磐石计划 01 · B 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	B	C	D	A	B	D	C	A

题号	9	10	11
答案	ABD	ACD	BD

12. 6；$(3,3)$

13. $1/3$；1

14. 2×3；23

15. 1；$\{1\}$

【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. 不同参数可能是同一个函数（5 分）

【答案】B

【分析】模型族的元素是函数，不是参数标签；要比较函数在整个给定输入集合的值。

【详解】笛卡尔积有 $2\times3=6$ 个元素。三组参数的 a+b 分别是 1、1、2，因此得到函数 x、x、2x，只有两个不同函数，选 B。参数不同并不自动使函数不同；这里在 x=1 即可区分两种函数。

【易错与修正】把参数组数当函数数；把笛卡尔积当并集；只在 x=0 比较导致所有函数看似一样。

【点睛】能解释参数到函数的多对一映射，并检查定义域。

【订正重建】为三组参数各写一列在 x=0、1 的函数值，据此合并相同函数。

【同题改动】输入空间仅保留 $\mathcal{X}=\{0\}$，参数空间不变。笛卡尔积大小与不同函数数分别为多少?

同题改动的核验见册末。回访：A1、A19；讲义第 1、9 章。

2. ReLU 零区间与复合不等式（5 分）

【答案】C

【分析】先做复合，再把激活等于零转成输入不大于零的不等式。

【详解】$(f\circ g)(x)=\rho(2(x+1)-1)=\rho(2x+1)$，故为零当且仅当 $2x+1\le0$，即 $x\le-1/2$，选 C。A 反向；B 阈值错；D 只保留边界遗漏整个负输入区域。

【易错与修正】把 ReLU=0 解成内部 =0；把闭端点漏掉。

【点睛】能从分段定义推导集合，而不是只求一个根。

【订正重建】完整写出复合函数的分段式，并说明零点集合为什么不是单个数。

【同题改动】求 $(g\circ f)(x)=0$ 是否有实数解。

同题改动的核验见册末。回访：A2、B16、B17；讲义第 3 章。

3. 长乘积的维度审计（5 分）

【答案】D

【分析】先在每项旁标行列，不需要知道矩阵元素，就能排除没有定义的式子。

【详解】A 为 $(1\times 2)(3\times 2)$，内维不等；B 为 $(2\times 3)(2\times 1)$，也不合法。C 合法但给三维向量；D 为 $(1\times 3)(3\times 2)(2\times 1)$，给标量，选 D。

【易错与修正】只凭字母熟悉程度判断；以为所有写在一起的矩阵都能相乘。

【点睛】能在多步乘积中持续跟踪维度，并用转置恒等式重组计算。

【订正重建】把四个选项都标出形状；对无定义的式子写出哪两个内维不等。

【同题改动】求 $W^\top v$ 的形状，并证明 $v^\top Wx=(W^\top v)^\top x$。

同题改动的核验见册末。回访：A3、A18、B20；讲义第 2 章。

4. 逐元素运算与范数的组合（5 分）

【答案】A

【分析】先逐位置相乘得到完整向量，再对该向量取范数。

【详解】$u\odot v=(3,-2,0)^\top$，一范数为 $3+2+0=5$，选 A。B 是内积绝对值；C、D 不是该向量的绝对值之和。

【易错与修正】先求内积 1 再取绝对值；认为乘 0 的坐标可以改变维数。

【点睛】能跨内积、外积、逐元素乘积与范数保持对象类型清楚。

【订正重建】并列计算 $u\odot v$、$u^\top v$、$uv^\top$ 的类型，至少写出外积第一行。

【同题改动】求 $\|u\odot v\|_2$ 与 $\|u\odot v\|_\infty$。

同题改动的核验见册末。回访：A3、A7、A18；讲义第 2、7 章。

5. 导数的每个因子来自哪里（5 分）

【答案】B

【分析】固定样本后只剩 w 可变。先把模型输出代入损失，再求复合导数。

【详解】$\lambda_{(2,1)}(w)=(2w-1)^2$，导数 $2(2w-1)\cdot 2=8w-4$，在 0 处为-4，选 B。A 漏内层因子 2；C 符号错；D 把损失值 1 当导数。也可用定义：$[(2h-1)^2-1]/h=4h-4\to -4$。

