磐石计划 01

基本数学语言与深度学习 · 随题复习讲义

v3.0 掌握型学习版。不是额外题库；本书保留 12 个知识专题，把练习入口统一对齐 A / B 两卷原 21 题和每题订正环节。遇到障碍时按题号查本书，不必先通读后才开始。

本书怎样支持你完成两卷

读公式时先问对象类型、定义域、谁固定谁变化；做运算时保留中间状态；做证明时写每个条件作用在哪里；订正时先遮盖恢复原题，再改变一个条件。以下讲解依据对话中的教材截图及前版复习讲义，经典论文只在明确标出的特例中扩展。

知识链	作答与订正入口	对应章节
符号与函数族	A1→B1→A19	第 1、9 章
矩阵、向量与展平	A3/A4→B3/B4→B14→B20	第 2、9 章
复合、激活与表达	A2/A8/A12→B2/B16→A20→B17	第 3、10 章
监督学习与风险	A5/A13→A17→B13→B19	第 4、8 章
导数、反传与优化	B5→A18→B8→B18→B21	第 5、6、11 章
范数、积分与误差	A7/A11/A14→B7/B11→B20	第 7、9 章
可辨识性与压缩	B1→A19→B12/B14→B20	第 1、9 章
稳定性、归纳偏置与算力	A10/A15→B10→A21→B21	第 11、12 章

先辨别：定义、恒等变形、定理、算法

“用平方衡量损失”是一个定义选择；“将预测代进平方”是恒等替换；“在给定条件下有限次更新”是需证明的定理；“按负梯度改变参数”是一种算法。不是每个公式都来自唯一必然的推导。后文分别说明如何计算、怎样想到和什么时候成立。

第 1 章　先认识对象，再理解公式

本章目标：能说清每个符号是什么类型、哪些量固定、哪些量正在变化。

1.1 数学公式不全是“推出来的”

定义是在约定一个对象叫什么、如何计算；恒等变形是在已有定义上按代数规则重写；定理是在假设下证明一个结论；算法是在多种可能方案中选择一种操作规则。学习时先判断句子属于哪一种，才能知道该问“为什么这样选”还是“这一步凭什么成立”。

类别	例子	应追问
定义	$\ell_y(z) = (z-y)^2$	为什么选平方而非绝对值?
恒等变形	$(wx+b-y)^2$	把预测代入哪里?
定理	$0 < \eta < 2/a$ 下特定二次损失下降	哪些假设使结论成立?
算法	$\theta^+ = \theta - \eta\nabla L$	这是一种更新规则，不是唯一选择

1.2 空间、元素、参数与模型族

$\mathcal{X}$ 是允许输入的集合，$x \in \mathcal{X}$ 是一个输入；Θ 是全部允许参数的集合，$\theta \in \Theta$ 是一组参数。若模型 $m(x;(w,b)) = wx+b$，则 $x \in \mathbb{R}$，$\theta = (w,b) \in \mathbb{R}^2$。

$$m : \mathbb{R} \times \mathbb{R}^2 \longrightarrow \mathbb{R}, \qquad (x,(w,b)) \longmapsto wx+b.$$

这里的 $\times$ 是笛卡尔积，组合输入与参数，不是把集合相乘成数。分号强调角色不同。固定参数 $\theta = (3,-2)$ 后，得到单变量函数 $x \mapsto 3x-2$；模型族则把所有参数对应的完整函数收集起来。

$$\mathcal{F} = \{m(\cdot;\theta) : \theta \in \Theta\}.$$

点号 $\cdot$ 是输入位置的占位符。θ^* 常标最优 / 目标参数；θ' 常是另一个候选参数而非导数；函数上的撇号 $L'(w)$ 才在此表示导数。符号不能脱离上下文解释。

1.3 例题：参数多，不同函数一定多吗？

设 $m(x;(a,b)) = (a+b)x$，参数可为任意实数。$(1,2)$ 与 $(0,3)$ 都给出 $3x$。因此从“参数空间”到“函数集合”的对应可以多对一。即便知道完整函数，也未必能恢复唯一参数。

判断两个函数相同，须在它们共同的定义域上对每个输入比较；只在几个训练点取值相同，不足以证明函数相同。

1.4 两种最优化符号

$$\min_{t\in[-2,2]} (t^2-1)^2 = 0, \qquad \operatorname*{argmin}_{t\in[-2,2]} (t^2-1)^2 = \{-1,1\}.$$

min 返回最小函数值；argmin 返回取得最小值的自变量集合。最小点可能不唯一，也可能不存在。例如在开区间 (0,1) 上，f(t)=t 的下确界是 0，却没有取得 0 的点。约束集合是问题的一部分，不能丢掉。

【在原题中用起来】回到 A1、A2、A6、B1、B15。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

第 2 章　矩阵没有“变成”向量：维度是运算语法

本章目标：能区别线性映射、输出对象、参数展平和改变模型结构。

2.1 四类运算必须分清

$$W \in \mathbb{R}^{p\times d},\ x \in \mathbb{R}^{d} \quad \Rightarrow \quad Wx \in \mathbb{R}^{p}.$$

W 是一条把 d 维向量变成 p 维向量的线性规则；Wx 是执行后的结果。W 本身没有变成向量。列向量可看作 d×1 矩阵，所以形状检查是 $(p \times d)(d \times 1) = (p \times 1)$。

运算	形状 / 结果	含义
$u^{\top}v$	标量	对应相乘，再全部相加
$u \odot v$	与 u、v 同形	每一位置独立相乘
$uv^{\top}$	行数为 u 维数、列数为 v 维数	外积，每个元素 $u_i v_j$
Wx	向量	每一行对 x 做一次内积

