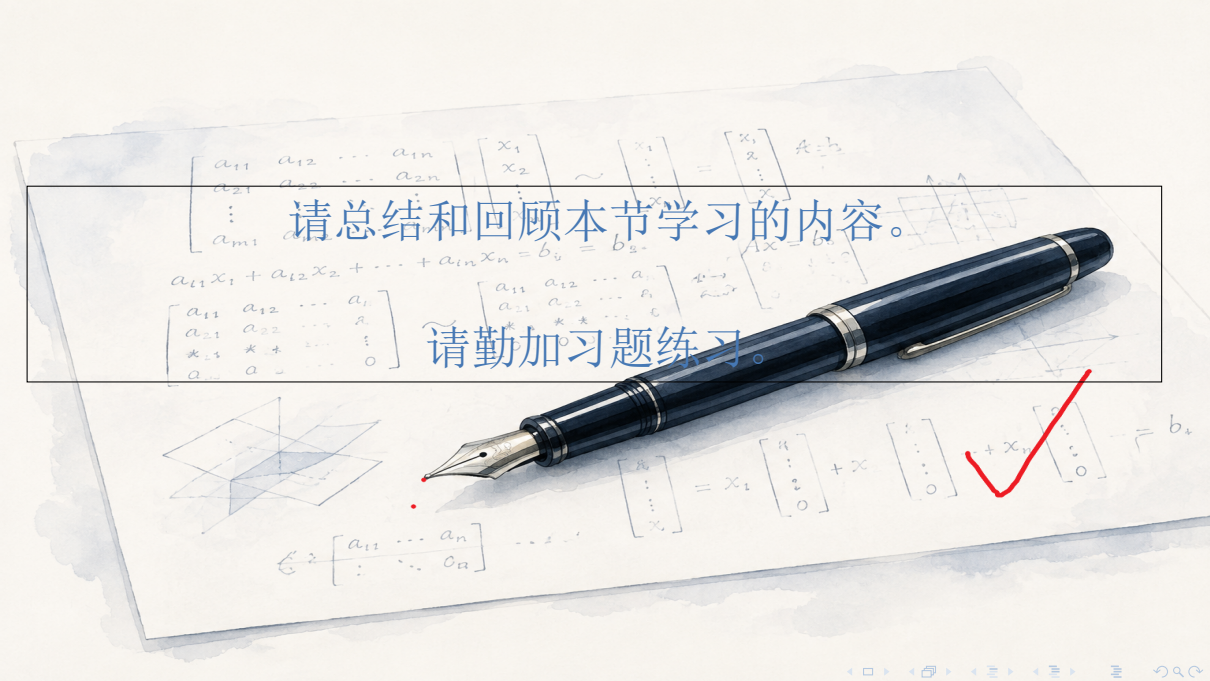
$$\|\boldsymbol{A}(\boldsymbol{w}+\delta\boldsymbol{w}) - \boldsymbol{A}\boldsymbol{w}\|_{V_m} \le \|\boldsymbol{A}\|_M \, \|\delta\boldsymbol{w}\|_{V_n}.$$

- 连续变换：各步矩阵都与同一向量范数相容，则 $\|\boldsymbol{A}_k \cdots \boldsymbol{A}_1 \boldsymbol{w}\|_V \le \left(\prod_{j=1}^{k} \|\boldsymbol{A}_j\|_M\right) \|\boldsymbol{w}\|_V$；可控制迭代误差累积、分析收敛性。
- 线性回归：参数扰动引起的预测变化满足 $\|\boldsymbol{X}(\boldsymbol{w}+\delta\boldsymbol{w}) - \boldsymbol{X}\boldsymbol{w}\|_2 \le \|\boldsymbol{X}\|_F \, \|\delta\boldsymbol{w}\|_2$.

Frobenius 范数的上界便于计算，但不一定最紧；下面引入诱导范数给出最紧放大率。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

上一节要点：矩阵范数

- 矩阵范数在范数公理之外还要满足次可乘性，它来自“复合映射”的放大。
- Frobenius 范数 $\|\boldsymbol{A}\|_F = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}$ 是矩阵范数。
- 相容性 $\|\boldsymbol{A}\boldsymbol{w}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{w}\|$ 控制放大与误差，但上界不一定最紧。

本节：矩阵的诱导范数

- 寻找最紧的放大倍数——诱导范数的定义。
- 诱导范数是最小的相容范数；列和、行和与谱范数三种算法。
- 谱范数的几何：单位球被映成椭球。

上一节末尾的相容性给出放大上界 $\|\boldsymbol{A}\boldsymbol{w}\|_{V_m} \le \|\boldsymbol{A}\|_M \|\boldsymbol{w}\|_{V_n}$，但 $\|\boldsymbol{A}\|_{m_1}, \|\boldsymbol{A}\|_F$ 等给的上界不一定最紧。回到第 3.4 节开头的图景：矩阵 $\boldsymbol{A}$ 是线性变换 $\mathcal{A}$ 的坐标表。于是最自然的问题是——

> 变换 $\mathcal{A}$ 究竟最多能把向量放大多少倍？
> 即求 $\dfrac{\|\mathcal{A}(\boldsymbol{w})\|}{\|\boldsymbol{w}\|}$ 在所有非零 $\boldsymbol{w}$ 上的最大值。

这个"最紧的放大倍数"本身就刻画了矩阵的大小，而且它自动与所用向量范数相容。这样得到的矩阵范数，称为诱导范数（或算子范数、从属范数）。

【定义 **3-10**】矩阵的诱导范数（算子范数）

设 $\|\cdot\|_V$ 是 $\mathbb{C}^n$ 上的向量范数。矩阵 $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 的 诱导范数定义为最大放大倍数：

$$\|\boldsymbol{A}\|_{V,\mathrm{ind}} = \max_{\boldsymbol{w}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{w}\|_V}{\|\boldsymbol{w}\|_V} = \max_{\|\boldsymbol{w}\|_V=1} \|\boldsymbol{A}\boldsymbol{w}\|_V .$$

（后一等号用到齐次性：把 $\boldsymbol{w}$ 单位化不改变比值。）

它自动满足三条关键性质

(1) 范数：非负、正定、齐次、三角不等式均可由向量范数的对应性质直接推出。**(2)** 相容：由定义，对任意 $\boldsymbol{w}$ 有 $\|\boldsymbol{A}\boldsymbol{w}\|_V \le \|\boldsymbol{A}\|_{V,\mathrm{ind}} \|\boldsymbol{w}\|_V$。**(3)** 次可乘：$\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{w}\|_V \le \|\boldsymbol{A}\|_{V,\mathrm{ind}} \|\boldsymbol{B}\boldsymbol{w}\|_V \le \|\boldsymbol{A}\|_{V,\mathrm{ind}} \|\boldsymbol{B}\|_{V,\mathrm{ind}} \|\boldsymbol{w}\|_V$。

相容性要求 $\|\boldsymbol{A}\boldsymbol{w}\|_V \le \|\boldsymbol{A}\|_M \|\boldsymbol{w}\|_V$ 对一切 $\boldsymbol{w}$ 成立，即 $\|\boldsymbol{A}\|_M$ 必须不小于最大放大倍数。而诱导范数取的正是这个最大值：