【易错与修正】将 y 或 x 也当作正在求导的变量；漏链式因子；导数与损失本身混淆。

【点睛】能解释导数数值和链式因子，并回到一次真实参数更新。

【订正重建】用链式法则和极限定义各求一次，并标明极限过程先要求 h 不为零。

【同题改动】从 $w_0=0$，用 $\eta=1/8$ 更新一次，求预测和损失。

同题改动的核验见册末。回访：A13、A18、B18；讲义第 4、5 章。

6. 偏置决定边界，约定决定端点（5 分）

【答案】D

【分析】先算分数，再按照本题零点约定分类；边界只由分数等于零确定。

【详解】在 (1,1) 处得分 $2-1-1=0$，按题设输出 1。边界 $2x_1-x_2-1=0$ 即 $x_2=2x_1-1$，选 D。A、C 用错零点约定；B 偏置移项错。

【易错与修正】沿用其他数学语境的 sign(0)=0；把边界本身判为负类；忽略偏置。

【点睛】能把阈值、偏置、不等式和几何边界互相转换。

【订正重建】由加权和与阈值的比较推回偏置 b=-阈值，写出正类、负类及边界。

【同题改动】将偏置从-1 改成 +1，写新边界并判断 (0,0) 的类别。

同题改动的核验见册末。回访：A16、B16；讲义第 3 章。

7. 尖峰误差让不同范数分离（5 分）

【答案】C

【分析】每个 n 下仍是一个矩形脉冲：高度 n，宽度 n 的负二次方；用这一几何结构计算，不逐点积分。

【详解】$\|f_n\|_1 = n \cdot n^{-2} = 1/n$，$\|f_n\|_2 = (n^2 \cdot n^{-2})^{1/2} = 1$，$\|f_n\|_\infty = n$，选 C。A 漏高度；B 混淆平方积分与开方；D 把高度当整体范数。于是平均绝对幅值趋零，但最高幅值增大。

【易错与修正】认为一种范数趋零就意味着所有范数趋零；只看高度不看宽度。

【点睛】能构造而非只口述平均误差小、最坏误差大的反例。

【订正重建】从定义推出一般 $\|f_n\|_q = n^{1-2/q}$，并分 $q<2, q=2, q>2$ 解释极限。

【同题改动】只把支撑宽度改为 n^{-3}，求二范数和最大幅值，并比较两者的极限。

同题改动的核验见册末。回访：A11、A14、B17；讲义第 7 章。

8. 负梯度方向与有限步长（5 分）

【答案】A

【分析】一次更新分三步：旧点梯度、参数改变量、新点的真实损失；不能拿一阶近似替代最后一步。

【详解】$L'(0) = -2$，$w_1 = 0 - (6/5)(-2) = 12/5$，$L(w_1) = (7/5)^2 = 49/25$，选 A。B 漏了导数系数 2；C 更新负号错；D 把参数当损失。新损失大于旧损失 1。

【易错与修正】看到负梯度就断言必下降；只算参数不验证真实损失；混用近似号与等号。

【点睛】能用精确余项或精确递推判断步长，不依赖口号。

【订正重建】对任意 w 推导 $L(w - \eta L'(w)) = (1-2\eta)^2 L(w)$，用它求全部非最优点严格下降的步长范围。

【同题改动】取 $\eta = 1$。损失是否严格下降？序列是否收敛到最优点？

同题改动的核验见册末。回访：A13、B18；讲义第 6 章。

9. 监督信号、反传与优化器（6 分）

【答案】ABD

【分析】检查“训练目标从哪里来”和“这一步计算什么”，避免以有无人工类别划分所有学习。

【详解】选 ABD。无监督可找结构，自监督目标可由原数据构造。GAN 中真实样本与生成样本的来源由采样过程已知，不要求人工猫狗标签，C 错误。[P12] 反向传播计算梯度；梯度下降等规则才把梯度转为参数更新。[P2]

【易错与修正】把所有标签都当人工类别；把反向传播当成学习率选择算法。

【点睛】能按训练信号来源和计算步骤分类，而不靠名词记忆。

【订正重建】分别举一个监督、无监督、自监督情境，并在一轮训练中标出前向、损失、反传、更新。

【同题改动】用句子前面的词预测下一词，目标由原句自动取得。是否可以同时说“存在输入—目标对”和“目标不需逐条人工标注”?