2.2 完整例题：先算线性层再激活

$$W = \begin{pmatrix} 1 & -2 \\ 0 & 3 \\ 2 & 1 \end{pmatrix}, \quad x = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix}.$$

$$Wx = \begin{pmatrix} 4 \\ -3 \\ 3 \end{pmatrix}, \quad Wx + b = \begin{pmatrix} 3 \\ -1 \\ 3 \end{pmatrix}, \quad \rho(Wx + b) = \begin{pmatrix} 3 \\ 0 \\ 3 \end{pmatrix}.$$

顺序不可交换：先对 x 激活通常不是同一个网络。运算每一步都保留形状，能发现大多数“答案像是对的”错误。

2.3 展平只是编码

一个网络的参数可能包括多个矩阵和偏置。把它们按固定顺序排成长向量 θ，只是为了统一优化记号。计算前向传播时，必须按原约定恢复每个矩阵的形状。

$$p_{\text{param}} = dh + h + ho + o$$

这是一个输入 d 维、隐藏 h 维、输出 o 维、两层均含偏置的全连接网络的实际存储参数数。它不自动等于函数空间的有效自由度。

2.4 仿射层的合并

$$y = W_2(W_1x + b_1) + b_2 = (W_2W_1)x + (W_2b_1 + b_2).$$

任意有限层仿射变换在前向函数上仍为仿射变换。没有非线性，深度不能突破这一函数类别；但多层参数化与单层参数化的训练轨迹、约束及冗余可以不同。

二次模型把参数写成矩阵，是选择了另一个函数族；展平则没有改变函数族。不要把“换模型”和“换存储方式”混为一谈。

【在原题中用起来】回到 A3、A4、A12、A19、B3、B4、B14。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

【三个“变化”务必分清】Wx 是运算得到向量；把 W 排成 θ 是存储编码；从 $w^{\top}x$ 改为 $x^{\top}Ax$ 是更换模型族。这三件事不能混说。

第 3 章　从激活到感知机：线性规则的能力边界

本章目标：会使用分段函数、分类不等式与几何间隔，不把经验事实当全称证明。

3.1 为什么非线性不可省略

ReLU 定义为 $\rho(t)=\max(0,t)$。它在每一侧都是线性的，但整条实数轴上的斜率发生改变，因此不是全局线性函数。两条基本恒等式可从 t≥0 与 t<0 分别验证：

$$\rho(t)+\rho(-t)=|t|,\qquad \rho(t)-\rho(-t)=t.$$

这些关系可将绝对值与线性通道写成 ReLU 网络，也是构造性题目的常用工具。并不是“用 ReLU 就处处弯曲”；一个固定激活模式内的网络仍可能是仿射的。

3.2 Sigmoid 与双曲正切

$$s(t)=\frac{1}{1+e^{-t}},\quad s(-t)=1-s(t),\quad s'(t)=s(t)(1-s(t)).$$

$$\tanh t=\frac{e^{t}-e^{-t}}{e^{t}+e^{-t}},\qquad \tanh(-t)=-\tanh t=2s(-2t)-1.$$

更常用的对应式为 $\tanh t=2s(2t)-1$。sigmoid 输出在 (0,1)，tanh 输出在 (-1,1)。函数趋于饱和时导数变小，但“导数小”不是最优证书。

3.3 感知机为何有偏置

$$\widehat{y}=\operatorname{sign}(w^{\top}x+b),\quad \operatorname{sign}(z)=\begin{cases}1, & z\ge 0,\\ -1, & z<0.\end{cases}$$

“加权和超过阈值 τ”为正类，可改写为 $w^{\top}x-\tau\ge 0$，所以 b=-τ。二维决策边界是直线，d 维是超平面。边界上的类别必须遵守题设，不能在证明中随意把 ≥ 改为 >。

3.4 经典间隔界的直觉 [P1]

若所有样本均能被某单位方向 u 以正间隔 γ 分开，且长度 ≤R，按非正间隔触发加法更新，那么沿 u 的投影每次至少增长 γ；由于更新前分类分数非正，向量长度平方每次至多增加 R^2。

$$k\gamma\le u^{\top}w^{(k)}\le\|w^{(k)}\|_2\le R\sqrt{k}\quad\Rightarrow\quad k\le(R/\gamma)^2.$$

关键不是背公式，而是看见两种增长速度的冲突。此界约束训练更新，不能单独控制总体分类错误率。若加入偏置，可增广输入为 (x,1)，但必须重新计算增广后的长度界和间隔。

3.5 XOR 的证明方法

对四个二进制点要求“不同为正，相同为负”，单个仿射分数对应四个不等式。把两个正类不等式相加，会与两个负类约束冲突。证明不存在时要用这种对全部参数有效的矛盾，而不是试若干参数。非线性特征 $|x_1\text{-}x_2|$ 则把两类映为 1 与 0，阈值即可区分。

【在原题中用起来】回到 A2、A4、A8、A16、A20、B2、B6、B16。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

第 4 章　监督学习的统一公式：从一条记录到训练目标

本章目标：会亲手构造损失函数，而不是把长公式作为整体记忆。

4.1 数据集是有限观察

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M}.$$

x_i 为第 i 次观察的输入，y_i 为标签，M 为观察条数。重复记录可有意义，故把数据看作多重集。在普通独立样本目标中，交换样本排列不改变平均损失；在时间序列任务里，序列内部的次序仍可能至关重要。

4.2 三种损失记号是同一计算的不同视角

$$\widehat{y} = m(x;\theta), \qquad \ell_y(\widehat{y}) = (\widehat{y} - y)^2,$$

$$\lambda_{(x,y)}(\theta) = \ell_y(m(x;\theta)), \qquad \Lambda(\theta) = \frac{1}{M}\sum_{i=1}^{M} \lambda_{(x_i,y_i)}(\theta).$$

固定真实标签看预测如何改变误差，用 ℓ；固定一条样本看参数如何改变误差，用 λ；对全部观察求平均，用 Λ。它们不是三个互不相关的训练算法。平均系数 1/M 不会改变固定数据集上的最小化点，但会缩放梯度，因此学习率也需考虑相同约定。

