磐石计划 01 · A 卷

答案解析与订正重建

本册配套 v3.0、21 题、150 分试卷。解析的作用是让你恢复方法，不只核对答案；先尝试，再逐级查看。客观题也必须写理由。

【使用顺序】分析只看第一步思路；详解按等号核对依据；点睛提取带条件的结论；遮住本册完成订正重建与同题改动，再翻末尾核验。

题号	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
答案	C	D	A	B	A	C	B	D

题号	**9**	**10**	**11**
答案	ABD	ACD	AC

12. $(2,1,5)$

13. $\frac{5}{2}(a-2)^2$；$\{2\}$

14. $1/\sqrt{3}$；1

15. $3/160$

【计分与掌握分开】首次独立得分、参考后订正得分分别记录，不相加。标准分值仍为 150 分；“看懂”“能复现”“能迁移”另记学习状态，不虚构统计性掌握概率。

建议把一次学习分成若干小段，卡住时查提示不视为失败。完成学习不等于必须在 120 分钟内读完所有解析；120 分钟仅用于之后的计时重做。

1. 模型是“输入＋参数”到输出的映射（5 分）

【答案】C

【分析】先不代数值，先数“可独立指定的对象”：输入有一个数，参数有两个数；输出仍是一个数。

【详解】$x \in \mathbb{R}$，$\theta = (w, b) \in \mathbb{R}^2$，所以完整自变量为 $(x, \theta) \in \mathbb{R} \times \mathbb{R}^2$，故选 C。A 把输入和输出都误写成二维；B 漏掉一个参数分量；D 把输出误作参数。固定 θ 后才得到单变量函数 $m(\cdot; \theta) : \mathbb{R} \to \mathbb{R}$。

【易错与修正】把参数空间当输出空间；把分号当运算符；把一个预测值与完整函数混为一谈。

【点睛】能在更换参数个数后重新写映射，并区分空间、元素、函数、函数值。

【订正重建】不用答案写出 $m : \mathcal{X} \times \Theta \to \mathcal{Y}$，并逐个解释 $m(2; (3, 1))$、$m(\cdot; (3, 1))$、Θ。

【同题改动】固定 $b = 1$，只训练 w。完整映射的输入空间怎样变化?

同题改动的核验见册末。回访：B1、A13、A19；讲义第 1 章。

2. 函数复合的方向（5 分）

【答案】D

【分析】把复合拆成两次输入输出交接，而不是从左边开始算。

【详解】$g(-1) = 2(-1) - 1 = -3$，再代入 f，得 $(f \circ g)(-1) = (-3)^2 = 9$，选 D。A 是中间结果；B 是倒序复合 $g(f(-1)) = 1$；C 也不符合原式的计算结果。

【易错与修正】把 $\circ$ 当乘法；把中间值 -3 当最终输出；交换复合次序。

【点睛】能写出内层函数输出，并把它作为一个整体代入外层。

【订正重建】写出 $(f \circ g)(x)$ 与 $(g \circ f)(x)$ 的完整表达式，不用一句“从右到左”代替实际代入。

【同题改动】求所有满足 $(f \circ g)(x) = (g \circ f)(x)$ 的实数 x。

同题改动的核验见册末。回访：B2、A18、A20；讲义第 1、3 章。

3. 转置、内积与输出类型（5 分）

【答案】A

【分析】先看形状：一行乘一列才能输出一个数；再逐项相乘并求和。

【详解】$\theta^\top x = 2\cdot 1 + (-1)\cdot 4 + 3\cdot(-1) = -5$，选 A。C 是 $\theta \odot x$，它还是三维向量；D 是参数本身。B 漏掉负号。$(1\times 3)(3\times 1) = (1\times 1)$。

【易错与修正】计算出了三个乘积就停下；认为写了转置只是外观变化，没有改变乘法形状。

【点睛】能不计算就判断输出是标量、向量还是矩阵，再正确运算。

【订正重建】并排算 $\theta^\top x$ 与 $\theta \odot x$，写出两者的形状，解释内积最后的求和。

【同题改动】把第三个参数改为 a，求模型输出为零的 a。

同题改动的核验见册末。回访：B3、B4、A19；讲义第 2 章。

4. 矩阵乘向量与逐元素激活（5 分）

【答案】B

【分析】逐层保留中间状态：线性乘法、加偏置、激活是三件事，不能合并着猜。

【详解】$Wx = (-2, 0)^\top$，$Wx + b = (-1, 1)^\top$，故 $\rho(Wx + b) = (0, 1)^\top$，选 B。A 停在激活之前；C 改变了坐标位置；D 把正分量也清零。W 一直是 2×3 矩阵，运算结果才是二维向量。