同题改动的核验见册末。回访：A9、A18；讲义第 4、5、12 章。

论文关联：[P2],[P12]。

10. 用具体运算检验等变与不变（6 分）

【答案】ACD

【分析】直接计算移位前后结果。把响应的坐标重排与响应的数值保持不变分开。

【详解】$C(Sx)=(x_3+2x_2,x_1+2x_3,x_2+2x_1)^\top=S(Cx)$，故 A 正确。取 $x=(1,0,0)^\top$，$Cx=(1,2,0)^\top$，$C(Sx)=(0,1,2)^\top$，二者不同，B 错。求和不受重排影响，C 正确；按题设禁用未来位置能限制信息访问，但不等于发现统计因果，D 正确。选 ACD。

【易错与修正】仅因名字是卷积就写不变；因果掩码被解读为因果发现。

【点睛】能亲算等变等式、反例和聚合不变，并认识其代价。

【订正重建】把 A 的恒等式逐坐标算完；为 B 找一组具体反例；说明 C 如何由 A 推出。

【同题改动】令 $r(x)=\sum_i(Cx)_i$，把 r 化简成 x 的表达式，说明这种聚合丢掉了哪些信息。

同题改动的核验见册末。回访：A10、B21；讲义第 12 章。

论文关联：[P11]。

11. 用反例检查四种看似合理的结论（6 分）

【答案】BD

【分析】正确的全称不等式需要证明，错误的全称命题只需一个符合条件的反例。

【详解】B 正确：$(\sum_i |x_i|)^2 \ge \sum_i x_i^2 \ge \max_i x_i^2$，开方即可。A 错，$(t^2-1)^2$ 在 0 驻点但非最小。C 错，取 u=(1,0)、v=(0,1)，拟范数有 $\|u+v\|_{1/2} = 4 > 2 = \|u\|_{1/2} + \|v\|_{1/2}$。D 正确，残差 $t \mapsto t+t$ 放大扰动。选 BD。

【易错与修正】检验几个数都成立就说全称成立；用不符合假设的反例；误称 q<1 也满足所有范数公理。

【点睛】能区分全称证明和反例，并说清范数使用条件。

【订正重建】为 A、C 各写一个有效反例，为 B 写全称证明；明确哪些范数条件被破坏。

【同题改动】找非零向量使 $\|x\|_1 = \|x\|_2 = \|x\|_\infty$，并说明为什么可以。

同题改动的核验见册末。回访：A6、A7、B18、B21；讲义第 6、7、11 章。

12. 二次交叉系数与对称约束（5 分）

【答案】6；$(3,3)$

【分析】将矩阵模型展开成单项式，再比较同类项系数。

【详解】$x^\top A x = 2x_1^2 + (a+b)x_1x_2 + x_2^2$，故 a+b=6。再有对称性才得到 a=b，从而 (a,b)=(3,3)。两空 2、3 分。仅由混合项不能分别确定 a、b。

【易错与修正】直接把 a 和 b 都设为 6；认为 x1x2 与 x2x1 独立；把对称约束当数据已经给出。

【点睛】能辨明数据条件和额外结构约束的不同作用。

【订正重建】先不加对称约束写全部矩阵，再解释对称约束如何消去冗余。

【同题改动】额外要求 $a-b=2$，但不要求对称，求 a、b。

同题改动的核验见册末。回访：A19、B20；讲义第 9 章。

13. 分布不同，也可能对某函数有同一期望（5 分）

【答案】1/3；1

【分析】先按重复次数构造经验概率，再算目标函数在每种取值上的值。

【详解】三条观测中 2 出现一次，故经验概率 1/3。$(Z-1)^2$ 在 Z=0 和 Z=2 时均为 1，因此经验期望为 1；总体期望也恰为 1。两空 2、3 分。分布不同不推出每个函数的期望都不同。

【易错与修正】拿真实概率 1/2 代替经验频率；认为只要分布不同，任何期望都不可能相同。

【点睛】能避免“分布不同所以期望必不同”的逻辑跳跃。

【订正重建】用同一张支持集表分别计算真实概率、经验概率和两种期望。

【同题改动】把目标函数改成 $g(Z)=Z$，求经验期望与真实期望。

同题改动的核验见册末。回访：A5、A17、B19；讲义第 8 章。

14. 参数矩阵展平不是更换模型（5 分）

【答案】2×3；23

【分析】每一层权重的行数等于输出维数，列数等于输入维数；偏置跟输出同维。

【详解】$W_1\in\mathbb{R}^{3\times 4}, b_1\in\mathbb{R}^3, W_2\in\mathbb{R}^{2\times 3}, b_2\in\mathbb{R}^2$。实际存储数 p=12+3+6+2=23。两空 2、3 分。展平仅是按约定编码；执行前向仍要恢复形状，且 23 不自动是可辨识函数自由度。

【易错与修正】把行列反过来；漏偏置；把参数向量当输出向量；把展平误认为线性模型。

【点睛】能从网络结构计算参数量，并区分形状、数量与有效自由度。

【订正重建】写四个对象形状、每个元素数与输出形状，并用一句话区分展平和换用二次模型。

【同题改动】去掉两层偏置，其他不变，参数数目和第二层形状如何变化?

