考号：_____　　姓名：_____

磐石计划 01

基本数学语言与深度学习联合考试试卷（A 卷）

本试卷共5页，21 小题，满分 150 分。计时重做可用 120 分钟；首次学习可分段完成。

【注意事项】在另纸或配套作答页作答。保留首次答案；独立尝试后可查分层提示，订正时遮住解析重新推导。提示使用不扣学习分，但须记录。

向量默认为列向量；$\rho(t)=\max\{0,t\}$，对向量逐分量作用；感知机约定 $\operatorname{sign}(0)=1$。函数的有限 q 范数按 $\|f\|_q=(\int |f|^q)^{1/q}$（$q\ge 1$）计算。本卷 $\|f\|_\infty$ 按逐点上确界。

一、单项选择题：共 8 小题，每小题 5 分，共 40 分。每题只有一个选项符合要求。

1. 设模型 $m(x;(w,b))=wx+b$，其中 x,w,b 均可取任意实数，记 $\theta=(w,b)$。下列对该模型定义域与输出空间的记法，正确的是（　　）

A. $m:\mathbb{R}^2\to\mathbb{R}^2$　　B. $m:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$　　C. $m:\mathbb{R}\times\mathbb{R}^2\to\mathbb{R}$　　D. $m:\mathbb{R}\to\mathbb{R}^2$

2. 已知 $f(t)=t^2$，$g(t)=2t-1$，则 $(f\circ g)(-1)$ 等于（　　）

A. -3　　B. 1　　C. 5　　D. 9

3. 设 $x=(1,4,-1)^\top$，$\theta=(2,-1,3)^\top$。线性模型 $m(x;\theta)=\theta^\top x$ 的输出为（　　）

A. -5　　B. 5　　C. $(2,-4,-3)^\top$　　D. $(2,-1,3)^\top$

4. 令 $\rho(t)=\max\{0,t\}$，对向量按分量作用。已知 $W=\begin{pmatrix}1&0&-1\\2&1&0\end{pmatrix}$，$x=(1,-2,3)^\top$，$b=(1,1)^\top$，则 $\rho(Wx+b)$ 等于（　　）

A. $(-1,1)^\top$　　B. $(0,1)^\top$　　C. $(1,0)^\top$　　D. $(0,0)^\top$

5. 训练数据多重集为 $\mathcal{D}=\{(0,0),(1,1),(1,1),(2,0)\}$。分类器 $h(x)=\mathbf{1}_{[1,\infty)}(x)$ 采用 0-1 损失 $\ell_y(\hat y)=\mathbf{1}_{\{\hat y\neq y\}}$。其经验平均损失为（　　）

A. $\frac14$　　B. $\frac13$　　C. $\frac12$　　D. $\frac34$

6. 设 $L(t)=(t^2-1)^2$，$t\in\mathbb{R}$。按照最小化点集合的定义，$\operatorname{argmin}_{t\in\mathbb{R}}L(t)$ 为（　　）

A. $\{0\}$　　B. $\{1\}$　　C. $\{-1,1\}$　　D. $\mathbb{R}$

7. 设 $x=(3,-4)^\top$，则 $\|x\|_1-\|x\|_\infty$ 等于（　　）

A. 0　　B. 3　　C. 5　　D. 7

8. 定义 $s(t)=\dfrac{1}{1+e^{-t}}$，$u(t)=\dfrac{e^t-e^{-t}}{e^t+e^{-t}}$。对任意实数 t，$s(t)+s(-t)+u(t)+u(-t)$ 恒等于（　　）

A. 0　　B. 2　　C. $s(t)$　　D. 1

二、多项选择题：共 3 小题，每小题 6 分，共 18 分。

全部选对得 6 分；有选错或未选得 0 分；非空且只选对部分时，选对 1 个得 2 分，选对 2 个得 4 分。

9. 关于模型表达、训练与泛化，下列说法正确的有（　　）

A. 利用已给定的输入—标签对 (x, y)，以标签评价预测并训练模型，属于监督学习。

B. 只提供输入样本而不提供外部标签，仍可通过聚类等方法寻找数据结构。

C. 具有通用逼近能力，就能保证梯度下降从任意初值找到全局最优参数。

D. 存在能够逼近目标函数的参数，不等于训练算法一定找到它，也不等于模型一定能泛化。

10. 关于网络结构所体现的假设，下列说法正确的有（　　）

A. 同一卷积核在各位置共享参数；忽略边界效应时，输入平移使特征相应平移，体现平移等变性。

B. 任何 Transformer 都天然禁止当前位置使用未来位置的信息，无须额外约束。

C. 自回归 Transformer 中的因果掩码，可以限制当前位置对未来位置的访问。

D. 符合任务结构的归纳偏置可能有益；不符合真实规律的结构约束也可能导致欠拟合。

11. 设 $g(t) = \mathbf{1}_{[0,2]}(t)$，$h(t) \equiv 1$，二者定义域均为 $\mathbb{R}$。以下积分均对整个实数轴进行。下列说法正确的有（　　）