4.3 例题：不用优化器也能识别最优

数据为 (1,4)、(3,10)，模型 m(x;a)=ax+1。平方损失平均为

$$\Lambda(a) = \frac{1}{2}[(a-3)^2 + (3a-9)^2] = 5(a-3)^2.$$

平方项非负，所以唯一最小化点是 a=3，最小值 0。这里不用求导也能解；梯度下降则提供对复杂多参数问题仍可实施的迭代方法。若换一个不含正确规律的模型族，即使训练充分也可能留下不可消除的误差。

4.4 训练目标与求解算法

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta} \Lambda(\theta).$$

该式定义理想的经验最优参数集合，不意味着真实训练算法必然到达全局最小。实际终止点可能仅是低损失参数或近似解。反向传播算梯度；梯度下降、Adam 等使用梯度更新，两件事不能等同。

4.5 从监督到自监督与生成学习

监督学习使用给定输入—标签对。无监督学习在没有外部任务标签时寻找结构；自监督从原始数据生成预测目标，例如遮蔽一部分输入再预测。GAN 的“真实 / 生成”标签可由数据来源自动产生，不意味着每张真实图片都必须人工标注语义类别。[P12]

【在原题中用起来】回到 A5、A9、A13、A17、B5、B9。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

第 5 章　导数、梯度与反向传播：每个因子从哪里来

本章目标：从变化率定义出发，能沿计算图写出并复用链式导数。

5.1 不背求导表也能理解导数

参数从 w 变到 w+h，损失变化为 L(w+h)-L(w)。平均变化率为这两个变化量之比；让 h 趋于 0，得到在 w 处的瞬时变化率。

$$L'(w) = \lim_{h\to 0} \frac{L(w+h) - L(w)}{h}.$$

例如 L(w)=(w-4)^2，分子展开后为 2(w-4)h+h^2。对 h≠0 先约分，再取极限：

$$L'(w) = \lim_{h\to 0}[2(w-4) + h] = 2(w-4).$$

h 趋于零不是一开始就令 h=0；否则差商分母为零。导数是变化率，不是函数值，也不是已经采取的参数增量。

5.2 一阶近似为何只用“约等于”

$$\Delta L \approx L'(w)\Delta w.$$

这是差商在小增量下接近导数的表达。对上述平方函数，精确式是

$$\Delta L = 2(w-4)\Delta w + (\Delta w)^2.$$

忽略的是二次余项，而不是没有误差。步长变大后，二次项可能与一阶项同样重要。若导数本来为零，一阶项消失，不能凭一阶式判断该点是最大、最小还是其他驻点。

5.3 偏导数与梯度

若 L 依赖 w、b，对 w 求偏导时暂时固定 b，记 $\partial L/\partial w$。把全部偏导按参数顺序排成列向量，得到梯度：

$$\nabla L(w,b) = \begin{pmatrix} \partial L/\partial w \\ \partial L/\partial b \end{pmatrix}, \qquad \Delta L \approx \nabla L^{\top}\Delta\theta.$$

此处是内积而不是逐元素乘法，因为损失变化是标量。参数展平顺序改变时，梯度也要按同一顺序重排。

5.4 链式法则：中间一层也不能漏

设 z=wx+b，L=½(z-y)^2。先对 z 求导为 z-y，再对 w 求 z 的变化率为 x，所以

$$\frac{\partial L}{\partial w} = (z-y)x, \qquad \frac{\partial L}{\partial b} = z - y.$$

模型有激活时须再多乘一个激活导数；有多条路径时须把各路径贡献相加。单个乘法链是链式法则，整个网络高效组织这些局部导数，才是反向传播的计算过程。[P2]

5.5 例题：一隐藏单元的完整反传

$$h = s(wx+b), \quad o = vh + c, \quad L = \frac{1}{2}(o-y)^2, \quad \delta = o - y.$$

$$\frac{\partial L}{\partial v} = \delta h, \quad \frac{\partial L}{\partial c} = \delta, \quad \frac{\partial L}{\partial w} = \delta v h(1-h)x, \quad \frac{\partial L}{\partial b} = \delta v h(1-h).$$

取 x=2、y=0、w=b=c=0、v=2，则 h=1/2、o=1、δ=1，梯度依次为 dL/dv=1/2、dL/dc=1、dL/dw=1、dL/db=1/2。共享的 δ 与 δvh(1-h) 避免每个参数都重做整条展开。

一次同步梯度下降须先在同一旧参数点求出所有梯度，再更新所有参数；逐个参数更新后立即用新值重算后续梯度，是另一种迭代方案。

【在原题中用起来】回到 A8、A18、B5、B9、B21。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

【矩阵梯度的形状】对 $z = Wx + b$，若隐藏误差信号为列向量 δ_h，则 $\nabla_W L = \delta_h x^\top$。它是外积，形状与 W 相同；$\delta_h \odot x$ 不一定有定义，即使有也不是同一运算。

第 6 章　梯度下降：从局部方向到经典收敛条件

本章目标：区分下降方向、下降步长、全程收敛和参数唯一性。

6.1 负号不是神秘规则

一阶变化近似为 $\nabla L^\top \Delta\theta$。选择 $\Delta\theta = -\eta\nabla L$、$\eta$>0，会使这一项变为负数或零：

$$\nabla L^\top \Delta\theta = -\eta\|\nabla L\|_2^2 \le 0.$$

因此负梯度在梯度非零时是局部下降方向。但参数实际移动多少仍由 η 控制；非线性的余项可能抵消预测下降。这是一种简单的算法选择，而不是从“想下降”唯一推出的规则。

6.2 二次函数的精确二级结论

$$L(w) = \frac{a}{2}(w - w^*)^2 + c, \quad a > 0, \qquad w^+ - w^* = (1 - a\eta)(w - w^*).$$

$$L(w^+) - c = (1 - a\eta)^2[L(w) - c].$$

对全部非最优点严格下降，当且仅当 $|1 - a\eta| < 1$，即 $0 < \eta < 2/a$。η=1/a 一步到最优；η=2/a 时误差反号但损失不变；η>2/a 时误差幅度变大。必须保留“此类二次函数”和 a>0。