同题改动的核验见册末。回访：A4、A12、B20；讲义第 2、9 章。

15. 约束最优不一定是驻点（5 分）

【答案】1；$\{1\}$

【分析】先检查允许的参数范围，再看目标函数在这个区间往哪个方向减小。

【详解】$L(t)=(t-2)^2$ 在 $[-1,1]$ 上递减，故右端点 $t=1$ 最优；最小值为 1，最小化点集合为 $\{1\}$。两空分别 2、3 分。无约束最优点 2 不允许使用；此处导数为 -2，并非零。

【易错与修正】直接解 L′=0 得到不合法点；把最小值与最小化点集合颠倒。

【点睛】能在有约束和最小值不取到的情形中使用正确优化记号。

【订正重建】画出参数约束区间，解释为什么最优点导数不为零不构成矛盾。

【同题改动】允许空间改成 $[-1,1)$，最小值能否取得？下确界与 argmin 是什么？

同题改动的核验见册末。回访：A6、B18；讲义第 1、6 章。

16. 非线性表征使 XOR 可分（12 分）

【答案】分问结果及证明见详解。

【分析】先证明原空间中不可能，再构造新特征。这样知道非线性到底改变了什么，而不是只见到一次成功拟合。

【详解】以下逐问恢复推理；括号内为分步给分。

(1) 2 分。四个预测依次为 $(-1,+1,+1,+1)$，只在 $(1,1)$ 处出错，共 1 个错误。

评分：预测序列 1 分；错误个数 1 分。

(2) 4 分。全部正确要求 $b<0$，$w_1+b\geq 0$，$w_2+b\geq 0$，$w_1+w_2+b<0$。前两个正类约束相加后有 $w_1+w_2+b\geq -b>0$，与最后约束矛盾。

评分：四约束 2 分；正确合并 1 分；矛盾结论 1 分。边界约定决定负类必须使用严格不等号。

(3) 4 分。$h_1+h_2=|x_1-x_2|$，在同类输入上为 0，异类输入上为 1。故全部参数恰为 $\{(a,c):c<0,\ a+c\geq 0\}$。必要性来自两类的得分要求，充分性由代回两个可能的特征值 0、1 立即验证。

评分：绝对值与取值 1 分；两约束 2 分；验证充分性 1 分。

(4) 2 分。去掉两个 ReLU 后，$h_1+h_2=(x_1-x_2)+(x_2-x_1)=0$，输出为常数 $\mathrm{sign}(c)$，不能表示两种标签。

评分：和恒为零 1 分；常数分类不能完成任务 1 分。

【易错与修正】尝试很多参数当成不可能性证明；负类边界写成非严格；构造参数只验证必要性不验证充分性。

【点睛】不可分性取决于特征表示；非线性映射可以改变几何结构，而单纯增加仿射层不能。

【订正重建】独立写四个分类约束和矛盾；再用两种隐藏特征值 0、1 推导全部参数范围。

【同题改动】把 XOR 标签全部翻转，仍使用同一隐藏特征，求 $\mathrm{sign}(a(h_1+h_2)+c)$ 正确的全部 (a,c)。

同题改动的核验见册末。回访：A12、A16、A20；讲义第 3 章。

论文关联：[P2]。原创构造题，体现 [P2] 中“隐藏单元学习内部表征”的思想；不是原论文中的完整定理复现。

17. 从折点结构构造平方逼近（12 分）

【答案】分问结果及证明见详解。

【分析】A20 制造折点，这题给它们一个任务：修正平方函数插值的误差。每轮比较旧端点、新中点，再用分段仿射的唯一性补齐整个区间。

【详解】以下逐问恢复推理；括号内为分步给分。

（1）3 分。

$$L_{a,b}(t) = (a+b)t - ab, \qquad L_{a,b}(t) - t^2 = (t-a)(b-t).$$

设 $r = t - a \in [0, b-a]$，则误差 $r(b-a-r)$ 在中点最大，等于 $(b-a)^2/4$，且区间内非负。

评分：割线式及恒等式 2 分；最大误差和取点 1 分。

（2）4 分。旧区间长度 $H = 2^{-(m-1)}$。旧端点上两插值相等，差为零；新中点上 $\mathrm{P_m}$ 已等于真实平方，旧插值比平方高 $H^2/4 = 4^{-m}$。两插值的差在每个新半区间都是仿射函数。

