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（B 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 $\mathcal{X}=\{0,1\}$，$\Theta=\{(1,0),(0,1),(1,1)\}$，$m(x;(a,b))=(a+b)x$。则 $\mathcal{X}\times\Theta$ 的元素个数与模型族 $\{m(\cdot;\theta):\theta\in\Theta\}$ 中不同函数的个数分别为（　　）

A. $3,3$　　B. $6,2$　　C. $6,3$　　D. $2,6$

2. 设 $f(t)=\rho(2t-1)$，$g(t)=t+1$，其中 $\rho(t)=\max\{0,t\}$。使 $(f\circ g)(x)=0$ 成立的全部实数 x 构成的集合为（　　）

A. $[-\frac{1}{2},\infty)$　　B. $(-\infty,\frac{1}{2}]$　　C. $(-\infty,-\frac{1}{2}]$　　D. $\{-\frac{1}{2}\}$

3. 设 $W\in\mathbb{R}^{3\times 2}$，$x\in\mathbb{R}^2$，$v\in\mathbb{R}^3$，向量均为列向量。下列式子中，有定义且结果为标量的是（　　）

A. $x^\top W$　　B. $W^\top x$　　C. $(Wx)\odot v$　　D. $v^\top W x$

4. 设 $u=(1,-2,0)^\top$，$v=(3,1,-4)^\top$。则 $\|u\odot v\|_1$ 等于（　　）

A. 5　　B. 1　　C. 7　　D. 9

5. 模型为 $m(x;w)=wx$，损失为 $\ell_y(\hat{y})=(\hat{y}-y)^2$。固定样本 $(x,y)=(2,1)$，定义 $\lambda_{(2,1)}(w)=\ell_1(m(2;w))$，则 $\lambda'_{(2,1)}(0)$ 等于（　　）

A. -2　　B. -4　　C. 4　　D. 1

6. 本题约定 $\operatorname{sign}(0)=1$。对感知机 $m(x)=\operatorname{sign}(2x_1-x_2-1)$，其在点 $(1,1)$ 的预测与其决策边界分别为（　　）

A. -1；$x_2=2x_1-1$　　B. 1；$x_2=2x_1+1$　　C. 0；$x_2=2x_1-1$　　D. 1；$x_2=2x_1-1$

7. 对每个正整数 n，在实数轴上定义 $f_n(t)=n\mathbf{1}_{[0,n^{-2}]}(t)$。按全实数轴上的积分定义函数范数，则 $\big(\|f_n\|_1,\|f_n\|_2,\|f_n\|_\infty\big)$ 等于（　　）

A. $(n^{-2},1,n)$　　B. (n^{-1},n,n)　　C. $(n^{-1},1,n)$　　D. $(n,1,n^{-1})$

8. 对损失 $L(w)=(w-1)^2$，从 $w_0=0$ 出发，用学习率 $\eta=\frac{6}{5}$ 作一次梯度下降。新参数与新损失分别为（　　）

A. $\frac{12}{5}$，$\frac{49}{25}$　　B. $\frac{6}{5}$，$\frac{1}{25}$　　C. $-\frac{12}{5}$，$\frac{49}{25}$　　D. $\frac{12}{5}$，$\frac{12}{5}$

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. 关于数据、训练信号和算法，下列说法正确的有（　　）

A. 无监督学习可以在未提供外部标签的数据中寻找聚类或其他结构。

B. 自监督学习可以从原始数据本身构造预测目标，而不必逐条人工标注该目标。

C. GAN 的判别器使用“真实 / 生成”训练信号，说明每个真实样本必须预先人工标注任务类别。

D. 反向传播用于高效计算梯度；梯度下降等优化方法再根据梯度更新参数。

10. 对 $x=(x_1,x_2,x_3)^\top$，定义循环移位 $Sx=(x_3,x_1,x_2)^\top$ 及共享系数映射 $Cx=(x_1+2x_3,x_2+2x_1,x_3+2x_2)^\top$。下列说法正确的有（　　）

A. 对所有 x，$C(Sx)=S(Cx)$，体现平移等变。

B. 对所有 x，$C(Sx)=C(x)$，所以 C 本身平移不变。

C. 对所有 x，$\sum_i (C(Sx))_i=\sum_i (Cx)_i$。

D. 自回归掩码禁止当前位置使用未来位置，是信息访问约束；不等于已经发现统计因果关系。

11. 关于范数、优化和稳定性，下列说法正确的有（　　）

A. 一个可微损失在某点的梯度为零，就说明该点一定是全局最小化点。

B. 对任意有限维实向量 x，均有 $\|x\|_1\ge\|x\|_2\ge\|x\|_\infty$。

C. 表达式 $\|x\|_{1/2}=(\sum_i |x_i|^{1/2})^2$ 对所有向量都满足三角不等式。

D. 残差连接或归一化本身，不足以保证每一层都使任意输入扰动严格收缩。

三、填空题：共 4 小题，每小题 5 分，共 20 分。

12. 设 $A=\begin{pmatrix}2 & a\\ b & 1\end{pmatrix}$。若对所有 $x\in\mathbb{R}^2$ 都有 $x^\top Ax=2x_1^2+6x_1x_2+x_2^2$，则 $a+b=$ _____；若再要求 $A=A^\top$，则 $(a,b)=$ _____。