6.3 例题：不要把近似当精确

$\mathrm{L(w)=2(w+1)^2}$，$\mathrm{w_0}$=1，$\mathrm{L\prime(1)}$=8。η=1/4 时 $\mathrm{w_1}$=-1、损失为 0；η=3/4 时 $\mathrm{w_1}$=-5、损失从 8 上升到 32。两次都沿负导数方向，只是第二次走得太远。

6.4 光滑上界与 PL 条件 [P6]

一般函数很难精确展开，所以给一个可证明或假设成立的二次上界。令 g=∇F(θ)，光滑性上界与 PL 条件分别为：

$$F(\theta + v) \le F(\theta) + g^\top v + \frac{L_s}{2}\|v\|_2^2, \qquad \frac{1}{2}\|g\|_2^2 \ge \mu(F(\theta) - F^*).$$

第一条控制函数不能比二次模型增长得更快；第二条排除“损失仍远离最优、梯度却任意过小”的情况。它们不是所有神经网络损失自动具备的性质。

$$F(\theta - \eta g) - F(\theta) \le -\eta(1 - L_s\eta/2)\|g\|_2^2.$$

取 η=$1/L_s$，再使用 PL，可得 $a_{k+1} \le (1 - \mu/L_s)a_k$，其中 $a_k = F(\theta_k) - F^*$。归纳得到几何衰减。这里“线性收敛”指固定比例，不是一次函数图像。

6.5 三个必须记住的边界

一般可微函数在导数零处未必最优；满足 PL 时，驻点才一定达到全局最小值。函数值收敛不自动说明参数唯一；例如 F(w,b)=$\mathrm{w^2}$ 与 b 无关，有整条最优参数线。非负光滑上界提供的步长范围是充分范围，不总是具体函数的精确必要范围。

【在原题中用起来】回到 A6、A13、B8、B11、B15、B18。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

第 7 章　范数、积分与关键不等式

本章目标：知道测量的对象、测量的区域和测量标准，能选择恰当工具。

7.1 从长度到 q 范数

$$\|x\|_q = \left(\sum_{i=1}^{d} |x_i|^q\right)^{1/q} \quad (q \ge 1), \quad \|x\|_\infty = \max_i |x_i|.$$

绝对值不可省略。对 x=(-6,8)，一范数 14、二范数 10、无穷范数 8。不同标准不是在给同一物理长度算不同答案，而是在定义不同的大小度量。0<q<1 时同一形式一般不满足三角不等式，不能直接作为标准范数使用。

$$\|x\|_\infty \le \|x\|_2 \le \|x\|_1 \le \sqrt{d}\,\|x\|_2.$$

第一条因最大分量平方不超过平方和；第二条因绝对值和的平方含非负交叉项；最后一条来自 Cauchy。左到右的比较针对同一个有限维向量；不能把它无条件照搬到任意无限测度函数空间。

7.2 Cauchy–Schwarz 与三角不等式

$$|u^\top v| \le \|u\|_2 \|v\|_2, \qquad \|u+v\|_2 \le \|u\|_2 + \|v\|_2.$$

Cauchy 可由非负平方 $\|u - tv\|_2^2 \ge 0$ 证明：v≠0 时取 t=(u^Tv)/||v||²，化简得 (u^Tv)²≤||u||²||v||²；v=0 显然。三角不等式则由展开 ||u+v||² 并用 Cauchy 得到。论文题中给出工具，不意味着只需背名字；应知道它把内积换成了什么。

7.3 指示函数把条件变成数字

$$\mathbf{1}_A(t) = \begin{cases} 1, & t \in A, \\ 0, & t \notin A. \end{cases}$$

分类 0-1 损失是“预测错误”事件的指示量。求平均便是错误率。记号 A 可为集合；有些教材也直接把一个逻辑条件写在下标，表示条件成立取 1。

7.4 函数的整体大小

$$\|f\|_q = \left(\int_\Omega |f(t)|^q \, dt\right)^{1/q}.$$

向量按有限位置求和；函数按指定区域积分。积分不仅累加高度，还考虑区域宽度 / 体积。f 属于 Lq 表示相应积分有限；严格的函数空间按几乎处处相等识别函数。

例：$f(t) = 3\mathbf{1}_{[0,1/9]}(t)$，则一范数为 1/3，二范数为 1，无穷范数为 3。尖峰虽然很窄，最大幅值仍大。常数函数 1 在整个实数轴上有界，却不属于任何有限 q 的 Lq；换成有限区间则不同。

7.5 sup 与 max，及数学阅读中的边界

在 [0,1) 上 f(t)=t，上确界 sup 为 1，但没有最大值 max。上确界是最小上界，不要求被达到。一般 L∞ 理论使用本性上确界；本套连续分段或区间指示例子中，它与通常的幅值上界没有造成答案冲突。遇到单个零测度点的异常值，必须区分两者。

7.6 简单工具为何能证明经典结果

感知机界用 Cauchy 比较两种增长；低秩逼近用 Cauchy 把矩阵误差联系到某个方向；逐层误差用三角不等式；这些不是不同学科互不相干的公式，而是相同基础工具在不同对象上的作用。

【在原题中用起来】回到 A7、A11、A14、A16、A21、B4、B7、B11。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

【尖峰反例回收】对 $f_n(t) = n\mathbf{1}_{[0,n^{-2}]}(t)$，高度与支撑长度同时影响范数。请在 B7 的订正中恢复一般 $n^{1-2/q}$，再用不同 q 解释平均与最坏误差为什么可以分离。