$\mathrm{T_m}$ 在相同三个节点依次取 0、1、0，且在每半区间仿射，故差函数处处为 $4^{-m}T_m$。因此 $c_m = 4^{-m}$。

评分：旧端点 1 分；中点差值 1 分；分段仿射端点判等 1 分；确定全域恒等式 1 分。

（3）2 分。

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

误差条件为 $2^{-2m-2} \le 2^{-10}$，故最小非负整数 $m = 4$。

评分：展开式 1 分；误差及最小 m 合计 1 分。

（4）3 分。保留状态 (T_{s-1}, P_{s-1})，一轮计算 $T_s = T(T_{s-1})$，再令 $P_s = P_{s-1} - 4^{-s}T_s$。每轮仅有固定个数的仿射操作与 ReLU；在 [0,1] 上状态非负，复制通道可用 ReLU 实现恒等传递。允许两层常宽模块完成一轮，因而总层数和单元数均为 $\mathrm{O(m)}$，而非把全部 $\mathrm{T_s}$ 分别重算的 $\mathrm{O(m^2)}$。

误差为 $4^{-(m+1)}$，所以充分取 $m = O(\log(1/\epsilon))$。这仅建立平方函数的显式构造，尚未证明任意连续函数的通用逼近。

评分：状态复用与常宽实现 1 分；复杂度和误差联系 1 分；限定函数范围 1 分。

【易错与修正】把粗网格宽度与细网格宽度混淆；漏误差中的 1/4；重算所有三角波导致 $\mathrm{O(m^2)}$；把平方特例说成完整通用逼近。

【点睛】深层网络不仅能制造许多折点，还可用这些折点构造有明确误差界的数值近似；逼近精度与资源之间可定量联系。

【订正重建】不看答案写出三节点的差值、望远镜求和与误差上界，并画一轮状态 (Ts,Ps) 如何复用。

【同题改动】在 [0,2] 上用 $Q_m(x) = 4P_m(x/2)$ 逼近平方，求最大误差及不超过 1/64 的最小 m。

同题改动的核验见册末。回访：A20、A14、B7；讲义第 10 章。

论文关联：[P5]。经典来源：[P5] 第 3.1 节 Proposition 2。这里证明的是其平方函数构造，不是 [P3] 的一般通用逼近定理。

18. 从局部下降到全程函数值收敛（12 分）

【答案】分问结果及证明见详解。

【分析】两条条件各有分工：光滑上界控制走一步的代价；PL 条件将梯度大小转为距最优值的差。要把它们连起来才能得到递推。

【详解】以下逐问恢复推理；括号内为分步给分。

(1) 3 分。记 $e=(w-2,b-1)^{\top}$，则 $\Lambda=\|e\|_2^2$，$g=\nabla\Lambda=2e$。更新后 $e^{+}=(1-2\eta)e$，因而 $\Delta\Lambda=-\eta(1-\eta)\|g\|_2^2$。非最优点 g 非零，严格下降恰当且仅当 $0<\eta<1$；$\eta=1/2$ 使 $e^{+}=0$。

评分：梯度 1 分；精确损失变化与严格范围 1 分；一步最优学习率 1 分。

(2) 3 分。

$$F(\theta-\eta g)-F(\theta)\le-\eta\|g\|_2^2+\frac{L_s\eta^2}{2}\|g\|_2^2=-\eta\left(1-\frac{L_s\eta}{2}\right)\|g\|_2^2.$$

在 $g\neq0$ 时，$0<\eta<2/L_s$ 足以保证下降。这是从给定光滑上界得出的充分条件，不声称对每个 F 都为必要条件。

评分：代入 1 分；整理二次项 1 分；范围及“充分”措辞 1 分。

(3) 4 分。

$$F(\theta_{k+1})\le F(\theta_k)-\frac{1}{2L_s}\|\nabla F(\theta_k)\|_2^2\le F(\theta_k)-\frac{\mu}{L_s}(F(\theta_k)-F^*).$$

两端减去 F* 得误差递推，反复代入即 $F(\theta_k)-F^*\le(1-\mu/L_s)^k[F(\theta_0)-F^*]$。当 $0<\mu<L_s$，因子在 (0,1) 中，表示每次以固定比例缩小误差，这叫线性收敛；不是误差对 k 的一次函数。$\mu=L_s$ 时给定条件使一步到最优值。