【易错与修正】把矩阵当作结果；先激活输入再做乘法；偏置维度没有跟随输出。

【点睛】能逐步写中间值，且能说明不是矩阵“变成”向量。

【订正重建】独立写出 $W, x, b, Wx, Wx + b, \rho(Wx + b)$ 的形状，并用每行一个内积解释输出。

【同题改动】只把 x_1 从 1 改成 3，其余不变，重算网络输出。

同题改动的核验见册末。回访：A12、A18、B14；讲义第 2、3 章。

5. 多重集、指示函数与经验平均（5 分）

【答案】A

【分析】先列四条记录，再决定每条记录是否出错；重复记录仍占一次权重。

【详解】四条记录的预测为 $0,1,1,1$，标签为 $0,1,1,0$，损失依次为 $0,0,0,1$。$\Lambda=\frac{1}{4}(0+0+0+1)=\frac{1}{4}$，选 A。B 是去重后错误计数；C、D 不符合逐条损失。**1** 把“预测错误”这一判断转成 0 或 1。

【易错与修正】把多重集去重；把分类器的输出当损失；忘记真实标签参与比较。

【点睛】能把逻辑判断、求和平均和经验期望三种写法互相转换。

【订正重建】写出预测列、标签列、0-1 损失列，以及经验分布下的期望式。

【同题改动】只把第二条 $(1,1)$ 改成 $(1,0)$，求新经验风险。

同题改动的核验见册末。回访：A17、B13、B19；讲义第 4、8 章。

6. 最小值与最小化点集合（5 分）

【答案】C

【分析】用“平方非负＋什么时候取零”，比直接求导更短，也更容易得到全部最优点。

【详解】$L(t)\geq 0$，且 $L(t)=0$ 当且仅当 $t^2=1$，所以最小值为 0，$\operatorname{argmin} L=\{-1,1\}$，选 C。A 中的 0 是驻点但不是最小点；B 漏掉另一最小点；D 包含非最优参数。

【易错与修正】把 min 的函数值写进 argmin；只找到一个最小点；把导数为零当充分条件。

【点睛】能区分最小值、最优参数、驻点，并保留参数约束。

【订正重建】分别写出 $\min L$、$\operatorname{argmin} L$、$L'(0)$，说明三者回答不同的问题。

【同题改动】参数只允许在 $[0,2]$ 中，再求最小值和最小化点集合。

同题改动的核验见册末。回访：B15、B18；讲义第 1、6 章。

7. 同一向量的三种大小（5 分）

【答案】B

【分析】同一组坐标，分别使用绝对值求和和最大绝对值；两种量不能混用。

【详解】$\|x\|_1 = 3+4 = 7$，$\|x\|_\infty = 4$，差为 3，选 B。二范数为 5，不是题目所求的差；忽略绝对值会造成负数抵消。

【易错与修正】把二范数 5 当所有范数；把下标当坐标；对负分量不取绝对值。

【点睛】能计算三种范数，并用缩放性质解释变式。

【订正重建】从一般式 $\|x\|_q = (\sum_i |x_i|^q)^{1/q}$ 推出 $q = 1, 2$ 的两式，再解释无穷范数。

【同题改动】不逐项重算，求 $\|-2x\|_1, \|-2x\|_2, \|-2x\|_\infty$。

同题改动的核验见册末。回访：B4、B11、B20；讲义第 7 章。

8. 从定义推出激活函数恒等式（5 分）

【答案】D

【分析】找成对出现的正负输入，而不是把四个分式通分到一起。

【详解】$s(-t) = \frac{1}{1+e^t} = 1 - s(t)$；由分子变号、分母不变得 $u(-t) = -u(t)$。总和为 1，选 D。A 忘了 Sigmoid 那一对的和，B 把两对都当成和为 1，C 仍依赖输入。

【易错与修正】只在 t=0 验证就宣称恒等；把 Sigmoid 误判为奇函数。

【点睛】能从定义恢复对称关系，而不是背某个输入处的数值。

【订正重建】不靠图像，分别从指数定义证明 $s(-t) = 1 - s(t)$ 和 $u(-t) = -u(t)$。

【同题改动】证明 $u(t) = 2s(2t) - 1$，并用它求 $u(0)$。

同题改动的核验见册末。回访：A18、B2；讲义第 3、5 章。

9. 逼近、训练与泛化的三道门（6 分）

【答案】ABD

【分析】区分“存在一个好函数”“算法找到它”“新数据也好”三个逻辑命题。

【详解】A、B 分别符合监督与无监督学习的题给情境；D 正确。C 错误：通用逼近的结论是适当网络族内的存在性，不给任意初值的全局优化和泛化保证，故选 ABD。[P3] 例如常用非线性网络能表示某目标，也可能因不合适步长而训练失败。