第 8 章　经验风险、真实风险与有限类泛化

本章目标：能把“训练得好能否推广”转成带假设与概率的精确结论。

8.1 P 与 P_D 不是同一个对象

P 描述总体生成规律，通常未知；P_D 给每次观测分配 1/M 的概率质量，重复点累加质量。期望是按所采用分布加权的平均。

$$\mathbb{E}_{P_{\mathcal{D}}}[g(X,Y)] = \frac{1}{M}\sum_{i=1}^{M} g(x_i, y_i).$$

不同分布并不意味着对每个函数的期望都不同。例如观测 0、0、2 的经验分布与 0、2 各一半的总体分布不同，但对 g(z)=(z-1)^2，两个期望都为 1。

8.2 从损失平均到风险

$$\widehat{R}(h) = \frac{1}{M}\sum_{i=1}^{M} \ell(Y_i, h(X_i)), \qquad R(h) = \mathbb{E}_P[\ell(Y, h(X))].$$

帽号强调用样本得到的量。某训练模型经验风险为零，但未观测区域可能出错。风险差可以为正、为零或为负；不能仅凭分布不同断言两种风险必然不相等。

8.3 例题：同样拟合，不同外推

在 x=0、2 上标签分别为 0、2。函数族 $f_c(x) = x + cx(x-2)$ 对任何 c 都训练零误差，但在 x=1 预测 1-c。若总体标签 Y=X，且 X=1 概率 1/2、另外两点各 1/4，则总体平方风险为 c^2/2，唯一最优 c=0。

这个例子没有说更简单总是真实；若真实规律是另一条恰好也经过训练点的曲线，样本仍不能自行告诉我们选哪一个。模型结构提供的是额外假设。

8.4 已知工具：Hoeffding 与并集界 [P7]

若损失在 [0,1] 内、M 条样本独立同分布，对抽样前固定的 h，可用 $\Pr(|\widehat{R}(h) - R(h)| > \epsilon) \le 2e^{-2M\epsilon^2}$。本手册把这条集中引理作为已知工具，不要求在本阶段证明。

$$\Pr\left(\bigcup_{j=1}^{J} E_j\right) \le \sum_{j=1}^{J} \Pr(E_j).$$

并集界来自对每一次结果的指示量关系：至少一个坏事件发生的指示 ≤ 坏事件指示之和，再取期望。事件之间不要求独立；不要把并集界和独立乘法公式混为一谈。

8.5 有限类的一致界与 ERM [P8]

对抽样前固定、含 J 个模型的类，把各模型偏差过大的事件求并集。可得全部模型同时偏差受控的概率至少 1-δ，偏差半径可取

$$\epsilon = \sqrt{\frac{\log(2J/\delta)}{2M}}.$$

ERM 选经验风险最小模型 $\widehat{h}$。既然好事件对全体成立，也就对训练后选出的 $\widehat{h}$ 成立。插入两个经验风险：

$$R(\widehat{h}) \le \widehat{R}(\widehat{h}) + \epsilon \le \widehat{R}(h^*) + \epsilon \le R(h^*) + 2\epsilon.$$

这是有限类版本。不能直接把 J 换成无穷大后得到有意义的结论；无限模型类需要更细的复杂度工具。本章也没有为无限范围平方损失自动提供相同有界集中界。

8.6 样本、置信与复杂度的取舍

为保证超额风险上界不超过 a，充分要求 $M \geq (2/a^2)\log(2J/\delta)$。a 减半，上界要求的样本量变为四倍；J 进入对数；更高置信意味着更小 δ、更多样本。此式是充分上界，不是任何任务的最小样本量。

【在原题中用起来】回到 A5、A9、A17、B13、B19。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

第 9 章　二次型、可辨识性与最佳低秩近似

本章目标：分清函数系数、参数存储、样本约束与压缩后的表达损失。

9.1 为什么向量参数换成矩阵参数

$$\theta^\top x = \sum_i \theta_i x_i, \qquad x^\top A x = \sum_{i,j} A_{ij} x_i x_j.$$

左侧线性模型给每个输入分量一个系数；右侧二次模型给特征间乘积一个系数。这里是主动改换模型，不是同一个参数在训练中突然改变形状。二次型对 x 是非线性的，但对固定二次特征上的系数仍是线性的。

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a & u \\ v & c \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = ax_1^2 + (u+v)x_1x_2 + cx_2^2.$$

9.2 只有对称部分影响二次型

任何实矩阵分为对称部分 S=(A+A$^{\mathrm{T}}$)/2 和反对称部分 K=(A-A$^{\mathrm{T}}$)/2。标量 x$^{\mathrm{T}}$Kx 等于自身转置，但转置又等于-x$^{\mathrm{T}}$Kx，所以该标量为零。

$$x^\top A x = x^\top \frac{A + A^\top}{2} x.$$

一般齐次二次型独立系数为 d(d+1)/2，实际存储完整矩阵则有 d^2 个数。模型参数数、有效函数自由度、数据能识别的自由度，是三个不同概念。

9.3 例题：两条数据遗漏一个方向

模型 $ax_1^2+bx_1x_2+cx_2^2$，训练点 (1,0) 标签 2、(0,1) 标签 3，只推出 a=2,c=3，完全不限制 b。在 (1,1) 的预测是 5+b。若补一条 (1,1) 标签 7，才推出 b=2。

数据数量大也不自动提供独立信息；重复同一点不能约束从未变化过的交叉特征。反过来，参数多于样本也不等于一定泛化差，结构约束和实际数据规律仍然重要。

9.4 低秩因子如何减少成本

$$A = UV^\top, \quad U, V \in \mathbb{R}^{d\times r}, \qquad x^\top A x = (U^\top x)^\top (V^\top x).$$