评分：代入 η 及第一步下降 1 分；用 PL 条件 1 分；归纳界 1 分；解释线性收敛 1 分。

(4) 2 分。PL 下若梯度为零，则 $0\ge\mu(F-F^*)\ge0$，故 F=F*。一般可微函数不满足此结论：$G'(0)=0$，但 G(0)=1，而 G(±1)=0。

评分：PL 驻点推出全局最优 1 分；反例 1 分。

【易错与修正】把充分步长范围写成所有函数必要范围；把线性收敛理解成误差一次函数；从函数值收敛直接推出参数唯一。

【点睛】光滑性控制一步更新的代价；PL 条件把梯度大小和剩余损失联系起来。二者连接才产生全程几何收敛。

【订正重建】按“代入负梯度—选 η—用 PL—两边减 F*—归纳”重写五步，标记每一步用了哪个条件。

【同题改动】考虑 $F(w,b)=(w-2)^2$。它可取 $L_s=\mu=2$。证明梯度下降虽可一步到最优值，最优参数却不唯一。

同题改动的核验见册末。回访：A6、A13、B8；讲义第 6 章。

论文关联：[P6]。经典来源：[P6] 第 2 节 Theorem 1 的证明路线；PL 思想可追溯至 Polyak 1963 年工作。

19. 从固定模型界到训练后选择的保证（12 分）

【答案】分问结果及证明见详解。

【分析】训练会挑模型，所以不能事后假装模型预先固定。先保证所有候选模型同时受控，再允许在这个事件里任意挑选。

【详解】以下逐问恢复推理；括号内为分步给分。

(1) 3 分。对每个 h 定义坏事件 $E_h = \{|\widehat{R}(h) - R(h)| > \epsilon\}$。最大偏差超标等价于至少一个 $\mathrm{E_h}$ 发生。由并集界与给定集中界得：

$$\Pr\left(\bigcup_{h\in\mathcal{H}} E_h\right) \le \sum_{h\in\mathcal{H}} \Pr(E_h) \le 2Je^{-2M\epsilon^2}.$$

评分：事件等价 1 分；并集界 1 分；代入得到结果 1 分。

(2) 3 分。要求 $2Je^{-2M\epsilon^2} \le \delta$。取对数并整理，可取

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

评分：概率阈值不等式 1 分；正确取对数 1 分；非负平方根及好事件概率 1 分。若 ϵ>1，结论仍真但可能缺乏实用精度。

(3) 4 分。

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

第一、三步使用同时对所有候选模型成立的偏差界；中间一步使用经验最小化性质。即便最优模型不唯一，任取一个也成立。

评分：三步各 1 分；说明中间是 ERM、两侧是同一个好事件上的统一界 1 分。

(4) 2 分。使 $2\epsilon \le a$ 的充分条件为 $M \ge \frac{2}{a^2}\log\frac{2J}{\delta}$；M 为整数时向上取整。训练后选择的模型依赖同一数据集，不能直接把“抽样前固定 h”的保证当成选择后保证；先建立全体同时成立的事件即可覆盖这个选择。

评分：样本量条件 1 分；说明数据依赖选择及一致事件 1 分。

【易错与修正】将置信度当单条预测正确率；忘记 J 代价；把随机选择的模型直接代入固定模型界；风险差中漏一个 ε。

【点睛】从一个模型的偏差控制到训练后选择的模型，需要应对模型选择本身；有限模型类可借助并集界支付 logJ 代价。

【订正重建】独立写坏事件、并集界和三段比较链；标明概率对训练集抽样而言，而不是对固定答案。

【同题改动】优化只找到 $\widetilde{h}$，满足 $\widehat{R}(\widetilde{h}) \le \min_h \widehat{R}(h) + \tau$。在同一好事件上推导真实风险界。

同题改动的核验见册末。回访：A9、A17、B13；讲义第 8 章。

论文关联：[P7],[P8]。经典来源：[P7] 的集中工具与 [P8] 的一致收敛思想。证明的是有限模型类特例，不证明 Hoeffding 引理或完整 VC 定理。