诱导范数是" 最省" 的相容范数

$$\|\boldsymbol{A}\|_{V,\mathrm{ind}} = \min\left\{ C:\ \|\boldsymbol{A}\boldsymbol{w}\|_V \le C\|\boldsymbol{w}\|_V \ \text{对一切}\boldsymbol{w} \right\}.$$

因此，凡与 $\|\cdot\|_V$ 相容的矩阵范数都不小于诱导范数；诱导范数是其中最小的。

这正解释了第 3.4 节前面为何说 Frobenius 等范数的上界" 不一定最紧"——它们相容，但一般严格大于诱导范数。诱导范数给出了放大率的精确刻画。

对 $p = 1, \infty$，最大放大倍数能写成显式公式，便于计算：

命题 3-5：诱导 1-范数（列和范数）与诱导 ∞-范数（行和范数）

$$\|\boldsymbol{A}\|_1 = \max_j \sum_i |a_{ij}| \quad \text{(列元素绝对值之和的最大者，外号列和范数)}$$

$$\|\boldsymbol{A}\|_\infty = \max_i \sum_j |a_{ij}| \quad \text{(行元素绝对值之和的最大者，外号行和范数)}$$

口诀

诱导 1-范数（列和范数）看列，诱导 ∞-范数（行和范数）看行：取各列（行）绝对值之和，再取最大。注意 $\|\boldsymbol{A}\|_1$ 与本讲的 $m_1 = \sum_{i,j} |a_{ij}|$（全体求和）、$\|\boldsymbol{A}\|_\infty$ 与 $m_\infty = n \max |a_{ij}|$（最大元的 n 倍）都不是一回事。

取向量 2-范数，最大放大倍数要用到矩阵的奇异值：

命题 3-6：诱导 2-范数（谱范数）

$$\|\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{w}\|_2=1} \|\boldsymbol{A}\boldsymbol{w}\|_2 = \sigma_{\max}(\boldsymbol{A}),$$

即 $\boldsymbol{A}$ 的最大奇异值（$\boldsymbol{A}^H\boldsymbol{A}$ 最大特征值的平方根）。它又称谱范数。

谱范数与 Frobenius 范数的分工

谱范数取最强奇异值 $\sigma_{\max}$，量" 最大放大率"；Frobenius 范数取全部奇异值 $\sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$，量" 总体能量"。谱范数更紧（它就是诱导范数），Frobenius 更易算——两者各有用处。

例 3-19：计算诱导 1-范数（列和范数）

设 $\boldsymbol{A} = \begin{pmatrix} 1 & -2 \\ -3 & 4 \end{pmatrix}$，求列和范数 $\|\boldsymbol{A}\|_1$。

例 3-19：计算诱导 1-范数（列和范数）

求 $\boldsymbol{A} = \begin{pmatrix} 1 & -2 \\ -3 & 4 \end{pmatrix}$ 的列和范数。

解各列绝对值之和：第 1 列 $|1| + |-3| = 4$，第 2 列 $|-2| + |4| = 6$。取最大者：

$$\|\boldsymbol{A}\|_1 = \max\{4, 6\} = 6.$$

例 3-20：计算诱导 ∞-范数（行和范数）

设 $\boldsymbol{A} = \begin{pmatrix} 1 & -2 \\ -3 & 4 \end{pmatrix}$，求行和范数 $\|\boldsymbol{A}\|_\infty$。

例 3-20：计算诱导 ∞-范数（行和范数）

求 $\boldsymbol{A} = \begin{pmatrix} 1 & -2 \\ -3 & 4 \end{pmatrix}$ 的行和范数。

解各行绝对值之和：第 1 行 $|1| + |-2| = 3$，第 2 行 $|-3| + |4| = 7$。取最大者：

$$\|\boldsymbol{A}\|_\infty = \max\{3, 7\} = 7.$$

同一个矩阵，列和范数得 6、行和范数得 7——两者一般不相等。

例 3-21：计算诱导 2-范数（谱范数）

设 $\boldsymbol{A} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$，求谱范数 $\|\boldsymbol{A}\|_2$。

例 3-21：计算诱导 2-范数（谱范数）

求 $\boldsymbol{A}=\begin{pmatrix}3 & 0\\ 0 & 1\end{pmatrix}$ 的谱范数。

解谱范数是最大奇异值。$\boldsymbol{A}^H\boldsymbol{A}=\mathrm{diag}(9,1)$，特征值为 $9,1$，故奇异值为 $3,1$，最大奇异值

$$\|\boldsymbol{A}\|_2=\sigma_{\max}(\boldsymbol{A})=3.$$

几何上看，$\boldsymbol{A}$ 把单位圆盘拉长为半轴 $3,1$ 的椭圆，最长半轴即谱范数。

例 3-22：计算复矩阵的三种诱导范数

设

$$\boldsymbol{B}=\begin{pmatrix}-1 & 2 & 3\\ 1 & 0 & \mathrm{i}\end{pmatrix}\in\mathbb{C}^{2\times 3},$$

求 $\|\boldsymbol{B}\|_1$、$\|\boldsymbol{B}\|_\infty$ 与 $\|\boldsymbol{B}\|_2$。

例 3-22：计算复矩阵的三种诱导范数

设 $\boldsymbol{B}=\begin{pmatrix}-1 & 2 & 3\\ 1 & 0 & \mathrm{i}\end{pmatrix}$，求三种诱导范数。

解（**1**）列和范数（逐列取模求和，再取最大）

$$\|\boldsymbol{B}\|_1=\max\big\{|-1|+|1|,\ |2|+|0|,\ |3|+|\mathrm{i}|\big\}=\max\{2,2,4\}=4.$$

（**2**）行和范数（逐行取模求和，再取最大）

$$\|\boldsymbol{B}\|_\infty=\max\big\{|-1|+|2|+|3|,\ |1|+|0|+|\mathrm{i}|\big\}=\max\{6,2\}=6.$$

注意复数取模 $|\mathrm{i}|=1$，这是本题与前几个实矩阵例题的主要差别。

例 3-22：计算复矩阵的三种诱导范数

设 $\boldsymbol{B}=\begin{pmatrix}-1 & 2 & 3\\ 1 & 0 & \mathrm{i}\end{pmatrix}$，求三种诱导范数。

解（**3**）谱范数 • 第一步：先算 $\boldsymbol{B}^H\boldsymbol{B}$

$\boldsymbol{B}^H=\begin{pmatrix}-1 & 1\\ 2 & 0\\ 3 & -\mathrm{i}\end{pmatrix}$（共轭转置：$\bar{\mathrm{i}}=-\mathrm{i}$），于是 $\boldsymbol{B}^H\boldsymbol{B}=\begin{pmatrix}2 & -2 & -3+\mathrm{i}\\ -2 & 4 & 6\\ -3-\mathrm{i} & 6 & 10\end{pmatrix}$，其中

- 对角元是各列模长平方：$(\boldsymbol{B}^H\boldsymbol{B})_{11}=|-1|^2+|1|^2=2$，$(\boldsymbol{B}^H\boldsymbol{B})_{22}=|2|^2+|0|^2=4$，$(\boldsymbol{B}^H\boldsymbol{B})_{33}=|3|^2+|\mathrm{i}|^2=10$；
- 非对角元是列向量的内积：$(\boldsymbol{B}^H\boldsymbol{B})_{12}=\overline{(-1)}\cdot 2+\bar{1}\cdot 0=-2$，$(\boldsymbol{B}^H\boldsymbol{B})_{13}=\overline{(-1)}\cdot 3+\bar{1}\cdot\mathrm{i}=-3+\mathrm{i}$，$(\boldsymbol{B}^H\boldsymbol{B})_{23}=\bar{2}\cdot 3+\bar{0}\cdot\mathrm{i}=6$；
- 其余由共轭对称得到（$\boldsymbol{B}^H\boldsymbol{B}$ 是 Hermitian 矩阵）。