A. $\|g\|_1 = 2$。

B. $\|g\|_2 = 2$。

C. 对每个有限实数 $q \geq 1$，均有 $g \in L^q(\mathbb{R})$。

D. 因为 h 有界，所以 $h \in L^2(\mathbb{R})$。

三、填空题：共 4 小题，每小题 5 分，共 20 分。

12. 设 $h = W_1 x + b_1$，$y = W_2 h + b_2$，其中 $W_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$，$b_1 = (1, 0)^\top$，$W_2 = (2, -1)$，$b_2 = 3$。若写成 $y = ax_1 + bx_2 + c$，则 $(a, b, c) =$ _____。

13. 对数据集 $\mathcal{D} = \{(1, 3), (2, 5)\}$，用模型 $m(x; a) = ax + 1$ 和平方损失训练。平均损失 $\Lambda(a) =$ _____；其最小化点集合为 _____。

14. 设 $f(t) = t$ $(0 \leq t < 1)$，在其余实数处 $f(t) = 0$。则 $\|f\|_2 =$ _____，$\sup_{t \in \mathbb{R}} |f(t)| =$ _____。

15. 某计算过程的误差满足 $e_0 = 0$、$0 \leq e_{k+1} \leq \frac{1}{2} e_k + \frac{1}{100}$。用此递推逐层估计得到 $e_4 \leq$ _____。要求填写能够由等号递推达到的上界。

四、解答题：共 6 小题，每小题 12 分，共 72 分。解答须写出文字说明、证明过程或演算步骤。

16.（12 分）

有标签样本 (x_i, y_i)，其中 $x_i \in \mathbb{R}^d$、$y_i \in \{-1, 1\}$，且 $\|x_i\|_2 \le R$。假设存在单位向量 u 和常数 $\gamma > 0$，使每个样本都满足 $y_i u^\top x_i \ge \gamma$。

从 $w^{(0)} = 0$ 开始，循环检查样本；当 $y_i (w^{(k)})^\top x_i \le 0$ 时，执行一次更新 $w^{(k+1)} = w^{(k)} + y_i x_i$。这里 k 只统计已执行的更新次数，不统计未更新的检查次数。

(1) 先取 $x = (1, -1)^\top, y = 1, w^{(0)} = 0$，按规则做一次更新，求 $w^{(1)}$ 及更新后该样本的得分 $y (w^{(1)})^\top x$。(2 分)

(2) 回到一般条件，分别证明 $u^\top w^{(k)} \ge k\gamma$ 和 $\|w^{(k)}\|_2^2 \le kR^2$。说明第二个不等式中触发更新的条件用在何处。(5 分)

(3) 结合前问，求仅用 R、γ 表示的更新次数上界，并在 $R = 2, \gamma = 1/2$ 时计算此界。(3 分)

(4) 该结论是否单独保证新样本上的分类错误率为零？说明理由。(2 分)

附：可使用 $|a^\top b| \le \|a\|_2 \|b\|_2$。

17.（12 分）

总体分布 P 的全部可能样本如下表。训练多重集为 $\mathcal{D} = \{(0,0), (1,1), (1,1), (1,1)\}$。采用 0-1 损失，考虑分类器 $h_A(x) = \mathbf{1}_{[1,\infty)}(x)$、$h_B(x) = \mathbf{1}_{\{1\}}(x)$、$h_C(x) = 0$。

x	y	$P(X = x, Y = y)$
0	0	$\frac{1}{4}$
1	1	$\frac{1}{4}$
2	0	$\frac{1}{2}$

(1) 分别求 $\mathbb{E}_{P_\mathcal{D}}[X]$ 与 $\mathbb{E}_P[X]$。(2 分)

(2) 分别求三个分类器的经验风险 Λ 与总体风险 R。(6 分)

(3) 在模型族 $\{h_A, h_B, h_C\}$ 中，分别写出经验风险与总体风险的最小化模型集合；指出训练未覆盖的事件，并解释为什么零训练风险不足以保证零总体风险。(4 分)