13. 随机变量 Z 的真实分布满足 $P(Z=0)=P(Z=2)=\frac{1}{2}$，实际观测为多重集 $\{0,0,2\}$。则 $P_{\mathcal{D}}(Z=2)=$ _____，$\mathbb{E}_{P_{\mathcal{D}}}[(Z-1)^2]=$ _____。

14. 某网络为 $h=\rho(W_1x+b_1)$、$y=W_2h+b_2$，其中 $x\in\mathbb{R}^4$，隐藏层宽度为 3，输出维度为 2。若全部权重与偏置均独立训练，展平成参数向量 $\theta\in\mathbb{R}^p$，则 W_2 的形状为 _____，$p=$ _____。

15. 设 $L(t)=(t-2)^2$，允许的参数空间为 $\Theta=[-1,1]$。在此参数空间内，L 的最小值为 _____，最小化点集合为 _____。

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）

输入 $x=(x_1,x_2)^\top\in\{0,1\}^2$。两分量不同时标签为 $+1$，相同时为 -1；约定 $\operatorname{sign}(z)=+1$ 当 $z\ge 0$，否则为 -1。

（1）按 $(0,0),(1,0),(0,1),(1,1)$ 的顺序，写出 $\operatorname{sign}(x_1+x_2-\frac{1}{2})$ 的预测，并求错误个数。（2 分）

（2）用不等式证明：不存在单个感知机 $\operatorname{sign}(w_1x_1+w_2x_2+b)$ 将四点全部正确分类。（4 分）

（3）令 $h_1=\rho(x_1-x_2)$、$h_2=\rho(x_2-x_1)$。求使 $\operatorname{sign}(ah_1+ah_2+c)$ 全部分类正确的所有 (a,c)，并说明条件既必要又充分。（4 分）

（4）若去掉上述隐藏层的两个 ReLU，是否还能选择 a,c 完成同一分类？说明原因。（2 分）

17.（12 分）

令 $T(t)=2\rho(t)-4\rho(t-\frac{1}{2})$ $(t\in[0,1])$，T_s 为 T 复合 s 次。令 P_m 是经过各点 $(j2^{-m},(j2^{-m})^2)$ $(j=0,\dots,2^m)$ 的连续分段线性插值，$P_0(t)=t$。

可直接使用：T_s 在每个区间 $[j2^{-(s-1)},(j+1)2^{-(s-1)}]$ 的两端为 0、中点为 1，并在两个半区间分别为仿射函数。

（1）设 $a<b$，连接 (a,a^2) 与 (b,b^2) 的直线为 $L_{a,b}(t)$。证明 $L_{a,b}(t)-t^2=(t-a)(b-t)$，并求区间 $[a,b]$ 上的最大误差。（3 分）

（2）对 $m\ge 1$，证明存在与 t 无关的系数 c_m，使 $P_{m-1}(t)-P_m(t)=c_mT_m(t)$，并求 c_m。（4 分）

（3）据此把 $P_m(t)$ 写成 t 与 $T_1(t),\dots,T_m(t)$ 的线性组合，求 $\|P_m-t^2\|_{\infty,[0,1]}$，并求误差不超过 2^{-10} 的最小非负整数 m。（2 分）

（4）若每步保留当前 $T_s(t)$ 和 $P_s(t)$，说明为什么可以只用 $O(m)$ 个 ReLU 单元及 $O(m)$ 层实现 P_m，从而得到误差 $\epsilon\to 0$ 时的 $O(\log(1/\epsilon))$ 构造。该结论是否已经证明了任意连续函数的通用逼近定理？（3 分）

18.（12 分）

先考虑 $\Lambda(w,b)=(w-2)^2+(b-1)^2$。再考虑一般可微函数 $F:\mathbb{R}^p\to\mathbb{R}$，最小值 F^* 可达到。已知存在 $0<\mu\le L_s$，对所有 θ, v 满足：

$$F(\theta+v)\le F(\theta)+\nabla F(\theta)^\top v+\frac{L_s}{2}\|v\|_2^2,$$

$$\frac{1}{2}\|\nabla F(\theta)\|_2^2\ge\mu\big(F(\theta)-F^*\big).$$

第一条为本题给定的光滑上界，第二条称为 PL 条件；不要求证明这两个已知条件。

(1) 求 $\nabla\Lambda(w,b)$。证明对 $\theta^+=(w,b)^\top-\eta\nabla\Lambda$，在全部非最优点严格下降的学习率恰为 $0<\eta<1$，并求一步到最优点的学习率。(3 分)