例 3-22：计算复矩阵的三种诱导范数

设 $\boldsymbol{B}=\begin{pmatrix}-1 & 2 & 3\\ 1 & 0 & \mathrm{i}\end{pmatrix}$，求三种诱导范数。

解（**3**）续 • 第二步：按第一行展开 $\det(\boldsymbol{B}^H\boldsymbol{B}-\lambda\boldsymbol{I})$ 对 $\boldsymbol{B}^H\boldsymbol{B}=\begin{pmatrix}2 & -2 & -3+\mathrm{i}\\ -2 & 4 & 6\\ -3-\mathrm{i} & 6 & 10\end{pmatrix}$，

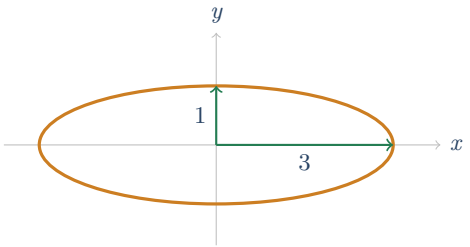
$$\begin{aligned}\det &= (2-\lambda)\big[(4-\lambda)(10-\lambda)-36\big]-(-2)\big[-2(10-\lambda)-6(-3-\mathrm{i})\big]+(-3+\mathrm{i})\big[-2\cdot 6-(4-\lambda)(-3-\mathrm{i})\big].\\ &= -\lambda^3+16\lambda^2-18\lambda=-\lambda\big(\lambda^2-16\lambda+18\big).\end{aligned}$$

令其为零：$\lambda_1=0$，或由二次方程求根公式

$$\lambda_{2,3}=\frac{16\pm\sqrt{16^2-4\cdot 18}}{2}=\frac{16\pm\sqrt{256-72}}{2}=\frac{16\pm\sqrt{184}}{2}=8\pm\sqrt{46}.$$

最大特征值 $\lambda_{\max}=8+\sqrt{46}\approx 14.78$，于是

$$\|\boldsymbol{B}\|_2=\sqrt{\lambda_{\max}(\boldsymbol{B}^H\boldsymbol{B})}=\sqrt{8+\sqrt{46}}\approx\sqrt{14.78}\approx 3.84.$$



单位圆上的点是 $\boldsymbol{w} = (\cos\theta, \sin\theta)^T$。左乘 $\boldsymbol{A} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$ 得 $\boldsymbol{A}\boldsymbol{w} = (3\cos\theta,\ \sin\theta)^T$。设像点 $(x, y) = (3\cos\theta, \sin\theta)$，由 $\cos^2\theta + \sin^2\theta = 1$ 消去 θ：$\left(\frac{x}{3}\right)^2 + y^2 = 1$——半轴 $3, 1$ 的椭圆。

谱范数在其中：$\|\boldsymbol{A}\boldsymbol{w}\|_2$ 是像点到原点的距离，在长半轴端点 $(\pm 3, 0)$ 处最大，故 $\|\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{w}\|_2=1} \|\boldsymbol{A}\boldsymbol{w}\|_2 = 3$。

几何结论

谱范数 = 单位球面被 $\boldsymbol{A}$ 映成的椭球面的最长半轴。

上一页的 $\boldsymbol{A} = \mathrm{diag}(3, 1)$ 只是特例。对任意 $\boldsymbol{A} \in \mathbb{C}^{m\times n}$，取奇异值分解

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H, \qquad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0),$$

其中 $\boldsymbol{U}, \boldsymbol{V}$ 为酉矩阵，$\sigma_1 \geq \cdots \geq \sigma_r > 0$ 为奇异值。
$\boldsymbol{A}$ 作用在单位球面 $\|\boldsymbol{w}\|_2 = 1$ 上，恰是三步：

- $\boldsymbol{V}^H$：酉变换，不改变长度，球面仍为球面；
- $\boldsymbol{\Sigma}$：沿坐标轴分别放大 σ_i 倍，球面变为轴对齐椭球；
- $\boldsymbol{U}$：再作酉变换，椭球整体旋转到最终朝向。

故像集是以 $\sigma_i \boldsymbol{u}_i$ 为半轴的（超）椭球（$\boldsymbol{u}_i$ 为 $\boldsymbol{U}$ 的第 i 列），最长半轴为 $\sigma_{\max} = \sigma_1$，在 $\boldsymbol{w}$ 取 $\boldsymbol{V}$ 的第一列（右奇异向量）时达到。

普遍结论

对任意 $\boldsymbol{A}$，$\|\boldsymbol{A}\|_2 = \sigma_{\max}$ 恒等于单位球面映成的椭球面的最长半轴；对角矩阵只是 $\boldsymbol{U} = \boldsymbol{V} = \boldsymbol{I}$ 的特例。$\boldsymbol{A}$ 奇异时椭球退化（某些半轴为零），最长半轴仍是 σ_1。

严格说，$\{\boldsymbol{A}\boldsymbol{w} : \|\boldsymbol{w}\|_2 \le 1\}$ 是单位球的线性像。由 $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$ 看：酉变换不改变形状，只有 $\boldsymbol{\Sigma}$ 按 σ_i 缩放各方向，故像集恒为椭球，但可能退化：

- $\boldsymbol{A}$ 满秩（$r = \min\{m, n\}$）：r 条正半轴的正规椭球；
- 秩亏 $r < \min\{m, n\}$：$\sigma_{r+1} = \cdots = 0$，这些方向被压成零，像集退化为 r 维椭球（$r = 1$ 为线段、$r = 2$ 为椭圆盘），内部为空；
- $\boldsymbol{A} = \boldsymbol{O}$：退化为单点 $\{\boldsymbol{0}\}$。