18.（12 分）

记 $s(t)=\frac{1}{1+e^{-t}}$。考虑两输入、两隐藏单元、一个线性输出的网络（$i,j\in\{1,2\}$）：

$$z_j=\sum_{i=1}^{2}W_{ji}x_i+b_j,\quad h_j=s(z_j),\quad o=v_1h_1+v_2h_2+c,\quad L=\frac{1}{2}(o-y)^2.$$

所有参数 W,b,v,c 均可训练，输入 x 与标签 y 固定。这里平方损失特意带系数 $\frac{1}{2}$。

（1）由定义求导，证明 $s'(t)=s(t)(1-s(t))$。（2 分）

（2）令 $\delta=o-y$。用链式法则分别求 $\frac{\partial L}{\partial v_j}$、$\frac{\partial L}{\partial c}$、$\frac{\partial L}{\partial b_j}$、$\frac{\partial L}{\partial W_{ji}}$，并说明计算中哪些中间量可以复用。（5 分）

（3）取 $x=(1,-1)^\top, y=1$，$W=0, b=0, v=(1,-1)^\top, c=0$。以学习率 $\eta=\frac{1}{10}$ 同时更新所有参数，求更新后的 W^+ 与 c^+。所有梯度均须在更新前的同一参数点计算。（3 分）

（4）保持 δ, v_j, x_i 有界，当 $|z_j|$ 很大时，通向第 j 个隐藏单元的梯度因子有何变化？这是否意味着该参数点一定是最优点？（2 分）

19.（12 分）

考虑 $m(x;A)=x^\top Ax$，$x=(x_1,x_2)^\top$，$A=\begin{pmatrix}a & u\\ v & c\end{pmatrix}$。训练集最初只有 $((1,0)^\top,1)$ 与 $((0,1)^\top,1)$，采用平方损失。

（1）展开模型，指出四个矩阵元素通过哪几个独立组合影响输出。（3 分）

（2）求使两条训练样本损失均为零的全部矩阵，并判断它们在新输入 $(1,1)^\top$ 上的预测是否唯一。（3 分）

（3）加入第三条样本 $((1,1)^\top,2)$ 后，零损失矩阵给出的函数是否唯一？参数矩阵是否唯一？再要求 $A=A^\top$ 后结论如何？（3 分）

（4）对一般 $d\times d$ 实矩阵，证明 $x^\top Ax=x^\top\frac{A+A^\top}{2}x$，并求一般齐次二次型独立系数个数。（3 分）

20.（12 分）

在 $[0,1]$ 上定义 $T(t)=2\rho(t)-4\rho(t-\frac{1}{2})$，其中 $\rho(t)=\max\{0,t\}$。令 $T_1=T$、$T_{k+1}=T\circ T_k$。线性段指斜率不变的最大区间，内部折点指开区间 $(0,1)$ 内斜率改变的位置。

（1）求 T、T_2 的分段表达式，并画出 T_2 的图象。（4 分）

（2）猜想并证明 T_k 的线性段数与内部折点数。证明中须说明相邻线性段不能合并。（5 分）

（3）单隐藏层模型为 $H(t)=c+\sum_{j=1}^{N}a_j\rho(b_jt+d_j)$。若它精确表示 T_5，用折点计数给出 N 的必要下界（不要求证明下界可达）；与连续复合五个 T 模块使用的 ReLU 单元数比较。该比较能否推出所有任务中深网络都更优？（3 分）

21.（12 分）

理想状态满足 $x_{k+1}=f_k(x_k)$，实际状态满足 $\widetilde{x}_{k+1}=f_k(\widetilde{x}_k)+\varepsilon_k$。设 $e_k=\|\widetilde{x}_k-x_k\|$，$e_0=0$，$\|\varepsilon_k\|\le\varepsilon$，且 $\|f_k(u)-f_k(v)\|\le K\|u-v\|$，$K\ge 0$。

（1）利用三角不等式证明 $e_{k+1}\le Ke_k+\varepsilon$。（3 分）

（2）求 e_L 的几何级数上界，分别讨论 $0\le K<1$、$K=1$、$K>1$。说明上界的指数增长是否意味着任意实际误差都指数增长。（5 分）

（3）深结构执行 L 个 $d\times d$ 稠密矩阵—向量乘法；浅结构执行一次 $d\to N$ 与一次 $N\to d$ 稠密投影。每个矩阵元素计一次乘加，忽略其余操作。取 $L=8,d=64,N=512$，计算两种成本，并说明总乘加次数是否单独决定延迟或任务精度。（4 分）

——全卷结束——