20. 低秩近似要同时有下界与可达构造（12 分）

【答案】分问结果及证明见详解。

【分析】不在四个冗余因子参数里蛮算。低秩矩阵会丢失某个方向；沿这个方向看原矩阵仍保留多少，再构造最好的保留方案。

【详解】以下逐问恢复推理；括号内为分步给分。

（1）3 分。

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

输出是标量。直接存储 d^2=1,000,000 个数，因子存储 2dr=20,000 个数，比例为 1/50。

评分：恒等变形 1 分；中间维度和输出 1 分；两项参数量 1 分。

（2）3 分。将 E 第 i 行写作 $e_i{}^{\mathrm{T}}$，对单位向量 z 应用 Cauchy：$(e_i^\top z)^2 \le \|e_i\|_2^2 \|z\|_2^2 = \sum_j E_{ij}^2$。逐行求和后开方即得

$$\|Ez\|_2^2 = \sum_i (e_i^\top z)^2 \le \sum_{i,j} E_{ij}^2 = \|E\|_F^2.$$

评分：行内积表示 1 分；逐行 Cauchy1 分；累加及开方 1 分。

（3）4 分。若 v=0 可任取单位 z；若 v 非零，取与 v 垂直的单位 z。均有 Bz=0。设 $z=(z_1,z_2)^{\mathrm{T}}$，则 $\|Dz\|_2^2 = 9z_1^2 + z_2^2 \ge z_1^2 + z_2^2 = 1$。因此

$$\|D - B\|_F \ge \|(D-B)z\|_2 = \|Dz\|_2 \ge 1.$$

取 $B = \begin{pmatrix} 3 & 0 \\ 0 & 0 \end{pmatrix}$，例如 $u=(3,0)^{\mathrm{T}}$、$v=(1,0)^{\mathrm{T}}$，则误差矩阵为 diag(0,1)，Frobenius 范数正好为 1，故为最佳。

评分：构造零空间方向 1 分；得出范数下界 2 分；构造可达矩阵 1 分。

（4）2 分。$(\alpha u)(\alpha^{-1}v)^\top = uv^\top$，非零因子可按 α 缩放而保持矩阵不变，故因子表示一般不唯一。前问最低误差为 1>0，也表明该压缩不能无损表示给定 D。

评分：缩放冗余 1 分；用严格正误差说明不无损 1 分。

【易错与修正】找到一个误差为 1 的矩阵就宣布最优；把最优矩阵与唯一因子混淆；把矩阵误差不加条件等同于任务误差。

【点睛】最佳低秩近似的下界可沿被低秩映射消去的方向建立；构造达到下界的矩阵，才能证明“最佳”。

【订正重建】先证明行内积的范数不等式，再找 Bz=0 的单位方向，最后给达到下界的矩阵。

【同题改动】由本题不等式证明对任意 x 有 $|x^\top(A-B)x| \le \|A-B\|_F \|x\|_2^2$；说明无界输入时矩阵误差小是否必使所有输出误差小。

同题改动的核验见册末。回访：A19、A21、B14；讲义第 9 章。

论文关联：[P9]。经典来源：[P9] 最佳低秩近似结果的二维对角、秩一特例；不要求使用奇异值分解，也不声称证明一般矩阵版本。

21. 残差、梯度与数值误差的三重辨析（12 分）

【答案】分问结果及证明见详解。

【分析】同一条链，既可以逐层相乘导数，也可先把状态相邻差求和再求导；比较两种写法后，检查“+1”有没有被其他项抵消。

【详解】以下逐问恢复推理；括号内为分步给分。

（1）2 分。逐式相加，$\sum_{k=0}^{L-1}(x_{k+1}-x_k)=\sum_{k=0}^{L-1}F_k(x_k)$，左端望远镜相消为 xL-x0，故得所证。

评分：相邻差求和 1 分；望远镜化简 1 分。

（2）4 分。

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

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

乘积式来自逐层链式法则；求和式来自对前问求导。每个 x_k 本身依赖 x_0，不能把 F′ 项后的 dx_k/dx_0 删掉。两式是同一导数的两种组织方式。

评分：两表达式各 1 分；两条推导路线各 1 分。

（3）3 分。若每层 $F_k(t)=-t$，则一层映射已为零映射，对 $L\geq 1$ 导数为 0。若每层 $F_k(t)=t$，则每层导数为 2，总导数为 2^L。恒等捷径不能无条件排除梯度消失或爆炸；“1+ 求和”中的求和可抵消 1，也可放大它。

评分：两个反例导数各 1 分；正确逻辑判断 1 分。

（4）3 分。此时整体一层为 $(1-\beta)t$，Lipschitz 系数 $1-\beta\in(0,1)$。于是

$$e_L\leq\varepsilon\sum_{j=0}^{L-1}(1-\beta)^j=\frac{\varepsilon[1-(1-\beta)^L]}{\beta}\leq\frac{\varepsilon}{\beta}.$$