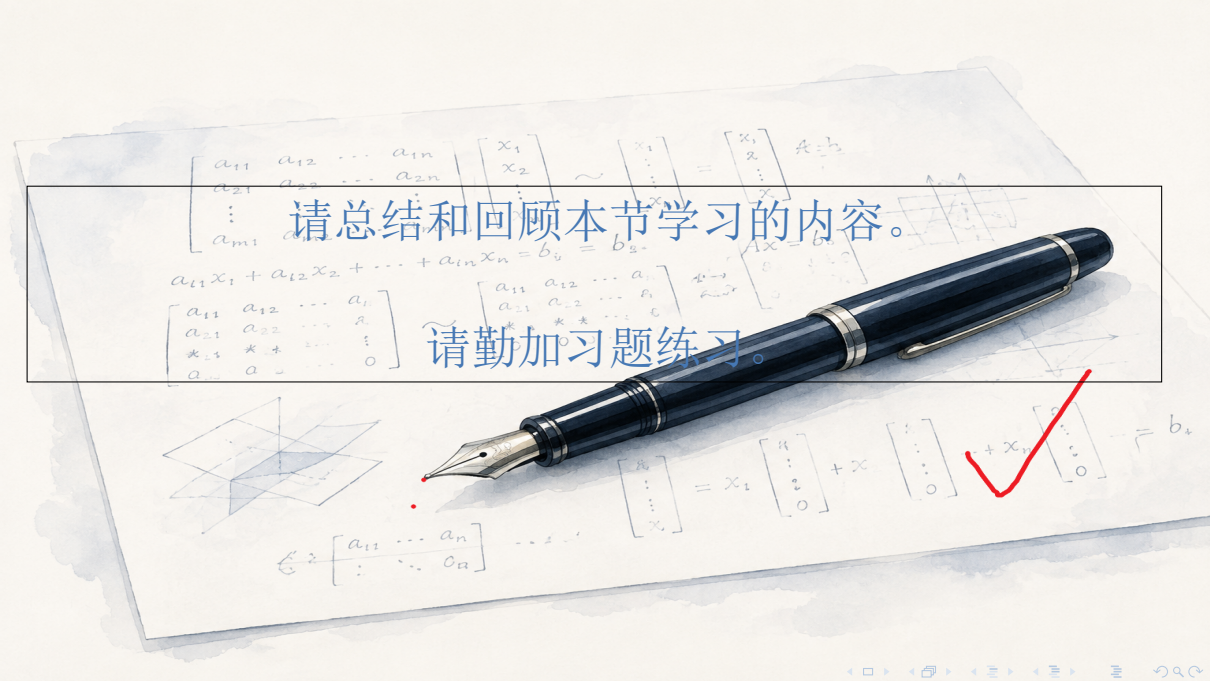
注意：椭球指的是整个单位球的像；单个 $\boldsymbol{w}$ 只给出一个点。

与谱范数的关系

退化不影响结论：$\max_{\|\boldsymbol{w}\|_2=1} \|\boldsymbol{A}\boldsymbol{w}\|_2 = \sigma_{\max}$ 仍然成立，最长（未退化的）半轴就是 $\sigma_{\max}$，在对应的右奇异向量处取到。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

上一节要点：矩阵的诱导范数

- 诱导范数 $\|\boldsymbol{A}\| = \max_{\|\boldsymbol{w}\|=1} \|\boldsymbol{A}\boldsymbol{w}\|$ 给出最紧的放大倍数，是最小的相容范数。
- 三种算法：列和范数、行和范数、谱范数（最大奇异值）。
- 几何图景：单位球被 $\boldsymbol{A}$ 映成椭球，最长半轴即谱范数。

本节：三种相容的矩阵范数

- 另取三种便于手算的相容范数 m_1、m_∞、m_2。
- 验证它们都是矩阵范数，并分别与 1、∞、2-范数相容。
- 与诱导范数对照，并给出一个特征值下界。

【定义 **3-11**】三种相容矩阵范数

设 $\boldsymbol{A}=(a_{ij})_{m\times n}$，$\boldsymbol{w}=(w_1,\ldots,w_n)^T$，定义

$$\|\boldsymbol{A}\|_{m_1} \triangleq \sum_{i,j}|a_{ij}|, \qquad \text{与 } \|\boldsymbol{w}\|_1 \text{ 相容;}$$

$$\|\boldsymbol{A}\|_{m_\infty} \triangleq n\max_{i,j}|a_{ij}|, \qquad \text{与 } \|\boldsymbol{w}\|_\infty \text{ 相容;}$$

$$\|\boldsymbol{A}\|_{m_2} \triangleq \left(\sum_{i,j}|a_{ij}|^2\right)^{1/2}, \qquad \text{与 } \|\boldsymbol{w}\|_2 \text{ 相容.}$$

三者都是矩阵范数，且分别与所示向量范数相容（验证见后续例题）。

例 3-23：计算矩阵一范数

设 $\boldsymbol{A}=\begin{pmatrix}1 & -2\\ 0 & 4\end{pmatrix}$。求 $\|\boldsymbol{A}\|_{m_1}$。

例 3-23：计算矩阵一范数

对 $\boldsymbol{A}=\begin{pmatrix}1 & -2\\ 0 & 4\end{pmatrix}$，计算 m_1-范数。

解按本讲定义，对所有元素的绝对值求和：

$$\|\boldsymbol{A}\|_{m_1}=|1|+|-2|+|0|+|4|=7.$$

例 3-24：计算最大元矩阵范数

设 $\boldsymbol{A}=\begin{pmatrix}1 & -2\\ 0 & 4\end{pmatrix}$。求 $\|\boldsymbol{A}\|_{m_\infty}$。

例 3-24：计算最大元矩阵范数

对 $\boldsymbol{A}=\begin{pmatrix}1 & -2\\ 0 & 4\end{pmatrix}$，计算 m_∞-范数。

解 $\boldsymbol{A}$ 有 $n=2$ 列，元素绝对值的最大值是 4，故

$$\|\boldsymbol{A}\|_{m_\infty}=n\max_{i,j}|a_{ij}|=2\cdot 4=8.$$

注意不要漏掉列数因子 n。

例 3-25：计算矩阵平方和范数

设 $\boldsymbol{A}=\begin{pmatrix}1 & -2\\ 0 & 4\end{pmatrix}$。求 $\|\boldsymbol{A}\|_{m_2}$。

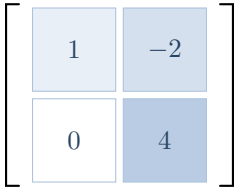
例 3-25：计算矩阵平方和范数

对 $\boldsymbol{A}=\begin{pmatrix}1 & -2\\ 0 & 4\end{pmatrix}$，计算 m_2-范数。

解各元素模平方求和再开方：

$$\|\boldsymbol{A}\|_{m_2}=\sqrt{|1|^2+|-2|^2+|0|^2+|4|^2}=\sqrt{21}.$$

$$\boldsymbol{A} = \begin{pmatrix} 1 & -2 \\ 0 & 4 \end{pmatrix}, \qquad n = 2$$



$$\|\boldsymbol{A}\|_{m_1} = 1 + 2 + 0 + 4 = 7$$

$$\|\boldsymbol{A}\|_{m_\infty} = n \max_{i,j} |a_{ij}| = 2 \cdot 4 = 8$$

$$\|\boldsymbol{A}\|_{m_2} = \sqrt{1 + 4 + 0 + 16} = \sqrt{21}$$

色块深浅表示 $|a_{ij}|$ 的大小；本讲的 m_∞ 定义包含列数因子 n。

把 $\boldsymbol{A}$ 的所有元素排成一个长向量 $\mathrm{vec}(\boldsymbol{A})$，则 m_1, m_2 与之逐字对应：

$$\|\boldsymbol{A}\|_{m_1} = \|\mathrm{vec}(\boldsymbol{A})\|_1, \qquad \|\boldsymbol{A}\|_{m_2} = \|\mathrm{vec}(\boldsymbol{A})\|_2 \text{ (即 Frobenius 范数).}$$

这是很好的记忆钩子，但边界要清楚：

m_∞ 的因子 n：拉直看不到乘法

$\|\mathrm{vec}(\boldsymbol{A})\|_\infty = \max_{i,j}|a_{ij}|$，而本讲 $\|\boldsymbol{A}\|_{m_\infty} = n\max_{i,j}|a_{ij}|$。多出的 n 是为满足矩阵范数特有的 次可乘性 $\|\boldsymbol{AB}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$：若去掉 n，取 $\boldsymbol{A} = \boldsymbol{B}$ 为全 1 矩阵，则 $\max|(\boldsymbol{AB})_{ij}| = n > 1 = \max|a_{ij}| \cdot \max|b_{ij}|$，次可乘性即失败。