中间是两个 r 维向量，最后是标量。存储从 d^2 变为 2dr；要真正减少存储需 2dr<d^2，即 r<d/2。仅说 r<d，并不总能保证因子存储量小于完整矩阵。计算还要包括最终内积，若题目只计投影须照约定解释。

因子存在缩放冗余：$(\alpha U)(\alpha^{-1}V)^\top = UV^\top$。更一般还可以存在基变换冗余，因此因子个数不直接等于可辨识自由度。

9.5 经典最佳低秩近似的二维证明 [P9]

令 D=diag(4,2)，用 B=uv$^{\mathrm{T}}$ 近似。存在单位向量 z 使 v$^{\mathrm{T}}$z=0，故 Bz=0。又 $||Dz||^2=16z_1^2+4z_2^2\geq 4$。由 ||Ez||≤||E||F 可得

$$\|D - B\|_F \geq \|(D-B)z\|_2 = \|Dz\|_2 \geq 2.$$

取 B=diag(4,0)，误差恰为 2，所以确为最佳秩一近似。证明分两半：给出所有候选都躲不过的下界，再给出达到下界的构造。只给一个候选误差为 2 还不能说明它最优。

这是 Eckart–Young 结果的特殊情形；一般矩阵版本需要更一般的奇异值与酉不变范数工具。本章不把局部例子冒充一般定理证明。

【在原题中用起来】回到 A19、B1、B12、B14、B20。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

第 10 章　深度与宽度：从折点计数到平方逼近

本章目标：让三角波构造同时解释表达效率和近似算法，而非只会背“深度有用”。

10.1 通用逼近回答“存在”，不回答“多省”

适当非线性网络能否逼近某个函数，是表达能力问题；需要多宽、多深、多少参数、能否训练到，是不同问题。Cybenko 的经典结果在特定 sigmoidal 条件下讨论连续函数逼近的存在性，不自动给出任意训练算法的成功保证。[P3]

10.2 两个 ReLU 构造一个三角波 [P4]

$$T(t) = 2\rho(t) - 4\rho(t - 1/2) \quad (0 \le t \le 1).$$

在左半段为 2t，右半段为 2-2t，输出始终落在 [0,1]。因此可以把输出再输入自身。一次复合变两次折叠：在每个旧线性段里，恰好一处达到 1/2，于是该段被一分为二。

归纳不变量是“每段一一覆盖 [0,1]”；相邻新段的斜率方向相反，不会重新合并。故 $\mathrm{T_k}$ 具有 2^k 个最大线性段与 $2^k - 1$ 个内部折点。

单隐藏层和式 $c + \sum_{j=1}^{N} a_j \rho(b_j t + d_j)$ 至多有 N 个内部折点。要精确表示 $\mathrm{T_k}$，必要条件 $\mathrm{N} \ge 2^k - 1$；k 个两 ReLU 模块只使用 2k 个单元。该必要下界不保证在不含显式线性通道的给定形式里必然可达。

10.3 三角波不只会制造复杂图像 [P5]

把 [0,1] 分成等长小区间，在每个区间用直线连接平方函数的两个端点。对于任意 [a,b]，割线与平方之差满足

$$L_{a,b}(t) - t^2 = (t - a)(b - t), \qquad \max_{[a,b]} (L_{a,b} - t^2) = \frac{(b-a)^2}{4}.$$

把区间对半细分，新插值在旧节点不变，在旧中点恰好降低前式给出的峰值；这正是一朵缩小的三角波。第 m 轮旧区间长度 $2^{-(m-1)}$，修正幅度因此为 4^{-m}。

$$P_{m-1} - P_m = 4^{-m} T_m, \qquad P_m(t) = t - \sum_{s=1}^{m} 4^{-s} T_s(t).$$

$$\| P_m - t^2 \|_{\infty,[0,1]} = 4^{-(m+1)}.$$

10.4 O(log(1/ϵ)) 从哪里来

每轮保留上一轮三角波值与插值值，只算一次新三角波，再减去对应修正。每轮固定宽度、固定数量的 ReLU 层，总量 O(m)。误差随 m 几何下降，所以误差要求 ϵ 对应 m 与 $\log(1/\epsilon)$ 同阶。

这里 O(m) 表示存在与 m 无关的常数 C，使资源不超过 Cm 加常数，并非精确等于 m。若把 $\mathrm{T_1}$ 到 $\mathrm{T_m}$ 每一个都从原始输入重新算，会发生重复工作；表达式短不代表实现已经高效。

10.5 两篇论文之间的连接

Telgarsky 构造让你看到复合可高效产生许多区段；Yarotsky 的平方构造进一步说明这些区段能服务于明确的逼近任务。前者的精确折点计数特例与后者的误差上界不同，不能混为一个定理；它们共享可复用的三角波模块。

【在原题中用起来】回到 A9、A20、B17。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

第 11 章　误差传播、残差与计算资源

本章目标：能解释逐层误差是否累积，以及为什么更少参数未必更低延迟。

11.1 先区分三种误差

逼近误差是模型族不能精确表示目标带来的差距；优化误差是算法没有达到该目标最优解；数值误差是有限精度、近似计算等使实际执行偏离理想运算。它们可能相互影响，但“网络有很多层”不意味着每层都有独立的训练目标误差。下面讨论的是理想状态与带局部扰动状态之差。

11.2 不同层可以放大，也可以抑制

$$\widetilde{x}_{k+1} = f_k(\widetilde{x}_k) + \varepsilon_k, \qquad x_{k+1} = f_k(x_k).$$

