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 $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$（$\pi_i > 0$），$G_i = -\log(-\log U_i)$，$U_i \sim \mathcal{U}(0,1)$ 相互独立。ST Gumbel-Softmax 中，

$$\boldsymbol{z} = \text{onehot}\Big(\operatorname*{argmax}_i v_i\Big), \qquad \boldsymbol{y} = \text{softmax}(\boldsymbol{v}/\tau), \qquad \boldsymbol{v} = \log \boldsymbol{\pi} + \boldsymbol{G},$$

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

(a) 求 G_i 的分布函数，并证明 $P\big(\operatorname{argmax}_i v_i = k\big) = \pi_k$。

(b) 证明 $\dfrac{\partial y_i}{\partial v_j} = \dfrac{1}{\tau} y_i(\delta_{ij} - y_j)$（$\delta_{ij}$ 为 Kronecker 符号），并证明 $\tau \to 0$ 时 $\boldsymbol{y} \to \boldsymbol{z}$（设 $\boldsymbol{v}$ 的最大分量唯一）。

(c) 说明 `z = y_hard - y_soft.detach() + y_soft`（`y_hard` 即 $\boldsymbol{z}$，`y_soft` 即 $\boldsymbol{y}$）的前向值与反向梯度；写出上游梯度为 $\boldsymbol{c}$ 时 ST 给出的 $\partial\mathcal{L}/\partial v_j$。若 $\boldsymbol{\pi} = \text{softmax}(\boldsymbol{s})$、$\mathcal{L} = \boldsymbol{c}^\top \boldsymbol{z}$（$\boldsymbol{c}$ 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. **(a)** $G \le g \iff -\log U \ge e^{-g} \iff U \le \exp(-e^{-g})$，所以 $F(g) = \exp(-e^{-g})$，$F'(g) = e^{-g}\exp(-e^{-g})$。第 k 个分量最大，等价于对所有 $i \ne k$ 有 $G_i < G_k + \log\pi_k - \log\pi_i$。对 $G_k = g$ 取条件，

$$\begin{aligned} P\big(\operatorname*{argmax}_i v_i = k\big) &= \int_{-\infty}^{\infty} F'(g) \prod_{i\ne k} F\big(g + \log\pi_k - \log\pi_i\big)\,\mathrm{d}g \\ &= \int_{-\infty}^{\infty} e^{-g} \exp\Big(-e^{-g}\sum_i \frac{\pi_i}{\pi_k}\Big)\,\mathrm{d}g, \end{aligned}$$

其中用到 $F(g + \log\pi_k - \log\pi_i) = \exp(-e^{-g}\pi_i/\pi_k)$。由 $\sum_i \pi_i = 1$，换元 $u = e^{-g}$ 得 $\int_0^\infty e^{-u/\pi_k}\,\mathrm{d}u = \pi_k$。

(b) 记 $S = \sum_l e^{v_l/\tau}$，则 $y_i = e^{v_i/\tau}/S$。由商的求导法则，$i = j$ 时 $\partial y_i/\partial v_i = \frac{1}{\tau}(y_i - y_i^2)$，$i \ne j$ 时 $\partial y_i/\partial v_j = -\frac{1}{\tau} y_i y_j$，合并即得 $\partial y_i/\partial v_j = \frac{1}{\tau} y_i(\delta_{ij} - y_j)$。设最大分量下标为 k，即 $v_k > v_j$，$j \ne k$，则 $y_i = e^{(v_i - v_k)/\tau}/(1 + \sum_{j\ne k} e^{(v_j - v_k)/\tau})$。由于 $j \ne k$ 时 $v_j - v_k < 0$，故 $\tau \to 0$ 时 $e^{(v_j - v_k)/\tau} \to 0$。因此 $y_k \to 1$，而对于 $i \ne k$ 有 $y_i \to 0$，即 $\boldsymbol{y} \to \boldsymbol{z}$。

(c) 前向：`y_soft.detach()` 与 `y_soft` 数值相同，相互抵消，`z` 等于 `y_hard`。反向：`y_hard` 由 argmax 得到，没有梯度；`detach()` 的部分不参与求导；只剩 `y_soft`，所以 $\partial\boldsymbol{z}/\partial\boldsymbol{v}$ 被

替换为 $\partial \boldsymbol{y}/\partial \boldsymbol{v}$。由 (b)，

$$\frac{\partial \mathcal{L}}{\partial v_j} \approx \sum_i c_i \frac{\partial y_i}{\partial v_j} = \frac{1}{\tau} y_j \Big(c_j - \sum_i c_i y_i \Big).$$