矩阵范数区别于向量范数的全部新意，就在于矩阵有乘法、还要与向量范数 相容——这两者都是" 拉直" 后看不见的结构。

向量范数量"长度"，而矩阵是线性映射：我们更关心它能把向量放大多少、两个映射复合后又放大多少。这迫使矩阵范数在范数三条件之外，多担一份责任。

次可乘不是规定，是"想要相容"逼出来的

一旦要求矩阵范数与向量范数相容，即 $\|\boldsymbol{A}\boldsymbol{w}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{w}\|$，那么对任意 $\boldsymbol{w}$：

$$\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{w}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\boldsymbol{w}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|\,\|\boldsymbol{w}\|,$$

于是必有 $\|\boldsymbol{A}\boldsymbol{B}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$。相容 $\Rightarrow$ 次可乘——这条公理是"想要相容"的必然代价。

多这一条换来三样好处：

- 幂可控：$\left\|\boldsymbol{A}^k\right\| \le \|\boldsymbol{A}\|^k$，迭代与级数收敛的判据由此而来；
- 连续性：矩阵多项式、指数 $e^{\boldsymbol{A}}$ 成为连续映射，才能谈微分；
- 扰动可估：误差如 $\left\|(\boldsymbol{A}+\boldsymbol{E})^{-1}-\boldsymbol{A}^{-1}\right\|$ 可以放缩，数值稳定性有了语言。

为避免记号混淆，取方阵 $\boldsymbol{A}=(a_{ij})\in\mathbb{C}^{n\times n}$ 作对照：

指标	本讲选取的矩阵范数	诱导矩阵范数（外号）
1	$\|\boldsymbol{A}\|_{m_1}=\sum_{i,j}\|a_{ij}\|$	$\|\boldsymbol{A}\|_1=\max_j\sum_i\|a_{ij}\|$（列和范数）
∞	$\|\boldsymbol{A}\|_{m_\infty}=n\max_{i,j}\|a_{ij}\|$	$\|\boldsymbol{A}\|_\infty=\max_i\sum_j\|a_{ij}\|$（行和范数）
2	$\|\boldsymbol{A}\|_{m_2}=\left(\sum_{i,j}\|a_{ij}\|^2\right)^{1/2}$	$\|\boldsymbol{A}\|_2=\sigma_{\max}(\boldsymbol{A})$（谱范数）

【结论准备】

令

$$\boldsymbol{w} = (|a_1|, \dots, |a_m|), \qquad \boldsymbol{v} = (|b_1|, \dots, |b_m|).$$

由 Cauchy–Schwarz 不等式，

$$(\boldsymbol{w}, \boldsymbol{v}) \le \sqrt{(\boldsymbol{w}, \boldsymbol{w})}\sqrt{(\boldsymbol{v}, \boldsymbol{v})},$$

即

$$\sum_{i=1}^{m} |a_i|\,|b_i| \le \left(\sum_{i=1}^{m} |a_i|^2\right)^{1/2} \left(\sum_{i=1}^{m} |b_i|^2\right)^{1/2}.$$

这个不等式我们并不陌生：第 3.1 节证向量 2-范数的三角不等式、第 3.4 节证 Frobenius 范数的次可乘性时都已用过它的一般形式；这里取绝对值形式，专为例 3-26 中 m_2 的次可乘性与相容性证明服务。

例 3-26：验证三种矩阵范数及其与对应向量范数的相容性*

设 $\boldsymbol{A} \in \mathbb{C}^{m\times n}$，$\boldsymbol{w} \in \mathbb{C}^{n}$。验证 m_1、m_∞、m_2 都是矩阵范数，并与对应向量范数相容：

- $\|\boldsymbol{A}\boldsymbol{w}\|_1 \leq \|\boldsymbol{A}\|_{m_1} \|\boldsymbol{w}\|_1$；
- $\|\boldsymbol{A}\boldsymbol{w}\|_\infty \leq \|\boldsymbol{A}\|_{m_\infty} \|\boldsymbol{w}\|_\infty$；
- $\|\boldsymbol{A}\boldsymbol{w}\|_2 \leq \|\boldsymbol{A}\|_{m_2} \|\boldsymbol{w}\|_2$。

例 3-26：验证三种矩阵范数及其与对应向量范数的相容性*

验证 m_1-矩阵范数及其与向量 1-范数的相容性。

解（**1**）正定性、齐次性和三角不等式由绝对值性质直接成立。设 $\boldsymbol{B}=(b_{kj})\in\mathbb{C}^{n\times l}$，则

$$\begin{aligned}\|\boldsymbol{AB}\|_{m_1}&=\sum_{i,j}\left|\sum_k a_{ik}b_{kj}\right|\\&\le\sum_{i,j,k}|a_{ik}|\,|b_{kj}|\\&\le\left(\sum_{i,k}|a_{ik}|\right)\left(\sum_{k,j}|b_{kj}|\right)\\&=\|\boldsymbol{A}\|_{m_1}\|\boldsymbol{B}\|_{m_1}.\end{aligned}$$

例 3-26：验证三种矩阵范数及其与对应向量范数的相容性*

证明 $\|\boldsymbol{A}\boldsymbol{w}\|_1 \leq \|\boldsymbol{A}\|_{m_1} \|\boldsymbol{w}\|_1$。

解（**1**）续此外，逐项使用 $|w_j| \leq \|\boldsymbol{w}\|_1$，得到

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{w}\|_1 &= \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij} w_j \right| \\ &\leq \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}| \, |w_j| \\ &\leq \left(\sum_{i,j} |a_{ij}| \right) \|\boldsymbol{w}\|_1 = \|\boldsymbol{A}\|_{m_1} \|\boldsymbol{w}\|_1 .\end{aligned}$$

例 3-26：验证三种矩阵范数及其与对应向量范数的相容性*

验证 m_∞-矩阵范数及其与向量 ∞-范数的相容性。

解（**2**）正定性、齐次性和三角不等式直接成立。设 $\boldsymbol{B}=(b_{ij})_{n\times l}$，则 $(\boldsymbol{AB})_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj}$。于是

$$\begin{aligned}\|\boldsymbol{AB}\|_{m_\infty} &= l\max_{i,j}\left|\sum_{k=1}^{n}a_{ik}b_{kj}\right| \\ &\le l\max_{i,j}\sum_{k=1}^{n}|a_{ik}|\,|b_{kj}| \\ &\le \left(n\max_{i,j}|a_{ij}|\right)\left(l\max_{i,j}|b_{ij}|\right) \\ &= \|\boldsymbol{A}\|_{m_\infty}\|\boldsymbol{B}\|_{m_\infty}.\end{aligned}$$

例 3-26：验证三种矩阵范数及其与对应向量范数的相容性*

证明 $\|\boldsymbol{A}\boldsymbol{w}\|_{\infty} \leq \|\boldsymbol{A}\|_{m_\infty} \|\boldsymbol{w}\|_{\infty}$。