设误差 e_k=$||\widetilde{x}_k$-$x_k||$、局部误差 $\le\epsilon$，并且 f_k 满足系数 K 的 Lipschitz 条件。相减、三角不等式、代入条件三步得到

$$e_{k+1} \le K e_k + \varepsilon.$$

$$e_L \le K^L e_0 + \varepsilon \sum_{j=0}^{L-1} K^j.$$

$0 \le K < 1$ 时旧误差衰减，但不断新增的误差形成非零上界；$K=1$ 时上界线性积累；$K>1$ 时最坏上界可能指数增长。只有额外构造同向并饱和的误差，才能证明某个上界可达到。

11.3 例题：相同层数，不同稳定性

局部误差 ≤ 0.02、初始无误差，经过 4 层：$K=1/2$ 时上界 $0.02(1+1/2+1/4+1/8)=0.0375$；$K=2$ 时上界 $0.02(1+2+4+8)=0.30$。差异来自放缩，而不是 4 这个层数本身。

11.4 恒等残差的两个表达 [P10]

$$x_{k+1} = x_k + F_k(x_k) \quad \Rightarrow \quad x_L = x_0 + \sum_{k=0}^{L-1} F_k(x_k).$$

$$\frac{dx_L}{dx_0} = \prod_{k=0}^{L-1} (1 + F_k'(x_k)) = 1 + \sum_{k=0}^{L-1} F_k'(x_k) \frac{dx_k}{dx_0}.$$

右侧的每个 x_k 仍依赖 x_0，因此链式因子不可漏。这是恒等捷径且相加后也没有额外激活时的标量公式；加上其他结构需要重新推导。

若 F_k=-x，导数一层就归零；若 F_k=x，导数为 2^L。这不否定残差结构的价值，而是说明“带有直通项”与“总梯度无条件被保证”不是同一句话。

11.5 乘加、存储、深度、延迟分别计

$$C_{\text{deep}} = Ld^2, \qquad C_{\text{shallow}} = 2dN.$$

公式针对稠密投影且忽略其他运算的成本模型。比较前应写清输入输出维度，不可把“一个隐藏层宽 N”的成本直接写成 N^2。一次乘加在有些 FLOPs 约定中计 2 次浮点运算，题面若计一次 MAC 就应始终按同一单位。

实际延迟还受并行度、带宽、批量和实现影响。不同模型的成本若未在相同精度或功能要求下比较，也不能直接判定谁“总体更好”。

【在原题中用起来】回到 A15、A21、B11、B21。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

【两个稳定目标可以不同】每层乘 1/2 时，加性误差累积有界，但初值到末层的导数为 2^{-L}，仍可趋零。数值误差稳定不等于反向梯度不消失。

第 12 章　归纳偏置与论文式证明：把知识连成方法

本章目标：不靠结论堆叠，而能识别工具、条件、证明路线和结论边界。

12.1 为什么模型架构是一种假设

样本只是有限观察，而候选函数往往极多。选择卷积、时间递推、低秩或某类损失，是在规定模型优先寻找什么规律。与任务匹配时可提高数据 / 计算效率；假设错误时也可能排除真实规律。归纳偏置不是“偏见越少越好”或“结构越强越好”。

12.2 卷积：等变并非不变

对离散信号定义 $(K * x)(t) = \sum_s K(s)x(t-s)$，平移 $S_a x(t) = x(t-a)$，假设卷积核有限支撑以避免求和收敛问题。直接代入：

$$(K * S_a x)(t) = \sum_s K(s)x(t-s-a) = (K * x)(t-a) = S_a(K * x)(t).$$

输入平移，输出跟着平移，这叫等变；输出完全不变叫不变。池化或聚合有时将位置响应转成较不敏感的分类结果，但边界、步长和具体实现可能破坏精确性质。

12.3 时间方向与注意力

单向 RNN 的当前状态只使用此前状态与已提供输入。自回归 Transformer 通过掩码限制看未来位置，而编码器式注意力可以允许双向访问。[P11] 信息流的因果约束不等于已经识别统计学的因果关系。选择正确架构前，要先定义部署时到底能获得哪些信息。

12.4 六种可复用证明动作

动作	什么时候用	本套连接
展开并比较	需要从参数得到系数、从更新得到长度	二次型、感知机
保持不变量并归纳	每轮变复杂但结构重复	三角波覆盖与段数
插入中间量	两个难比较的对象之间有易比较桥梁	总体风险 →经验风险 →总体风险
构造反例	否定“所有”“必然”“自动保证”	驻点最优、残差稳定、零训练误差
先下界再构造	要证明某实现“最佳”	二维低秩近似
望远镜求和	相邻步骤差可大量消去	平方插值修正、残差展开

12.5 不要把论文全篇与一个特例混同

一次试卷能完整证明一个重要特例，已经能训练实质思维。但证明一个二维秩一例子不等于一般奇异值定理；证明平方函数不等于所有连续函数；有限 J 个模型的一致界不等于完整 VC 理论；精确折点计数不等于论文对所有浅网络的误差下界。阅读原文时要核对对象类别、范数、概率、近似精度与资源约定。

12.6 一个严谨答案的最小结构

先列对象与假设；写要证明的量；选工具并说明为何能用；给连续推理；处理零值、边界、非唯一等特例；最后用一句话解释结论不能推出什么。高分证明不是写得长，而是没有关键逻辑缺口。