（$\boldsymbol{v} = \boldsymbol{s} - \log \sum_l e^{s_l} + \boldsymbol{G}$，$\partial v_i/\partial s_j = \delta_{ij} - \pi_j$，而 $\sum_i \partial \mathcal{L}/\partial v_i = 0$，所以对 s_j 的梯度与对 v_j 的相同。）

真实梯度：$\mathbb{E}[\mathcal{L}] = \sum_i \pi_i c_i$，$\frac{\partial}{\partial s_j}\mathbb{E}[\mathcal{L}] = \pi_j\big(c_j - \sum_i c_i \pi_i\big)$。ST 估计一般**不是**无偏的。反例：$K = 2$，$\boldsymbol{s} = (0,0)$（$\boldsymbol{\pi} = (\frac{1}{2}, \frac{1}{2})$），$\boldsymbol{c} = (1,0)$，$\tau = 1$，真实梯度为 $\frac{1}{4}$。记 $X = v_1 - v_2 = G_1 - G_2$，ST 估计为 $y_1(1-y_1) = \sigma'(X)$（σ 为 sigmoid）。由 (a)（取 $K = 2$），$P(X < a) = \sigma(a)$，即 X 的密度为 $\sigma'(x)$，于是

$$\mathbb{E}\big[\sigma'(X)\big] = \int_{-\infty}^{\infty} \sigma'(x)^2 \, \mathrm{d}x \overset{u=\sigma(x)}{=} \int_0^1 u(1-u) \, \mathrm{d}u = \frac{1}{6} \neq \frac{1}{4}.$$

事实上，ST 估计的期望是松弛目标 $\mathbb{E}[\boldsymbol{c}^\top \boldsymbol{y}]$ 的梯度，而不是 $\mathbb{E}[\boldsymbol{c}^\top \boldsymbol{z}]$ 的梯度。$\tau \to 0$ 时偏差趋于零（数值上 $\tau = 0.1$ 时期望约 0.248），但方差约按 $1/\tau$ 增长（$\tau = 0.1$ 时约 0.35，$\tau = 0.02$ 时约 2.0），这就是课上 “τ 小：偏差小但方差大” 的权衡。

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

(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 $F(\boldsymbol{w}) \in \mathbb{R}^n$，平方损失 $\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\|F(\boldsymbol{w}) - \boldsymbol{y}\|^2$。记 $J = \partial F/\partial \boldsymbol{w} \in \mathbb{R}^{n \times p}$，$K = JJ^\top$。证明：若 $\lambda_{\min}(K(\boldsymbol{w})) \geq \mu > 0$ 对所有 $\boldsymbol{w}$ 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，$\|\boldsymbol{w}_T - \boldsymbol{w}_0\| \leq 2\sqrt{2L\,\mathcal{L}(\boldsymbol{w}_0)}/\mu$，并证明 $\{\boldsymbol{w}_t\}$ 收敛到 $\mathcal{L}$ 的某个全局最小点。

解答. **(a)** $f''(x) = 2 + 6\cos 2x$，$f''(\pi/2) = -4 < 0$，所以 f 不是凸函数。在 $[-10, 10]$ 上取细网格计算 $\frac{1}{2}f'(x)^2/f(x)$（$x \neq 0$）：$x \to 0$ 时趋于 8，$|x| \to \infty$ 时趋于 2，最小值约为 0.175，在 $x \approx \pm 2.2$ 处取到。$|x| \geq 10$ 时 $|f'(x)| \geq 2|x| - 3 \geq 1.7|x|$，$f(x) \leq x^2 + 3 \leq 1.03x^2$，比值 ≥ 1.40，不会更小。所以使 PL 成立的最大常数为 $\mu \approx 0.1755$（Karimi et al. 给出的 $\mu = 1/32$ 也成立，但更保守）。

(b) 记 $\boldsymbol{r} = F(\boldsymbol{w}) - \boldsymbol{y}$，由链式法则 $\nabla \mathcal{L} = J^\top \boldsymbol{r}$，于是

$$\|\nabla \mathcal{L}\|^2 = \boldsymbol{r}^\top JJ^\top \boldsymbol{r} = \boldsymbol{r}^\top K \boldsymbol{r} \;\geq\; \lambda_{\min}(K)\,\|\boldsymbol{r}\|^2 \;\geq\; \mu\,\|\boldsymbol{r}\|^2 = 2\mu\,\mathcal{L}(\boldsymbol{w}),$$