【易错与修正】认为表达能力越强就自动训练越好；把未提供人工标签误解为不能学习任何东西。

【点睛】能给出训练拟合不确定外推的具体反例。

【订正重建】用三句话分别解释逼近、优化、泛化，并各写一个不能由前者推出后者的理由。

【同题改动】只有训练点 $(0,0),(1,1)$，比较 $f(x)=x$ 和 $g(x)=x^2$。零训练误差能否确定未来的规律?

同题改动的核验见册末。回访：A17、B16、B19；讲义第 4、8、10 章。

论文关联：[P3]。

10. 结构假设不是无条件保证（6 分）

【答案】ACD

【分析】把每个架构名称翻译成一个具体的信息访问或参数共享约束。

【详解】选 ACD。卷积在忽略边界等条件下的响应随输入平移而平移，称等变，不是每一输出坐标都不变。Transformer 是否遮住未来取决于掩码，编码器不必采用自回归限制；B 错误。合理偏置缩小搜索空间，不匹配任务时也会排除真实规律。[P11]

【易错与修正】等变与不变混用；认为任何注意力都有同一掩码；把序列信息方向当因果发现。

【点睛】能用等式而不是口号区分等变、不变与因果访问限制。

【订正重建】分别用 $F(Sx)=SF(x)$ 与 $F(Sx)=F(x)$ 写出等变和不变；解释 S 是什么。

【同题改动】若向量映射 C 满足 $C(Sx)=S(Cx)$，S 只是重排坐标，证明 $\sum_i (C(Sx))_i=\sum_i (Cx)_i$。

同题改动的核验见册末。回访：B10、B21；讲义第 12 章。

论文关联：[P11]。

11. 有界不等于全域可积（6 分）

【答案】AC

【分析】先看函数非零的区域，再算高度的 q 次方乘以区域长度。

【详解】$\int_{\mathbb{R}} |g|^q dt = 2$，故 $\|g\|_q = 2^{1/q}$。A、C 正确；B 漏了开方，二范数应为 $\sqrt{2}$。D 错误，$\int_{\mathbb{R}} 1^2 dt = \infty$。函数有界只能控制高度，不能保证整个无限区间上的累积有限。选 AC。

【易错与修正】忘记积分区域；把积分值当范数；把有界与可积当同义词。

【点睛】能解释范数有限所依赖的区域和指数，而不只计算结果。

【订正重建】写出 $g \in L^q(\mathbb{R})$ 的有限性条件，解释 dx 表示按长度或体积累计，而不是按概率取平均。

【同题改动】把 h 的定义域改为 $[0,2]$，再求它在该区域上的二范数。

同题改动的核验见册末。回访：A14、B7、B17；讲义第 7 章。

12. 仿射层折叠与偏置传播（5 分）

【答案】$(2,1,5)$

【分析】把第一层完整代入第二层，权重乘在输入上，也要乘在第一层偏置上。

【详解】$h = (x_1 + x_2 + 1, x_2)^\top$，$y = 2h_1 - h_2 + 3 = 2x_1 + x_2 + 5$。故 $(a,b,c) = (2,1,5)$。一般地 $W_2(W_1 x + b_1) + b_2 = (W_2 W_1)x + W_2 b_1 + b_2$。分值为 2、1、2 分。

【易错与修正】只乘权重不乘偏置；把两个偏置直接相加；认为可折叠意味着训练轨迹相同。

【点睛】能推导折叠公式，并用分段斜率说明非线性的作用。

【订正重建】从代入式重建一般合并公式，标记所有矩阵与偏置的形状。

【同题改动】在 h 之后加 ReLU，沿输入 $x = (t,0)^\top$ 写出 y，判断能否在全域折叠为仿射函数。

同题改动的核验见册末。回访：B14、B16、A18；讲义第 2 章。

13. 从记录构造损失，而非背总公式（5 分）

【答案】 $\frac{5}{2}(a-2)^2$；$\{2\}$

【分析】 先独立写两个预测及其误差，最后才平均；平方和非负可直接定位最优点。

【详解】第一条预测为 $a+1$，误差 $a-2$；第二条预测为 $2a+1$，误差 $2a-4$。因此 $\Lambda(a)=\frac{1}{2}[(a-2)^2+(2a-4)^2]=\frac{5}{2}(a-2)^2$，最小化点集合为 $\{2\}$。两空 3、2 分。