解（**2**）续另一方面，

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{w}\|_{\infty} &= \max_i |a_{i1}w_1 + \cdots + a_{in}w_n| \\ &\leq \max_i \left(|a_{i1}| + \cdots + |a_{in}|\right) \max_j |w_j| \\ &\leq \left(n \max_{i,j} |a_{ij}|\right) \|\boldsymbol{w}\|_{\infty} \\ &= \|\boldsymbol{A}\|_{m_\infty} \|\boldsymbol{w}\|_{\infty}.\end{aligned}$$

故 $\|\cdot\|_{m_\infty}$ 与 $\|\cdot\|_{\infty}$ 相容。

例 3-26：验证三种矩阵范数及其与对应向量范数的相容性*

验证 m_2-矩阵范数及其与向量 2-范数的相容性。

解（**3**）正定性和齐次性直接成立。对三角不等式，将 $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{m\times n}$ 按列写成 $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$、$\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$，则

$$
\begin{aligned}
\|\boldsymbol{A}+\boldsymbol{B}\|_{m_2}^2 &= \sum_{j=1}^{n} \|\boldsymbol{a}_j + \boldsymbol{b}_j\|_2^2 \\
&\le \sum_{j=1}^{n} \left(\|\boldsymbol{a}_j\|_2 + \|\boldsymbol{b}_j\|_2\right)^2 \\
&\le \|\boldsymbol{A}\|_{m_2}^2 + 2\left(\sum_j \|\boldsymbol{a}_j\|_2^2\right)^{1/2}\left(\sum_j \|\boldsymbol{b}_j\|_2^2\right)^{1/2} + \|\boldsymbol{B}\|_{m_2}^2 \\
&= \left(\|\boldsymbol{A}\|_{m_2} + \|\boldsymbol{B}\|_{m_2}\right)^2.
\end{aligned}
$$

解（**3**）续对乘法不等式，由 Cauchy–Schwarz 不等式，

$$
\begin{aligned}
\|\boldsymbol{AB}\|_{m_2}^2 &= \sum_{i,j} \left| \sum_k a_{ik} b_{kj} \right|^2 \\
&\le \sum_{i,j} \left(\sum_k |a_{ik}|^2 \right) \left(\sum_k |b_{kj}|^2 \right) \\
&= \left(\sum_{i,k} |a_{ik}|^2 \right) \left(\sum_{k,j} |b_{kj}|^2 \right) \\
&= \|\boldsymbol{A}\|_{m_2}^2 \, \|\boldsymbol{B}\|_{m_2}^2 .
\end{aligned}
$$

例 3-26：验证三种矩阵范数及其与对应向量范数的相容性*

证明 m_2-矩阵范数与向量 2-范数相容。

解（**3**）续当 $\boldsymbol{B}=\boldsymbol{w}\in\mathbb{C}^{n\times 1}$ 时，m_2-范数的乘法不等式退化为

$$\|\boldsymbol{A}\boldsymbol{w}\|_2=\|\boldsymbol{A}\boldsymbol{B}\|_{m_2}\leq\|\boldsymbol{A}\|_{m_2}\|\boldsymbol{B}\|_{m_2}=\|\boldsymbol{A}\|_{m_2}\|\boldsymbol{w}\|_2\,.$$

例 3-27：特征值的下界

设 $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ 可逆，λ 是 $\boldsymbol{A}$ 的任一特征值。证明

$$|\lambda| \geq \frac{1}{\left\|\boldsymbol{A}^{-1}\right\|}.$$

例 3-27：特征值的下界

设 $\boldsymbol{A}$ 可逆，λ 为任一特征值，证明 $|\lambda| \geq 1/\left\|\boldsymbol{A}^{-1}\right\|$。

证明取 λ 对应的特征向量 $\boldsymbol{x} \neq \boldsymbol{0}$，即 $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$。因 $\boldsymbol{A}$ 可逆，$\lambda \neq 0$，且 $\boldsymbol{x} = \boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x})$。
由相容性 $\left\|\boldsymbol{A}^{-1}\boldsymbol{v}\right\| \leq \left\|\boldsymbol{A}^{-1}\right\| \|\boldsymbol{v}\|$，取 $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{x}$：

$$\|\boldsymbol{x}\| = \left\|\boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x})\right\| \leq \left\|\boldsymbol{A}^{-1}\right\| \|\boldsymbol{A}\boldsymbol{x}\| = \left\|\boldsymbol{A}^{-1}\right\| |\lambda| \|\boldsymbol{x}\|.$$

因为 $\boldsymbol{x} \neq \boldsymbol{0}$，故 $\|\boldsymbol{x}\| > 0$，约去得 $1 \leq |\lambda| \left\|\boldsymbol{A}^{-1}\right\|$，即

$$|\lambda| \geq \frac{1}{\left\|\boldsymbol{A}^{-1}\right\|}.$$

例 3-27：特征值的下界

设 $\boldsymbol{A}$ 可逆，λ 为任一特征值，证明 $|\lambda| \geq 1/\left\|\boldsymbol{A}^{-1}\right\|$。

两点说明

- 结论也可读作“相容范数不小于谱半径”：

$$\left\|\boldsymbol{A}^{-1}\right\| \;\geq\; \rho(\boldsymbol{A}^{-1}) = \frac{1}{\min_i |\lambda_i|}.$$

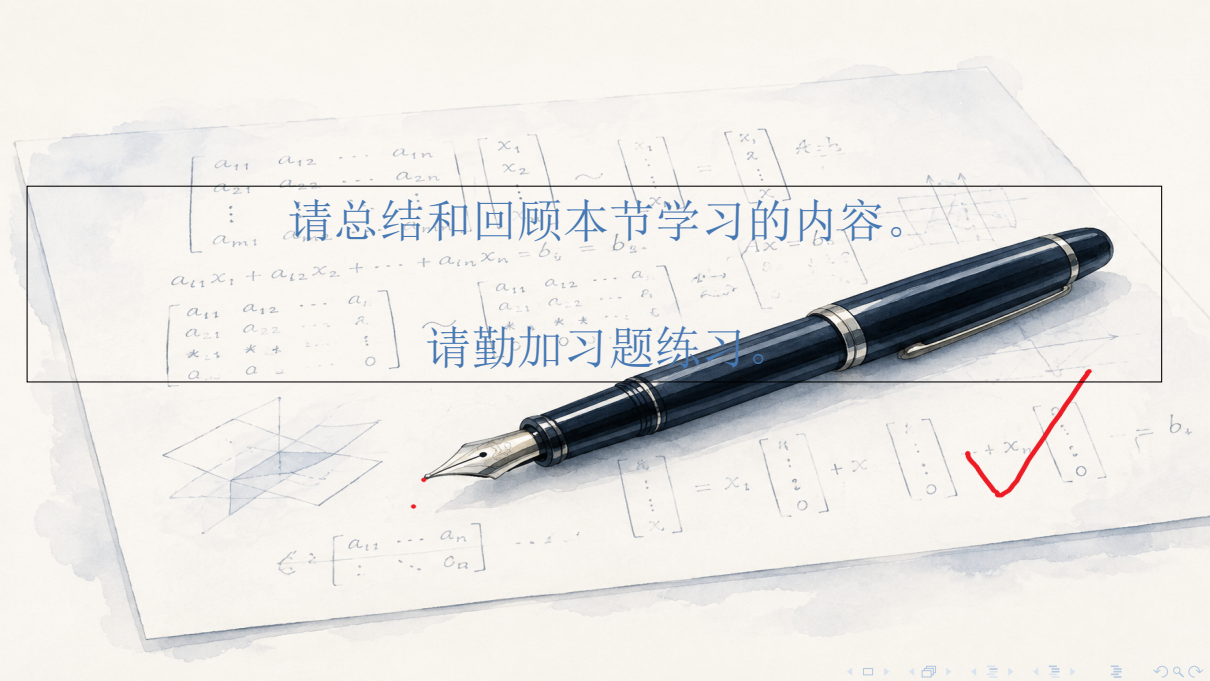
- 若取诱导 2-范数，则 $\left\|\boldsymbol{A}^{-1}\right\|_2 = 1/\sigma_{\min}(\boldsymbol{A})$，结论化作

$$\sigma_{\min}(\boldsymbol{A}) \;\leq\; |\lambda| \;\leq\; \sigma_{\max}(\boldsymbol{A}),$$