(2) 对一般 F，取 $v=-\eta\nabla F(\theta)$，推导一次更新的损失变化上界；给出确保非零梯度点下降的学习率充分范围。(3 分)

(3) 取固定学习率 $\eta=1/L_s$，证明 $F(\theta_k)-F^*\le(1-\mu/L_s)^k[F(\theta_0)-F^*]$。解释这里“线性收敛”为什么表现为几何衰减。(4 分)

(4) 在 PL 条件下，驻点是否一定为全局最优点？不具备 PL 条件的一般可微损失能否沿用该结论？用 $G(t)=(t^2-1)^2$ 说明。(2 分)

19.（12 分）

设 $\mathcal{H}=\{h_1,\dots,h_J\}$ 是抽样前固定的有限模型类。训练样本 (X_i,Y_i) 独立同分布于 P，损失 $\ell(Y,h(X))\in[0,1]$。定义

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

可直接使用：对每个预先固定的 h，$\Pr(|\widehat{R}(h)-R(h)|>\epsilon)\le 2e^{-2M\epsilon^2}$；以及并集界 $\Pr(\bigcup_j E_j)\le\sum_j\Pr(E_j)$。这里概率针对训练集的随机抽样。

(1) 证明 $\Pr(\max_{h\in\mathcal{H}}|\widehat{R}(h)-R(h)|>\epsilon)\le 2Je^{-2M\epsilon^2}$。(3 分)

(2) 设 $0<\delta<1$，求一个 $\epsilon(M,J,\delta)$，使全部模型的风险偏差同时不超过 ϵ 的概率至少为 $1-\delta$。(3 分)

(3) 设 $\widehat{h}\in\operatorname{argmin}_{h\in\mathcal{H}}\widehat{R}(h)$，$h^*\in\operatorname{argmin}_{h\in\mathcal{H}}R(h)$。在前问的“好事件”上证明 $R(\widehat{h})\le R(h^*)+2\epsilon$。(4 分)

(4) 为使上述超额风险上界不超过 $a>0$，写出关于 M 的一个充分条件；说明为什么不能仅把单个固定模型的概率界直接用于训练后挑出的 $\widehat{h}$。(2 分)

20.（12 分）

设 $Q_A(x) = x^\top Ax$，限制 $A = UV^\top$，$U, V \in \mathbb{R}^{d\times r}$。实际存储参数量按标量个数计。另给定 $D = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$，用 $B = uv^\top$（$u, v \in \mathbb{R}^2$，允许零向量）近似它。

定义 Frobenius 范数 $\|E\|_F = (\sum_{i,j} E_{ij}^2)^{1/2}$；$uv^\top$ 的秩至多为 1，意为列空间维数至多为 1。

(1) 证明 $Q_A(x) = (U^\top x)^\top (V^\top x)$，标明中间向量维度，并在 $d = 1000, r = 10$ 时比较直接存储与因子存储的参数量。(3 分)

(2) 利用 Cauchy–Schwarz 不等式证明：对任意矩阵 E 与单位向量 z，$\|Ez\|_2 \le \|E\|_F$。(3 分)

(3) 证明任意 $B = uv^\top$ 都满足 $\|D - B\|_F \ge 1$，并构造达到等号的 B。这是二维情形下的最佳秩一近似。(4 分)

(4) 对非零实数 α，比较 (u, v) 与 $(\alpha u, \alpha^{-1} v)$。结合前问，说明低秩参数化能否同时自动保证因子唯一与无损表达。(2 分)

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

21.（12 分）

考虑标量残差链 $x_{k+1} = x_k + F_k(x_k)$，$k = 0, \dots, L-1$，其中 F_k 均可微。捷径与相加后的映射都是恒等映射，所有 x_k 视为初值 x_0 的函数。

(1) 证明 $x_L = x_0 + \sum_{k=0}^{L-1} F_k(x_k)$。(2 分)

(2) 用链式法则，分别从逐层递推和前问求和式出发，推出 $\frac{dx_L}{dx_0}$ 的乘积表达式与“1+ 求和”表达式。解释后者每项中不能遗漏的链式因子。(4 分)

(3) 分别取每层 $F_k(t) = -t$ 与 $F_k(t) = t$，求 dx_L/dx_0（$L \ge 1$），判断恒等捷径是否足以无条件保证梯度既不消失也不爆炸。(3 分)

(4) 改取 $F_k(t) = -\beta t$，$0 < \beta < 1$；若每层另有绝对值不超过 ε 的加性计算误差，初始误差为零，求深度 L 后误差的上界，并说明这次收缩来自什么条件。(3 分)

——全卷结束——