【易错与修正】 不减真实标签；把预测加起来再平方；漏除样本数；把 argmin 写成最小损失 0。

【点睛】 能从一条记录开始搭建平均损失，再用梯度更新。

【订正重建】 以 $\ell_y(y')$、$\lambda_{(x,y)}(a)$、$\Lambda(a)$ 三层记号重写，再求 $\Lambda'(a)$。

【同题改动】 从 $a_0=0$ 用学习率 $1/5$ 做一步梯度下降，求新参数和损失。

同题改动的核验见册末。回访：B5、B8、B18；讲义第 4、6 章。

14. L2 误差与上确界不是同一对象（5 分）

【答案】 $1/\sqrt{3}$；1

【分析】 平方积分测整体大小，上确界测最高幅值；后者不要求在定义域内取到。

【详解】 $\|f\|_2=(\int_0^1 t^2dt)^{1/2}=1/\sqrt{3}$。对任意 t 有 $|f(t)|<1$，但令 $t\to 1^-$ 可任意接近 1，因此 $\sup|f|=1$；没有最大值。两空 3、2 分。

【易错与修正】 把“没取到 1”写成上确界不存在；积分忘平方或忘开方；忽略端点定义。

【点睛】 能区分 sup 与 max，并说明单点改动对积分和逐点值的不同影响。

【订正重建】 分别证明 1 是上界且任何小于 1 的数不是上界；再独立计算积分。

【同题改动】 只把 $f(1)$ 改成 1。二范数、上确界和最大值分别如何变化?

同题改动的核验见册末。回访：B7、B17；讲义第 7 章。

15. 局部误差如何变为几何级数（5 分）

【答案】$3/160$

【分析】先连续展开四步，不提前猜是四倍；每条早期误差还会被后续系数缩放。

【详解】$e_1 \le 1/100$，$e_2 \le 3/200$，$e_3 \le 7/400$，$e_4 \le 15/800 = 3/160$。一般地上界为 $\frac{1}{100}\sum_{j=0}^{L-1}(1/2)^j$。若各步都取等号，所得值可达到，故不是随意放大的界。

【易错与修正】直接乘层数；把不等式写成每次实际必然相等；套级数时漏掉初始项 1。

【点睛】能从递推写出求和，并把可达上界与实际值区分。

【订正重建】列出每条误差在第四层的权重，并解释为什么最晚一条误差权重为 1。

【同题改动】把系数 $1/2$ 改为 1，保持局部误差和初值不变，求四步可达上界。

同题改动的核验见册末。回访：A21、B21；讲义第 11 章。

16. 从一次更新走到感知机间隔界（12 分）

【答案】分问结果及证明见详解。

【分析】单个坐标可能反复改变，难以直接控制。题目给了正确分离方向 u，于是比较沿 u 的进展与总长度：一个至少按 k 增长，另一个至多按根号 k 增长。

【详解】以下逐问恢复推理；括号内为分步给分。

(1) 2 分。旧得分为 0，触发更新。$w^{(1)} = 0 + 1 \cdot (1,-1)^\top = (1,-1)^\top$，新得分 $1 \cdot [(1,-1)(1,-1)^\top] = 2$。

评分：新参数 1 分，新得分 1 分。

(2) 5 分。每次更新均满足 $u^\top w^{(k+1)} = u^\top w^{(k)} + y_i u^\top x_i \geq u^\top w^{(k)} + \gamma$。从零向量开始归纳，得到 $u^\top w^{(k)} \geq k\gamma$。

$$\|w^{(k+1)}\|_2^2 = \|w^{(k)}\|_2^2 + 2y_i (w^{(k)})^\top x_i + \|x_i\|_2^2 \leq \|w^{(k)}\|_2^2 + R^2.$$

中间交叉项非正正是触发更新的条件；从 $\|w^{(0)}\|^2 = 0$ 得 $\|w^{(k)}\|_2^2 \leq kR^2$。

评分：投影单步与累加 2 分；平方展开 1 分；用更新触发条件 1 分；范数平方累加 1 分。

(3) 3 分。

$$k\gamma \leq u^\top w^{(k)} \leq \|u\|_2 \|w^{(k)}\|_2 \leq R\sqrt{k}.$$

若 $k = 0$ 结论已成立；若 $k > 0$，除以 $\sqrt{k}$ 后平方，得 $k \leq R^2/\gamma^2$。更新数为整数，还可取该数的向下取整。当 $R = 2, \gamma = 1/2$ 时，上界为 16。

