

Deep Learning (WJ64007N): Homework #1

姓名 Name: 张三

学号 Student ID: 2026000000

Due: 2026-10-25 23:59:59

题目 1 (ST Gumbel-Softmax, 20 pts). 设类别概率为 $\pi = (\pi_1, \dots, \pi_K)$ ($\pi_i > 0$), $G_i = -\log(-\log U_i)$, $U_i \sim \mathcal{U}(0, 1)$ 相互独立。ST Gumbel-Softmax 中,

$$\mathbf{z} = \text{onehot}\left(\arg\max_i v_i\right), \quad \mathbf{y} = \text{softmax}(\mathbf{v}/\tau), \quad \mathbf{v} = \log \pi + \mathbf{G},$$

其中温度 $\tau > 0$ 。前向使用 $\mathbf{z}$, 反向用 $\mathbf{y}$ 的梯度代替 $\mathbf{z}$ 的梯度。

- (a) 求 G_i 的分布函数, 并证明 $P(\arg\max_i v_i = k) = \pi_k$ 。
- (b) 证明 $\frac{\partial y_i}{\partial v_j} = \frac{1}{\tau} y_i (\delta_{ij} - y_j)$ (δ_{ij} 为 Kronecker 符号), 并证明 $\tau \rightarrow 0$ 时 $\mathbf{y} \rightarrow \mathbf{z}$ (设 $\mathbf{v}$ 的最大分量唯一)。
- (c) 说明 $\mathbf{z} = \mathbf{y_hard} - \mathbf{y_soft.detach()} + \mathbf{y_soft}$ ($\mathbf{y_hard}$ 即 $\mathbf{z}$, $\mathbf{y_soft}$ 即 $\mathbf{y}$) 的前向值与反向梯度; 写出上游梯度为 $\mathbf{c}$ 时 ST 给出的 $\partial \mathcal{L} / \partial v_j$ 。若 $\pi = \text{softmax}(\mathbf{s})$, $\mathcal{L} = \mathbf{c}^\top \mathbf{z}$ ($\mathbf{c}$ 为常向量), ST 对 s_j 的梯度与上式相同, 它是 $\partial \mathbb{E}[\mathcal{L}] / \partial s_j$ 的无偏估计吗?

解答.

题目 2 (PL 条件, 15 pts). 称 $\mathcal{L}$ 满足 μ -PL 条件 ($\mu > 0$), 若对所有 $\mathbf{w}$ 都有 $\frac{1}{2} \|\nabla \mathcal{L}(\mathbf{w})\|^2 \geq \mu(\mathcal{L}(\mathbf{w}) - \mathcal{L}^*)$, 其中 $\mathcal{L}^* = \min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$ 。

- (a) 设 $f(x) = x^2 + 3 \sin^2 x$ 。证明 f 不是凸函数, 并通过数值计算说明 f 满足 PL 条件, 估计使其成立的最大常数 μ 。
- (b) 设模型有 p 个参数, 在 n 个训练样本上的输出为 $F(\mathbf{w}) \in \mathbb{R}^n$, 平方损失 $\mathcal{L}(\mathbf{w}) = \frac{1}{2} \|F(\mathbf{w}) - \mathbf{y}\|^2$ 。记 $J = \partial F / \partial \mathbf{w} \in \mathbb{R}^{n \times p}$, $K = JJ^\top$ 。证明: 若 $\lambda_{\min}(K(\mathbf{w})) \geq \mu > 0$ 对所有 $\mathbf{w}$ 成立, 则 $\mathcal{L}$ 满足 μ -PL; 并说明 $p < n$ 时这一条件不可能成立。
- (c) 设 $\mathcal{L} \geq 0$, $\mathcal{L}^* = 0$, $\mathcal{L}$ 是 L -光滑的且满足 μ -PL, GD 的步长为 $1/L$ 。证明对任意 T , $\|\mathbf{w}_T - \mathbf{w}_0\| \leq 2\sqrt{2L\mathcal{L}(\mathbf{w}_0)}/\mu$, 并证明 $\{\mathbf{w}_t\}$ 收敛到 $\mathcal{L}$ 的某个全局最小点。

解答.

题目 3 (He 初始化, 15 pts). 考虑无偏置的 L 层 ReLU 网络 $\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{x}^{(l-1)}$, $\mathbf{x}^{(l)} = \text{ReLU}(\mathbf{z}^{(l)})$, 第 l 层宽度为 n_l (n_0 为输入维数)。 $\mathbf{W}^{(l)}$ 的元素独立同分布, 均值为 0、方差为 σ_l^2 , 分布关于 0 对称, 且与 $\mathbf{x}^{(l-1)}$ 独立。