即特征值落在最小与最大奇异值之间。

请总结和回顾本节学习的内容。

请勤加习题练习。



1 矩阵梯度及其应用 3
- 梯度的定义与记号 6
- 向量变量的梯度公式 11
- 矩阵变量的梯度公式 24

2 线性回归 39
- 线性回归与一般意义下的回归问题 40
- 与王文杰讨论：用本章的数学记号理解神经网络下的回归问题 46
- 回归问题中的记号准备 52
- 线性回归与最小二乘法 66
- 正则化线性回归（Ridge 与 LASSO） 100
- LASSO 回归与 Python 代码示例 117

3 范数理论及其应用 127
- 向量范数 128
- 矩阵诱导范数与加权线性回归 168
- 向量范数的等价性* 187
- 矩阵范数 200
- 矩阵的诱导范数 210
- 三种相容的矩阵范数 232
- Frobenius 范数的迹运算* 258

上一节要点：三种相容的矩阵范数

- m_1 取全部元素模之和；m_∞ 取列数乘最大元；m_2 取平方和开方。
- 三者都是矩阵范数，并分别与 1、∞、2-范数相容。
- m_2 正是把矩阵拉直为向量后的 2-范数；相容性给出 $\|\boldsymbol{A}\boldsymbol{w}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{w}\|$。

本节：Frobenius 范数的迹运算（略讲）

- 同一个范数的另一种写法：$\|\boldsymbol{A}\|_F = \left[\mathrm{tr}(\boldsymbol{A}^H\boldsymbol{A})\right]^{1/2}$。
- 三个好处：坐标无关、连通奇异值、一行证出正交不变。
- 用它计算范数，并估计矩阵与向量范数的相容上界。

Frobenius 范数已在第 3.4 节定义过（平方和形式，并验证了它是矩阵范数），第 3.6 节又把它列为相容范数 m_2。本小节给它一个等价的迹形式——这不是新定义，而是同一个范数的另一种写法：

$$\|\boldsymbol{A}\|_F \triangleq \|\boldsymbol{A}\|_{m_2} = \Big[\operatorname{tr}(\boldsymbol{A}^H\boldsymbol{A})\Big]^{1/2} = \Big[\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}^H)\Big]^{1/2}.$$

为何绕到迹？不是为算得快（先求 $\boldsymbol{A}^H\boldsymbol{A}$ 反而更绕），而是它带来三样平方和写法给不了的好处——坐标无关、连通奇异值、正交不变。下面分三帧逐一说明。

回忆第 3.4 节的图景：矩阵是线性变换 $\mathcal{A}$ 在某组基下的坐标表。记 $\mathcal{A}$ 在原基 $\mathcal{B}$ 下的矩阵为 $\boldsymbol{A}_{\mathcal{B}}$；换一组基 $\mathcal{B}'$，同一个 $\mathcal{A}$ 的矩阵就变成另一个数表 $\boldsymbol{A}_{\mathcal{B}'}$。

平方和看不出不变，迹一眼看出

逐元素写法 $\sum_{i,j}|a_{ij}|^2$ 随基而变的形式藏在 a_{ij} 里，无法直接读出它是否不变。而迹在换基（相似变换）下不变：$\mathrm{tr}(\boldsymbol{P}^{-1}\boldsymbol{M}\boldsymbol{P})=\mathrm{tr}(\boldsymbol{M})$。所以 $\mathrm{tr}(\boldsymbol{A}^H\boldsymbol{A})$ 这个写法，把" 不依赖坐标" 写在了脸上。

结论：即便矩阵从 $\boldsymbol{A}_{\mathcal{B}}$ 变成 $\boldsymbol{A}_{\mathcal{B}'}$，Frobenius 范数仍量的是变换 $\mathcal{A}$ 本身的大小——它是个几何量，不随描述方式而变。这正是" 范数" 该有的资格。

$\boldsymbol{A}^H\boldsymbol{A}$ 半正定，其特征值是 $\boldsymbol{A}$ 的奇异值平方。迹等于特征值之和，于是

$$\|\boldsymbol{A}\|_F^2 = \operatorname{tr}(\boldsymbol{A}^H\boldsymbol{A}) = \sum_i \lambda_i(\boldsymbol{A}^H\boldsymbol{A}) = \sum_i \sigma_i(\boldsymbol{A})^2, \qquad \|\boldsymbol{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}.$$

从" 数元素" 升级为" 谱语言"

Frobenius 范数 = 全部奇异值的 **2**-范数；而谱范数 $\|\boldsymbol{A}\|_2 = \sigma_{\max}$ 只取最大奇异值。两者一个是" 全体能量的平方和开方"、一个是" 最强方向"，关系一目了然——这是后面比较"Frobenius 上界 vs 最大放大率" 的理论根子。逐元素平方和的写法做不到这一点。

要证" 对酉矩阵 $\boldsymbol{U}$ 有 $\|\boldsymbol{UA}\|_F = \|\boldsymbol{A}\|_F$"，用迹只需一行：

$$\|\boldsymbol{UA}\|_F^2 = \operatorname{tr}\left((\boldsymbol{UA})^H(\boldsymbol{UA})\right) = \operatorname{tr}(\boldsymbol{A}^H\boldsymbol{U}^H\boldsymbol{UA}) = \operatorname{tr}(\boldsymbol{A}^H\boldsymbol{A}) = \|\boldsymbol{A}\|_F^2.$$

靠的是迹的循环性

关键是 $\operatorname{tr}(\boldsymbol{XY}) = \operatorname{tr}(\boldsymbol{YX})$，让 $\boldsymbol{U}^H\boldsymbol{U} = \boldsymbol{I}$ 自然消去。若把 Frobenius 范数只定义成" 平方和开方"，这条好性质就得逐元素硬算，几乎证不动。正是" 想要正交不变性" 反过来要求用迹来定义。

例 3-28：用迹计算 Frobenius 范数

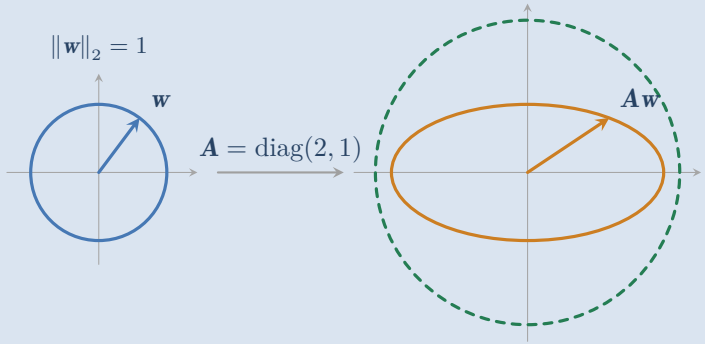
设 $\boldsymbol{A} = \mathrm{diag}(3, 4)$。利用迹公式求 $\|\boldsymbol{A}\|_F$。