评分：合并两界并使用 Cauchy 不等式 1 分；处理零次更新并推出一般次数界 1 分；代入得到 16 得 1 分。

(4) 2 分。不能。假设约束的是已给定训练样本的几何间隔，尚无新样本分布、覆盖或泛化条件。证明限制更新次数，不意味着真实分布错误率为零。

评分：否定结论 1 分；指出训练几何条件与新样本风险的区别 1 分。

【易错与修正】把检查次数当更新次数；没有在展开范数时使用触发条件；直接除以零次更新数；把训练收敛当总体保证。

【点睛】“正确方向投影至少按 k 增长，总长度至多按 $\sqrt{k}$ 增长”不可能永远同时成立。这是收敛界的核心对比。

【订正重建】遮住解答，只保留两个空框：$u^\top w^{(k)} \geq \square$，$\|w^{(k)}\|_2^2 \leq \square$。独立补齐证明，最后写出所用假设。

【同题改动】每次更新改为 $w^+ = w + \eta y_i x_i$，其中固定 $\eta > 0$。重新比较两种增长，判断次数界是否改变。

同题改动的核验见册末。回访：B16、B20；讲义第 3、7 章。

论文关联：[P1]。经典来源：[P1] 第 3.1 节 Theorem 1。采用“非正间隔即更新”的边界约定，证明的是同型更新次数界。

17. 经验风险与真实风险分别用谁的权重（12 分）

【答案】分问结果及证明见详解。

【分析】模型预测不变，只换权重：同一错误表按观测频数加权是训练风险，按总体概率加权是真实风险。

【详解】以下逐问恢复推理；括号内为分步给分。

(1) 2 分。经验期望为 $(0+1+1+1)/4 = 3/4$；总体期望为 $0\cdot 1/4 + 1\cdot 1/4 + 2\cdot 1/2 = 5/4$。

评分：两个期望各 1 分。

(2) 6 分。

分类器	经验风险	总体风险	总体中错误的事件
h_A	0	$1/2$	$X = 2$
h_B	0	0	无
h_C	$3/4$	$1/4$	$X = 1$

评分：六个风险数值各 1 分。重复观测不能去重，也不能拿总体概率替代经验权重。

(3) 4 分。经验最小化模型集合为 $\{h_A, h_B\}$，总体最小化集合为 $\{h_B\}$。训练集中没有 $X = 2$，但该事件总体概率为 $1/2$。h_A 在此出错，故训练零风险不足以保证总体零风险。

评分：两个集合各 1 分；指出未覆盖事件及其权重 1 分；由反例得出逻辑结论 1 分。

【易错与修正】混用两套概率；忘记训练中完全没观察到 x=2；最小化点不唯一却只写一个模型。

【点睛】经验平均与总体平均可能排序不同；观测点上完全一致的模型，仍可能在未观测事件上产生不同预测。

【订正重建】独立做出三模型、三输入的预测表；不看解答写两个期望和两类最优集合。

【同题改动】向原训练多重集补入一条 $(2, 0)$，三个模型的经验风险分别是多少?

同题改动的核验见册末。回访：B13、B19；讲义第 4、8 章。

18. 链式法则怎样组织成反向传播（12 分）

【答案】分问结果及证明见详解。

【分析】不要展开成一个巨大的式子。先求损失对输出的导数，再沿每条边传回去；不同入边共用同一个隐藏误差信号。

【详解】以下逐问恢复推理；括号内为分步给分。

（1）2 分。

$$s'(t)=\frac{e^{-t}}{(1+e^{-t})^2}=\frac{1}{1+e^{-t}}\left(1-\frac{1}{1+e^{-t}}\right)=s(t)(1-s(t)).$$

评分：正确求导 1 分；改写成 s 的形式 1 分。

（2）5 分。

$$\frac{\partial L}{\partial v_j}=\delta h_j,\qquad \frac{\partial L}{\partial c}=\delta,$$

$$\delta_j=\delta v_j h_j(1-h_j),\qquad \frac{\partial L}{\partial b_j}=\delta_j,\qquad \frac{\partial L}{\partial W_{ji}}=\delta_j x_i.$$

δ 是输出误差信号；同一隐藏单元的 δ_j 可供该单元全部入边权重和偏置复用。前向算出的 h_j 也可复用，不必重复从头展开损失。

评分：四类导数各 1 分；明确公共误差信号的复用 1 分。只写最终矩阵公式但含义和维度正确同样给分。

（3）3 分。旧参数下 $z=(0,0)^\top$、$h=(1/2,1/2)^\top$，$o=0$、$\delta=-1$；因此 $\delta_1=-1/4,\delta_2=1/4$。

$$\nabla_W L=\begin{pmatrix}-1/4 & 1/4\\ 1/4 & -1/4\end{pmatrix},\quad W^+=\begin{pmatrix}1/40 & -1/40\\ -1/40 & 1/40\end{pmatrix},\quad c^+=\frac{1}{10}.$$