- 如果把每一层的权重都初始化为同一个常数 (例如全为 0), 训练时会出现什么问题?
- 证明 $\text{Var}(z_i^{(l)}) = \frac{1}{2} n_{l-1} \sigma_l^2 \text{Var}(z_j^{(l-1)})$ ($l \geq 2$), 并由此说明 ReLU 网络应取 $\sigma_l^2 = 2/n_{l-1}$ 。比较 $L = 50$ 、每层宽度为 n 时取 $\sigma^2 = 1/n$ 与 $\sigma^2 = 2/n$ 的 $\text{Var}(z^{(L)})/\text{Var}(z^{(1)})$ 。
- 记 $\delta^{(l)} = \partial \mathcal{L} / \partial \mathbf{z}^{(l)}$ 。假设 $\mathbf{W}^{(l)}$ 与 $\delta^{(l)}$ 独立, $\delta_i^{(l)}$ 同分布且均值为 0, $\text{ReLU}'(z_j^{(l-1)})$ 以各 $\frac{1}{2}$ 的概率取 0 或 1 且与其余量独立。证明 $\text{Var}(\delta_j^{(l-1)}) = \frac{1}{2} n_l \sigma_l^2 \text{Var}(\delta_i^{(l)})$, 并与 (b) 的条件比较。

解答.

题目 4 (反向传播计算, 15 pts). 考虑 2 个输入、2 个隐藏单元 (ReLU)、1 个输出的网络与平方损失:

$$\mathbf{z} = \mathbf{W}_1 \mathbf{x} + \mathbf{b}_1, \quad \mathbf{h} = \text{ReLU}(\mathbf{z}), \quad \hat{y} = \mathbf{w}_2^\top \mathbf{h} + b_2, \quad \mathcal{L} = \frac{1}{2}(\hat{y} - y)^2,$$

其中 $\mathbf{W}_1 = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$, $\mathbf{b}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\mathbf{w}_2 = \begin{pmatrix} 1.5 \\ -2 \end{pmatrix}$, $b_2 = 0.5$, 样本为 $\mathbf{x} = (1, -1)^\top$, $y = 3$ 。

- 计算 $\mathbf{z}$ 、 $\mathbf{h}$ 、 $\hat{y}$ 和 $\mathcal{L}$ 。
- 计算 $\partial \mathcal{L} / \partial \mathbf{w}_2$ 、 $\partial \mathcal{L} / \partial b_2$ 、 $\partial \mathcal{L} / \partial \mathbf{W}_1$ 、 $\partial \mathcal{L} / \partial \mathbf{b}_1$ (与参数同形状)。哪些梯度为零? 为什么?
- 取学习率 $\eta = 0.1$, 用 (b) 的梯度同时更新所有参数, 重新计算损失。

解答.

题目 5 (STE 与 REINFORCE, 15 pts).

- 设 $\mathcal{L}(x) = (\text{round}(x) - 1)^2$, 当前 $x = 0.3$ 。求真实导数 $d\mathcal{L}/dx$ 和 STE (前向用 round, 反向令 $d\text{round}(x)/dx := 1$) 给出的梯度; 用学习率 $\eta = 0.15$ 按 STE 梯度做一步梯度下降, 求新的 x 和损失。
- 设 $z \in \{0, 1\}$, $P(z = 1) = p = \sigma(\theta)$, σ 为 sigmoid 函数。证明 $\frac{\partial}{\partial \theta} \log p_\theta(z) = z - p$ ($p_\theta(z)$ 为 z 的概率); 并证明对任意 $f: \{0, 1\} \rightarrow \mathbb{R}$ 与常数 b (均与 θ, z 无关), $\hat{g}_b = (f(z) - b)(z - p)$ 都是 $\frac{d}{d\theta} \mathbb{E}[f(z)]$ 的无偏估计。
- 取 $\theta = 0$, $f(0) = 99$, $f(1) = 101$, 分别计算 $b = 0$ 与 $b = 100$ 时 $\hat{g}_b$ 的方差。对一般的 p 、 $f(0)$ 、 $f(1)$, 求使方差最小的 b 。

解答.

题目 6 (SGD 的收敛性, 20 pts). 设 $\mathcal{L}$ 是 L -光滑的, 下界为 $\mathcal{L}^*$ 。SGD 的迭代为 $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$, 其中随机梯度无偏 ($\mathbb{E}[\mathbf{g}_t | \mathbf{w}_t] = \nabla \mathcal{L}(\mathbf{w}_t)$), 且满足

$$\mathbb{E}[\|\mathbf{g}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|^2 | \mathbf{w}_t] \leq \sigma^2.$$

(a) 证明 $\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1}) | \mathbf{w}_t] \leq \mathcal{L}(\mathbf{w}_t) - \eta(1 - \frac{L\eta}{2})\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \frac{L\eta^2}{2}\sigma^2$ 。

(b) 设 $\eta \leq 1/L$, 证明

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 \leq \frac{2(\mathcal{L}(\mathbf{w}_0) - \mathcal{L}^*)}{\eta T} + L\eta\sigma^2.$$

(c) 固定 η 而 $T \rightarrow \infty$ 时, 上界趋于什么? 若对每个 T 取 $\eta = 1/\sqrt{T}$ (设 $T \geq L^2$, 迭代记为 $\mathbf{w}_t^{(T)}$), 收敛速度是多少? 并证明 $\lim_{T \rightarrow \infty} \min_{0 \leq t \leq T-1} \mathbb{E} \|\nabla \mathcal{L}(\mathbf{w}_t^{(T)})\|^2 = 0$ 。

解答.