【在原题中用起来】回到 A10、B9、B10。完成对应的【订正重建】和【同题改动】，不另追加独立计分试题。

二级结论索引：每条都须带条件

【复合顺序】$(f\circ g)(x)=f(g(x))$，一般不可交换。

【仿射折叠】$W_2(W_1x+b_1)+b_2=(W_2W_1)x+W_2b_1+b_2$；只说前向函数，不保证训练轨迹相同。

【内积与外积】$u^\top v$ 是数，$uv^\top$ 是矩阵，$u\odot v$ 是同形逐元素乘积。

【最优对象】min 返回值，argmin 返回达到最小值的参数集合；约束域可使集合为空或多解。

【经验期望】$\mathbb{E}_{P_{\mathcal{D}}}g=\frac{1}{M}\sum_i g(x_i,y_i)$；多重集重复项仍计数。

【范数顺序】有限维实向量有 $\|x\|_\infty\le\|x\|_2\le\|x\|_1$；不直接照搬到无限测度函数空间。

【二次损失步长】$L=\frac{a}{2}(w-w^*)^2+c, a>0$ 时，全部非最优点严格下降当且仅当 $0<\eta<2/a$。

【Sigmoid 导数】$s'=s(1-s)$；饱和因子趋小不说明参数最优。

【二次型有效参数】$x^\top Ax=x^\top(A+A^\top)x/2$；齐次二次系数数为 $d(d+1)/2$。

【感知机次数界】半径 $\le$R 且存在单位分离向量、正间隔 γ 时，零初值规则的更新数 $\le \mathrm{R}^2/\gamma^2$。

【误差级数】$e_L\le K^Le_0+\varepsilon\sum_{j=0}^{L-1}K^j$；K=1 须单独处理，上界不等于实际值。

【折点下界】N 个一维单隐藏层 ReLU 至多 N 个内部折点；三角波复合 k 次有 2^k-1 个。

【平方插值】二分网格的 $P_m=t-\sum_{s=1}^{m}4^{-s}T_s$，在 [0,1] 的最大误差 $4^{-(m+1)}$。

【PL 收敛】给定光滑上界与 PL 条件，η=1/Ls 时函数值差几何下降；不保证参数唯一。

【有限类风险】预先固定 J 个模型、独立同分布样本、损失在 [0,1]，统一偏差可取 $\sqrt{\log(2J/\delta)/(2M)}$。

【低秩输出误差】$|x^\top(A-B)x|\le\|A-B\|_F\|x\|_2^2$；需输入大小限制才能给一致精度。

【残差导数】$dx_L/dx_0=\prod_k(1+F_k'(x_k))$；捷径项可被抵消，也可被放大。

【平移结构】$C(Sx)=S(Cx)$ 是等变；$r(Sx)=r(x)$ 才是不变。全局求和可从重排等变得到不变。

论文来源与改编边界

以下只列本册使用的来源；原文编号与教学题编号不同。正文重新给出全部考查假设。基础恒等式不冒充论文专属定理，教学特例不替代原论文一般结论。

【P1】Freund, Y.; Schapire, R. E. (1999). Large Margin Classification Using the Perceptron Algorithm. Machine Learning 37, 277–296.

原文定位 / 教学范围：第 3.1 节、Theorem 1；复述 Block / Novikoff 的经典间隔界。

DOI / 标识：10.1023/A:1007662407062

【P2】Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. (1986). Learning representations by back-propagating errors. Nature 323, 533–536.

原文定位 / 教学范围：公式 (1)–(6)：前向计算、平方损失及链式求导。

DOI / 标识：10.1038/323533a0

【P3】Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2, 303–314.

原文定位 / 教学范围：通用逼近的存在性背景；本资料不声称证明其完整定理。

DOI / 标识：10.1007/BF02551274

【P4】Telgarsky, M. (2016). Benefits of depth in neural networks. COLT, PMLR 49, 1517–1539.

原文定位 / 教学范围：第 3.3 节、Lemmas 3.8–3.11：三角波及迭代构造；试卷只证精确表示的简化情形。

DOI / 标识：arXiv:1602.04485

【P5】Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. Neural Networks 94, 103–114.

原文定位 / 教学范围：第 3.1 节、Proposition 2：平方函数的深 ReLU 逼近；采用 arXiv v3 的命题编号。

DOI / 标识：10.1016/j.neunet.2017.07.002

【P6】Karimi, H.; Nutini, J.; Schmidt, M. (2016). Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak–Łojasiewicz Condition. ECML PKDD, 795–811.

原文定位 / 教学范围：第 2 节、Theorem 1：光滑性与 PL 条件下的梯度下降收敛。

DOI / 标识：10.1007/978-3-319-46128-1_50

【P7】Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association 58(301), 13–30.

原文定位 / 教学范围：有界独立变量的集中界；试卷将双侧均值界作为已知工具。

DOI / 标识：10.1080/01621459.1963.10500830

【P8】Vapnik, V. N.; Chervonenkis, A. Ya. (1971). On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. Theory of Probability and Its Applications 16(2), 264–280.

原文定位 / 教学范围：一致收敛背景；试卷只证有限模型类版本，不涉及完整 VC 维理论。

DOI / 标识：10.1137/1116025

【P9】Eckart, C.; Young, G. (1936). The Approximation of One Matrix by Another of Lower Rank. Psychometrika 1, 211–218.

原文定位 / 教学范围：最佳低秩近似经典结果；试卷只证二维对角矩阵的秩一特例。

DOI / 标识：10.1007/BF02288367

【P10】He, K.; Zhang, X.; Ren, S.; Sun, J. (2016). Identity Mappings in Deep Residual Networks. ECCV, 630–645.

原文定位 / 教学范围：第 2 节公式 (3)–(5)：恒等捷径的前向展开与反向传播；试卷采用标量可微特例。

DOI / 标识：10.1007/978-3-319-46493-0_38

【**P11**】Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems 30.

原文定位 / 教学范围：编码器与解码器注意力、位置与因果掩码背景，不在本卷考完整注意力推导。

DOI / 标识：arXiv:1706.03762

【**P12**】Goodfellow, I. J., et al. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems 27.

原文定位 / 教学范围：判别器学习“数据 / 生成”来源，不要求人工任务类别。

DOI / 标识：arXiv:1406.2661

范围说明

本书覆盖本次对话已讨论的数学语言与深度学习入门主题。卷内证明足以训练相关方法，但不能据此声称已经掌握完整 VC 理论、所有低秩近似定理或整套深度学习。是否真正内化，需用遮盖重建、同题变式和之后的无提示回访来判断。