收缩来自整体映射的斜率绝对值小于 1，而非“残差”名称。持续新增误差时上界可稳定在非零量级。

评分：收缩系数 1 分；几何级数界 1 分；解释条件 1 分。

【易错与修正】遗漏 dxk/dx0；见到 +1 就断言梯度 ≥1；将数值误差有界当梯度不消失。

【点睛】残差形式提供代数上的直通项；总传播仍由完整导数及各项相互作用决定。结构优势与无条件保证必须分开。

【订正重建】从两种路线重新推导同一导数，用 Fk=-t 和 Fk=t 逐一检验，最后独立推误差几何级数。

【同题改动】取每层 $F_k(t)=-t/2$。同时求 dx_L/dx_0 与加性误差上界，并解释这能否既数值稳定又梯度趋零。

同题改动的核验见册末。回访：A18、A21、B11；讲义第 5、11 章。

论文关联：[P10]。经典来源：[P10] 第 2 节公式 (3)–(5) 的标量可微版本；反例用于限定结论的适用条件。

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】分别为 3、1；所有候选函数在唯一输入 0 上都输出 0。函数相同的判断依赖定义域。

【2 题】$g(f(x)) = \rho(2x-1)+1 \geq 1$，无解。复合顺序会连零点的有无都改变。

【3 题】$W^\top v \in \mathbb{R}^2$；利用 $(AB)^\top = B^\top A^\top$ 得 $(W^\top v)^\top = v^\top W$。

【4 题】分别为 $\sqrt{13}$、3；外积第一行为 $(3,1,-4)$，外积是 3×3 矩阵。

【5 题】$w_1 = 0-(1/8)(-4) = 1/2$，预测 $2w_1 = 1$，平方损失 0。

【6 题】边界为 $x_2 = 2x_1+1$，原点得分 1，属于正类。

【7 题】二范数为 $(n^2 n^{-3})^{1/2} = n^{-1/2} \to 0$，最大幅值仍为 n，趋于无穷。有限 q 范数小仍不必逐点最坏误差小。

【8 题】非最优初值下 $w_{k+1}-1 = -(w_k-1)$，损失不变且在两点之间往返；0<η<1 才对全部非最优点严格下降。

【9 题】可以；前者说明优化可以写成输入—目标预测，后者说明监督信号的来源是自监督。两种说法并不矛盾。

【10 题】$r(x) = 3(x_1+x_2+x_3)$。如 $(1,0,0)$ 和 $(0,1,0)$ 有相同聚合结果，位置排列的信息不再保留；不变性不等于保留一切有用信息。

【11 题】任取只有一个非零分量的向量，例如 (-3,0)，三种范数均为 3。多个非零分量则 1 范数与 2 范数的平方存在正交叉项。

【12 题】联立 a+b=6、a-b=2，得 a=4、b=2；与 (3,3) 矩阵给出同一二次函数。

【13 题】经验期望 2/3，总体期望 1；这次可以区分，但一个期望相同仍不够确定分布相同。

【14 题】参数数目为 12+6=18，第二层权重仍 2×3；改变偏置是否训练会改变模型族，单纯展平则不改变。

【15 题】下确界为 1，但 t=1 不允许，故没有最小值，argmin 为空集 $\varnothing$。

【16 题】同分量时要正类，c≥0；不同分量时要负类，a+c<0。两条件共同充要。

【17 题】$|Q_m(x)-x^2| = 4|P_m(x/2)-(x/2)^2|$，最大误差 4^{-m}。要求 $4^{-m} \leq 4^{-3}$，最小 m=3。

【18 题】梯度为 (2(w-2),0)，η=1/2 更新后 w=2，b 不变。所有 (2,b) 都是最优参数；因此函数值收敛与参数唯一不是同一结论。

【19 题】$R(\widetilde{h}) \leq \widehat{R}(\widetilde{h})+\epsilon \leq \widehat{R}(h^*)+\tau+\epsilon \leq R(h^*)+2\epsilon+\tau$。估计误差与优化误差在这条界中分开出现。

【20 题】设 E=A-B。Cauchy 给 $|x^\top Ex| \leq \|x\|_2\|Ex\|_2$，缩放单位向量不等式得 $\|Ex\|_2 \leq \|E\|_F\|x\|_2$。故所证成立；需限制输入范数才能得到一致的小输出误差。

【21 题】整体层斜率 1/2，导数为 $2^{-L} \to 0$；加性误差界为 $2\varepsilon(1-2^{-L}) \leq 2\varepsilon$。两者可同时成立：数值扰动控制与反向梯度保留是不同目标。

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。

【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

【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