评分：求出旧点误差信号 1 分；W 更新 1 分；c 更新 1 分。先更新 v 再据其求 W 梯度，不属于本题规定的同时更新。

（4）2 分。当 $z_j\to\pm\infty$，$h_j(1-h_j)\to 0$。其余乘数有界时，该隐藏单元的入边梯度趋小；这可能只是激活饱和，不能推出全局或局部最优。有限 z_j 时 sigmoid 导数仍严格为正，不宜直接写成恒等于零。

评分：饱和因子及其极限 1 分；否定最优保证并解释 1 分。

【易错与修正】漏 Sigmoid 导数；漏输入 xi；将矩阵梯度写成向量；先改 v 再算 W 梯度导致不再是同一点的同时更新。

【点睛】反向传播是对复合计算图应用链式法则，并复用共享的中间导数；它计算梯度，不等于优化器本身。

【订正重建】画五个符号框 L,o,h_j,z_j,W_{ji}，每条箭头旁写局部导数。再把 $\nabla_W L$ 写成外积并检查形状。

【同题改动】仅把输出损失由 $\frac{1}{2}(o-y)^2$ 改为 $(o-y)^2$。所有梯度和等效学习率怎样变化?

同题改动的核验见册末。回访：B5、B18、B21；讲义第 5 章。

论文关联：[P2]。经典来源：[P2] 公式 (1)–(6) 的两隐藏单元特例；输出层改为线性，保留相同链式计算原理。

19. 函数自由度与参数冗余（12 分）

【答案】分问结果及证明见详解。

【分析】先展开二次型，观察 u 和 v 以什么组合出现。样本能约束函数系数，却永远看不到对输出无贡献的方向。

【详解】以下逐问恢复推理；括号内为分步给分。

(1) 3 分。

$$x^\top Ax = ax_1^2 + (u+v)x_1x_2 + cx_2^2.$$

有效的独立组合为 $a, u+v, c$。四个可存储元素不是四个独立函数系数。

评分：正确展开 2 分；指出三组合 1 分。

(2) 3 分。两条约束分别给出 $a=1, c=1$，u, v 任意，故 $A = \begin{pmatrix} 1 & u \\ v & 1 \end{pmatrix}$。在新点的输出为 $2+u+v$，可为任意实数，不唯一。

评分：全部矩阵形式 2 分；新点不唯一及表达式 1 分。

(3) 3 分。新增样本要求 $u+v=0$, 全部参数为 $A = \begin{pmatrix} 1 & t \\ -t & 1 \end{pmatrix}$。它们在所有输入上给出同一函数 $x_1^2+x_2^2$, 但参数矩阵不唯一。若又对称，则 $t=0$，唯一为 I_2。

评分：函数唯一 1 分；矩阵不唯一并给参数式 1 分；对称后唯一 1 分。

(4) 3 分。置 $S=(A+A^\top)/2$、$K=(A-A^\top)/2$，则 $K^\top=-K$。标量 $r=x^\top Kx$ 等于其转置，同时 $r=x^\top K^\top x=-r$，故 $r=0$。于是 $x^\top Ax = x^\top Sx$。

$$d + \frac{d(d-1)}{2} = \frac{d(d+1)}{2}.$$

评分：反对称部分贡献为零的证明 2 分；独立系数计数 1 分。

【易错与修正】把 4 个存储数当 4 个可辨识函数系数；认为函数相同必然参数相同；漏掉矩阵反对称部分。

【点睛】参数不可辨识与预测函数不可辨识应分开。二次型的反对称部分在任何输入上都不可由输出识别。

【订正重建】用三行区分“最初两样本”“加入第三样本”“再加对称约束”，分别记录函数和参数是否唯一。

【同题改动】第三条标签由 2 改为 4，求所有零损失矩阵；再求其中唯一的对称矩阵。

同题改动的核验见册末。回访：B1、B12、B20；讲义第 2、9 章。

20. 从复合到指数增长的折点（12 分）

【答案】分问结果及证明见详解。

【分析】归纳时不仅猜段数，还要保留可继续使用的性质：每段单调覆盖 [0,1]。这保证下一轮恰好切成两段。

【详解】以下逐问恢复推理；括号内为分步给分。

(1) 4 分。

$$T(t)=\begin{cases}2t, & 0\le t\le 1/2,\\ 2-2t, & 1/2\le t\le 1,\end{cases}$$

$$T_2(t)=\begin{cases}4t, & 0\le t\le 1/4,\\ 2-4t, & 1/4\le t\le 1/2,\\ 4t-2, & 1/2\le t\le 3/4,\\ 4-4t, & 3/4\le t\le 1.\end{cases}$$