即 $\frac{1}{2}\|\nabla \mathcal{L}\|^2 \geq \mu \mathcal{L}(\boldsymbol{w}) \geq \mu\big(\mathcal{L}(\boldsymbol{w}) - \mathcal{L}^*\big)$（因为 $\mathcal{L}^* \geq 0$）。又 $\operatorname{rank} K \leq \operatorname{rank} J \leq p$；若 $p < n$，$n \times n$ 的 K 不满秩，$\lambda_{\min}(K) = 0$，条件不可能成立。

(c) 由下降引理与 $\mathcal{L} \geq 0$，$\|\nabla \mathcal{L}(\boldsymbol{w}_t)\|^2 \leq 2L\big(\mathcal{L}(\boldsymbol{w}_t) - \mathcal{L}(\boldsymbol{w}_{t+1})\big) \leq 2L\mathcal{L}(\boldsymbol{w}_t)$，所以 $\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\| = \frac{1}{L}\|\nabla \mathcal{L}(\boldsymbol{w}_t)\| \leq \sqrt{2\mathcal{L}(\boldsymbol{w}_t)/L}$。由 PL 与光滑性的线性收敛 $\mathcal{L}(\boldsymbol{w}_t) \leq$

$(1-\mu/L)^t\mathcal{L}(\boldsymbol{w}_0)$，记 $q=\sqrt{1-\mu/L}$，则

$$\sum_{t=0}^{\infty}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|\le\sqrt{\frac{2\mathcal{L}(\boldsymbol{w}_0)}{L}}\sum_{t=0}^{\infty}q^t=\frac{\sqrt{2\mathcal{L}(\boldsymbol{w}_0)/L}}{1-q}\le\sqrt{\frac{2\mathcal{L}(\boldsymbol{w}_0)}{L}}\cdot\frac{2L}{\mu}=\frac{2\sqrt{2L\mathcal{L}(\boldsymbol{w}_0)}}{\mu},$$

其中用到 $\sqrt{1-a}\le 1-\frac{a}{2}$，即 $1-q\ge\frac{\mu}{2L}$。（这里 $\mu\le L$：由光滑性 $0\le\mathcal{L}(\boldsymbol{w}-\frac{1}{L}\nabla\mathcal{L})\le\mathcal{L}(\boldsymbol{w})-\frac{1}{2L}\|\nabla\mathcal{L}\|^2$，再结合 PL 即得，所以 q 有定义。）由三角不等式即得 $\|\boldsymbol{w}_T-\boldsymbol{w}_0\|\le 2\sqrt{2L\mathcal{L}(\boldsymbol{w}_0)}/\mu$。

收敛：总路程有限，所以对 $m>t$，$\|\boldsymbol{w}_m-\boldsymbol{w}_t\|\le\sum_{s\ge t}\|\boldsymbol{w}_{s+1}-\boldsymbol{w}_s\|\to 0$（$t\to\infty$），$\{\boldsymbol{w}_t\}$ 是 Cauchy 列，收敛到某个 $\boldsymbol{w}_\infty$。$\mathcal{L}$ 连续且 $\mathcal{L}(\boldsymbol{w}_t)\to 0$，故 $\mathcal{L}(\boldsymbol{w}_\infty)=0=\mathcal{L}^*$，$\boldsymbol{w}_\infty$ 是全局最小点，且离 $\boldsymbol{w}_0$ 不超过 $2\sqrt{2L\mathcal{L}(\boldsymbol{w}_0)}/\mu$。

题目 3 (He 初始化，15 pts). 考虑无偏置的 L 层 ReLU 网络 $\boldsymbol{z}^{(l)}=\boldsymbol{W}^{(l)}\boldsymbol{x}^{(l-1)}$，$\boldsymbol{x}^{(l)}=\mathrm{ReLU}(\boldsymbol{z}^{(l)})$，第 l 层宽度为 n_l（n_0 为输入维数）。$\boldsymbol{W}^{(l)}$ 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 $\boldsymbol{x}^{(l-1)}$ 独立。

(a) 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?