例 3-28：用迹计算 Frobenius 范数

对 $\boldsymbol{A} = \mathrm{diag}(3, 4)$，用迹公式计算 Frobenius 范数。

解 $\boldsymbol{A}^H\boldsymbol{A} = \mathrm{diag}(9, 16)$，所以

$$\|\boldsymbol{A}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{A}^H\boldsymbol{A})} = \sqrt{9 + 16} = 5.$$



橙色椭圆是单位圆盘经 $\boldsymbol{A}$ 变换后的边界；绿虚线圆的半径 $\|\boldsymbol{A}\|_F = \sqrt{5}$ 给出上界：$\|\boldsymbol{A}\boldsymbol{w}\|_2 \le \|\boldsymbol{A}\|_F \|\boldsymbol{w}\|_2$。

例 3-29：数值核对范数相容不等式

取 $\boldsymbol{A}=\mathrm{diag}(2,1)$、$\boldsymbol{w}=(1,1)^T$。计算并比较 $\|\boldsymbol{A}\boldsymbol{w}\|_2$ 与 $\|\boldsymbol{A}\|_F\|\boldsymbol{w}\|_2$。

例 3-29：数值核对范数相容不等式

对给定的对角矩阵与向量，核对 Frobenius 范数的相容不等式。

解 $\boldsymbol{A}\boldsymbol{w} = (2,1)^T$，故

$$\|\boldsymbol{A}\boldsymbol{w}\|_2 = \sqrt{5}, \qquad \|\boldsymbol{A}\|_F \|\boldsymbol{w}\|_2 = \sqrt{5}\sqrt{2} = \sqrt{10}.$$

因此对这组数据有 $\sqrt{5} \leq \sqrt{10}$。

例 3-30：Frobenius 上界是否紧

令 $\boldsymbol{A}=\operatorname{diag}(2,1)$。在 $\|\boldsymbol{w}\|_2=1$ 的条件下，求 $\|\boldsymbol{A}\boldsymbol{w}\|_2$ 的最大值，并与相容性上界 $\|\boldsymbol{A}\|_{m_2}\|\boldsymbol{w}\|_2$ 比较。这个矩阵范数上界是否总是紧的？

例 3-30：Frobenius 上界是否紧

对 $\boldsymbol{A} = \operatorname{diag}(2, 1)$，比较单位向量上的最大放大与 Frobenius 范数上界。

解若 $\boldsymbol{w} = (w_1, w_2)^T$ 且 $\|\boldsymbol{w}\|_2 = 1$，则

$$\|\boldsymbol{A}\boldsymbol{w}\|_2^2 = 4\,|w_1|^2 + |w_2|^2 \le 4\big(|w_1|^2 + |w_2|^2\big) = 4.$$

取 $\boldsymbol{w} = (1, 0)^T$ 可取到等号，所以最大值为 2。而 $\|\boldsymbol{A}\|_{m_2} = \sqrt{5}$，相容性给出的上界 $\sqrt{5}$ 有效但不紧。对这个矩阵，诱导 2-范数（谱范数）恰为最大放大率 2。

命题 3-7：有限维矩阵空间中的范数与收敛

在 $\mathbb{C}^{m\times n}$ 中，任意两种矩阵范数等价：存在 $0 < c_1 \le c_2$，使得

$$c_1 \|\boldsymbol{A}\|_\beta \le \|\boldsymbol{A}\|_\alpha \le c_2 \|\boldsymbol{A}\|_\beta .$$

因此矩阵序列的收敛与所选矩阵范数无关：

$$\boldsymbol{A}^{(k)} \to \boldsymbol{A} \quad \Longleftrightarrow \quad \left\|\boldsymbol{A}^{(k)} - \boldsymbol{A}\right\| \to 0.$$

证明将矩阵按元素展为 $\mathbb{C}^{mn}$ 中的向量；矩阵范数满足向量范数的前三条公理。由有限维空间范数等价定理及范数判别收敛定理即得。

定理 3-8：置换矩阵与酉矩阵的作用

对于 $\boldsymbol{A}_{m\times n}$ 及置换矩阵 $\boldsymbol{P}_{m\times m}$、$\boldsymbol{Q}_{n\times n}$，有

$$\|\boldsymbol{PA}\|_F = \|\boldsymbol{A}\|_F, \qquad \|\boldsymbol{AQ}\|_F = \|\boldsymbol{A}\|_F.$$

更一般地，上式对满足 $\boldsymbol{P}^H\boldsymbol{P} = \boldsymbol{I}$、$\boldsymbol{QQ}^H = \boldsymbol{I}$ 的相应酉矩阵成立。

证明

$$\begin{aligned}\|\boldsymbol{PA}\|_F^2 &= \operatorname{tr}\big((\boldsymbol{PA})^H(\boldsymbol{PA})\big)\\ &= \operatorname{tr}(\boldsymbol{A}^H\boldsymbol{P}^H\boldsymbol{PA}) = \operatorname{tr}(\boldsymbol{A}^H\boldsymbol{A}) = \|\boldsymbol{A}\|_F^2,\\ \|\boldsymbol{AQ}\|_F^2 &= \operatorname{tr}\big((\boldsymbol{AQ})(\boldsymbol{AQ})^H\big)\\ &= \operatorname{tr}(\boldsymbol{AQQ}^H\boldsymbol{A}^H) = \operatorname{tr}(\boldsymbol{AA}^H) = \|\boldsymbol{A}\|_F^2.\end{aligned}$$

例 3-31：最小二乘分类器与 one-hot 编码

考虑一个有 K 个类别的分类问题。令 $\boldsymbol{Y}=(y_{ik})\in\{0,1\}^{n\times K}$ 为 one-hot 标签矩阵（每行恰有一个 1），$\boldsymbol{X}\in\mathbb{R}^{n\times p}$、$\boldsymbol{W}\in\mathbb{R}^{p\times K}$，并以 $\boldsymbol{XW}$ 作为类别分数矩阵。写出 Frobenius 平方损失与新样本的分类规则，并说明类别分数是否为概率。

例 3-31：最小二乘分类器与 one-hot 编码

拟合 one-hot 类别标签并按分数判类；预测分数是否就是概率？

解记 $\boldsymbol{w}_k$ 为 $\boldsymbol{W}$ 的第 k 列，则

$$\mathcal{J}(\boldsymbol{W}) = \frac{1}{nK}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_F^2 = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(y_{ik} - \boldsymbol{x}_i^T\boldsymbol{w}_k\right)^2,$$

这是对 nK 个样本-类别误差取平均（$K=1$ 时即普通 MSE）。新样本的分数向量为 $\boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{W}^T\boldsymbol{x}$，故 $\widehat{c}(\boldsymbol{x}) = \arg\max_{1\le k\le K} s_k(\boldsymbol{x})$；例如分数 $(0.2, 1.1, 0.3)^T$ 判为第 2 类。这些分数不受 $[0,1]$ 约束且和不必为 1，不是概率；若需概率输出，可改用 softmax 与交叉熵。

请总结和回顾本节学习的内容。

请勤加习题练习。