图象由折线依次连接 $(0,0),(1/4,1),(1/2,0),(3/4,1),(1,0)$。端点的重复归属无妨，因为相邻公式取值相同。

评分：T 分段式 1 分；T_2 四段式 2 分；正确图象 1 分。

(2) 5 分。猜想段数 2^k、内部折点数 2^k-1。初始 $k=1$ 成立，且各段严格单调地覆盖 $[0,1]$。假设第 k 次成立，每个旧段内恰有一点使 $T_k=1/2$；再作用 T，这个旧段分成两段，各段仍覆盖 [0,1]，于是段数加倍。

新段的方向交替为上升、下降；旧段公共端点原本输出 0 或 1，再作用 T 均变成局部低点或高点。相邻段斜率一正一负，不能合并。由归纳得所猜结论。

评分：正确猜想 1 分；初值和归纳假设 1 分；每段分为两段且覆盖不变 2 分；排除合并并完成归纳 1 分。

(3) 3 分。每项 $\rho(b_jt+d_j)$ 至多带来一个内部折点，和式只能在这些位置改变斜率，所以折点至多 N。T_5 有 31 个内部折点，故必要条件 $N\ge 31$。五个 T 模块共使用 10 个 ReLU 单元。

这是该结构的必要下界，不是声明 N=31 一定足够；也不是所有任务或所有近似误差水平下深网都优于浅网。

评分：折点数推出下界 1 分；10 个单元 1 分；正确限定结论 1 分。

【易错与修正】只统计新增折点却没排除旧折点消失；把必要下界当最小可达宽度；把个别构造推广为所有任务。

【点睛】单隐藏层的一维折点数受单元数线性限制；适当的复合可让折点数随深度指数增长。

【订正重建】不看原图重画 T 和 T2；用“归纳性质—切分—不能合并”三段完成证明，并给出浅层折点上界。

【同题改动】将 T 复合 6 次，给出线性段数、内部折点数、单隐藏层必要宽度下界，以及模块实现的 ReLU 总数。

同题改动的核验见册末。回访：B17；讲义第 3、10 章。

论文关联：[P4]。经典来源：[P4] 第 3.3 节三角波构造。考查精确表示的简化分离，不声称证明论文的完整逼近下界。

21. 误差传播与计算成本的联合推理（12 分）

【答案】分问结果及证明见详解。

【分析】将实际状态与理想状态相减，先得到一个可以反复使用的单步不等式，再用级数汇总；成本则从矩阵形状另外计数。

【详解】以下逐问恢复推理；括号内为分步给分。

（1）3 分。

$$e_{k+1} = \|f_k(\widetilde{x}_k) - f_k(x_k) + \varepsilon_k\| \le \|f_k(\widetilde{x}_k) - f_k(x_k)\| + \|\varepsilon_k\| \le K e_k + \varepsilon.$$

评分：相减取范数 1 分；三角不等式 1 分；用 Lipschitz 与局部误差条件 1 分。

（2）5 分。

$$e_L \le \varepsilon \sum_{j=0}^{L-1} K^j = \begin{cases} \varepsilon(1-K^L)/(1-K), & K \ne 1, \\ L\varepsilon, & K = 1. \end{cases}$$

当 $0 \le K < 1$，上界不超过 $\varepsilon/(1-K)$；当 K=1，上界线性增长；当 K>1，上界按指数阶增长。上界不是每次运行的实际值，误差可更小、抵消，或局部放大系数没有同时达到最坏值。

评分：求和式 1 分；闭式及 K=1 特例 1 分；三种增长情形 2 分；区分上界与实际 1 分。

（3）4 分。深结构为 $Ld^2 = 8 \times 64^2 = 32768$ 次乘加；浅结构为 $2dN = 2 \times 64 \times 512 = 65536$ 次。后者是前者的两倍。

这只能比较题给成本模型。运行延迟还依赖顺序依赖、并行吞吐和数据搬运；题目也未规定两模型有相同表达能力或相同任务精度。

评分：两项成本各 1 分；说明延迟不能单独决定 1 分；说明精度未受保证 1 分。

【易错与修正】把上界当实际轨迹；K=1 仍除以 K-1；把 MAC 次数、FLOPs、延迟和任务准确率互当同义词。

【点睛】稳定性由扰动放大系数与新增误差共同决定；总运算量、顺序计算长度、运行延迟和任务误差不是同一指标。

【订正重建】从状态定义推导误差递推；分三类写几何级数；单独写深浅结构两式，并注明计数单位为乘加。

【同题改动】初始误差改为 $e_0 > 0$，推导一般上界；再令 $K = 1/2, \varepsilon = 1/100, e_0 = 1/10$，求第四层上界。

同题改动的核验见册末。回访：B20、B21；讲义第 7、11 章。

同题改动核验区：做完重建后再看

这些是原 21 题的订正环节，不是追加 21 道独立计分题。核验后遮住文字，用自己的符号与语言再解释一次。

【1 题】可调参数只有一个实数，完整映射是 $m:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$；$m(2;(3,1))=7$ 是值，$m(\cdot;(3,1))$ 是 $x\mapsto 3x+1$ 这个函数。

【2 题】$(2x-1)^2=2x^2-1$，化为 $2(x-1)^2=0$，故仅 $x=1$。两个函数在一个点相等不意味着它们恒等。

【3 题】$2-4-a=0$，得 $a=-2$。内积是 -5；原逐元素乘积是 $(2,-4,-3)^\top$。

【4 题】新 $Wx=(0,4)^\top$，加偏置得 $(1,5)^\top$，两分量均非负，输出 $(1,5)^\top$。

【5 题】新损失为 $0,1,0,1$，故风险为 $1/2$；同一输入可因观测标签不同而出现不同损失。

【6 题】仍有最小值 0，但最小化点集合变为 $\{1\}$；$L'(t)=4t(t^2-1)$，故 $L'(0)=0$ 而 $L(0)=1$。

【7 题】由齐次性 $\|cx\|=|c|\|x\|$，三者依次为 14、10、8。

【8 题】$2s(2t)-1=(1-e^{-2t})/(1+e^{-2t})=(e^t-e^{-t})/(e^t+e^{-t})$，故 $u(0)=0$。

【9 题】两函数在训练点相同，但在 $x=1/2$ 分别为 $1/2,1/4$；仅这两点不能区分。

【10 题】重排不会改变求和，故后接全局求和可将这个等变表示变成不变的标量。具体卷积的边界、步长等仍须另行检查。

【11 题】范数变为 $\sqrt{\int_0^2 1dt}=\sqrt{2}$；是否可积依赖函数和积分域。

【12 题】$y=2\rho(t+1)+3$；在 $t<-1$ 与 $t>-1$ 的斜率分别为 0 和 2，不能是同一个仿射函数。

【13 题】$\Lambda'(a)=5(a-2)$，$a_1=0-(1/5)(-10)=2$，损失为 0；这种一步到达来自本题二次结构，并非任意网络都如此。

【14 题】单点对积分无贡献，二范数仍 $1/\sqrt{3}$，上确界仍 1；这次最大值 1 在 t=1 取得。

【15 题】四条误差都不衰减，$e_4\le 4/100=1/25$，等号递推可达。

【16 题】投影至少为 $k\eta\gamma$，长度平方至多为 $k\eta^2R^2$；合并后 $k\eta\gamma\le\eta R\sqrt{k}$，正学习率消去，次数界仍为 R^2/γ^2。这不等于所有优化算法都对学习率不敏感。

【17 题】新数据共 5 条，$\Lambda(h_A)=1/5$，$\Lambda(h_B)=0$，$\Lambda(h_C)=3/5$；此时经验最优模型唯一为 h_B。这只解决题设支持集例子，不代表任意新增样本都消除泛化问题。

【18 题】损失整体乘 2，梯度整体乘 2；将新学习率改为旧学习率的一半，可在同一旧参数点得到相同更新。原 $\nabla_W L=\delta_h x^\top$，为 2×2；这与 Hadamard 乘积不同。

【19 题】$a=c=1,u+v=2$，矩阵为 $\begin{pmatrix}1 & t\\ 2-t & 1\end{pmatrix}$。对称要求 t=1，得全 1 矩阵；唯一函数是 $(x_1+x_2)^2$。

【20 题】依次为 64、63、$N\ge 63$、12。下界只证明至少需要这么多，不证明 N=63 足够。

【21 题】$e_L\le K^Le_0+\varepsilon\sum_{j=0}^{L-1}K^j$。第四层为 $1/160+3/160=1/40$。局部误差不停止时，即使 K<1，界也未必趋零。

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

【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