(b) 证明 $\mathrm{Var}(z_i^{(l)})=\frac{1}{2}\,n_{l-1}\,\sigma_l^2\,\mathrm{Var}(z_j^{(l-1)})$（$l\ge 2$），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 $\mathrm{Var}(z^{(L)})/\mathrm{Var}(z^{(1)})$。

(c) 记 $\boldsymbol{\delta}^{(l)}=\partial\mathcal{L}/\partial\boldsymbol{z}^{(l)}$。假设 $\boldsymbol{W}^{(l)}$ 与 $\boldsymbol{\delta}^{(l)}$ 独立，$\delta_i^{(l)}$ 同分布且均值为 0，$\mathrm{ReLU}'(z_j^{(l-1)})$ 以各 $\frac{1}{2}$ 的概率取 0 或 1 且与其余量独立。证明 $\mathrm{Var}(\delta_j^{(l-1)})=\frac{1}{2}\,n_l\,\sigma_l^2\,\mathrm{Var}(\delta_i^{(l)})$，并与 (b) 的条件比较。

解答. **(a)** 同一层的所有隐藏单元计算相同的输出、得到相同的梯度，更新后仍然相同，永远无法区分，相当于每层只有一个神经元（对称性无法打破）。若全为 0，ReLU 网络的所有梯度都为零，参数根本不会更新。所以必须随机初始化。

(b) $z_i^{(l)}=\sum_j W_{ij}x_j$，各项均值为零且互不相关（$\mathbb{E}[W_{ij}]=0$，$W$ 与 $\boldsymbol{x}$ 独立），故 $\mathrm{Var}(z_i^{(l)})=\sum_j\mathbb{E}[W_{ij}^2]\mathbb{E}[x_j^2]=n_{l-1}\sigma_l^2\,\mathbb{E}[x_j^2]$。$z_j^{(l-1)}$ 关于 0 对称（W 对称），所以 $\mathbb{E}[z^2\mathbf{1}\{z>0\}]=\mathbb{E}[z^2\mathbf{1}\{z<0\}]$，$\mathbb{E}[x_j^2]=\mathbb{E}[z_j^2\mathbf{1}\{z_j>0\}]=\frac{1}{2}\mathbb{E}[z_j^2]=\frac{1}{2}\mathrm{Var}(z_j^{(l-1)})$。代入即得。要使方差逐层不变，需要 $\frac{1}{2}n_{l-1}\sigma_l^2=1$，即 $\sigma_l^2=2/n_{l-1}$。取 $\sigma^2=c/n$ 时比值为 $(c/2)^{L-1}$：$c=1$ 为 $2^{-49}\approx 1.8\times 10^{-15}$，信号几乎消失；$c=2$ 为 1。

(c) 由链式法则 $\delta_j^{(l-1)}=b_j\,u_j$，其中 $b_j=\mathrm{ReLU}'(z_j^{(l-1)})$，$u_j=\sum_i W_{ij}^{(l)}\delta_i^{(l)}$。由 $\mathbb{E}[W_{ij}]=0$ 及独立性，$\mathbb{E}[u_j]=0$，且交叉项 $\mathbb{E}[W_{ij}W_{kj}]=0$（$i\ne k$），所以 $\mathbb{E}[u_j^2]=n_l\sigma_l^2\mathrm{Var}(\delta_i^{(l)})$。又 $b_j^2=b_j$，$\mathbb{E}[b_j^2]=\frac{1}{2}$，故 $\mathrm{Var}(\delta_j^{(l-1)})=\mathbb{E}[b_j^2]\mathbb{E}[u_j^2]=\frac{1}{2}n_l\sigma_l^2\mathrm{Var}(\delta_i^{(l)})$。反向保持方差需要 $\sigma_l^2=2/n_l=2/n_{\text{out}}$，而前向需要 $2/n_{l-1}=2/n_{\text{in}}$，两者只在 $n_{\text{in}}=n_{\text{out}}$ 时一致。若只满足前向条件，反向每层因子为 n_l/n_{l-1}，逐层相乘后只剩 n_L/n_0，不会产生深度相关的指数爆炸或消失。因此在宽度相近的深层 ReLU 网络中，He 初始化能够同时较好地保持前向和反向尺度稳定。

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：

$$\boldsymbol{z} = W_1\boldsymbol{x} + \boldsymbol{b}_1, \qquad \boldsymbol{h} = \mathrm{ReLU}(\boldsymbol{z}), \qquad \hat{y} = \boldsymbol{w}_2^{\top}\boldsymbol{h} + b_2, \qquad \mathcal{L} = \tfrac{1}{2}(\hat{y} - y)^2,$$

其中 $W_1 = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$，$\boldsymbol{b}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$，$\boldsymbol{w}_2 = \begin{pmatrix} 1.5 \\ -2 \end{pmatrix}$，$b_2 = 0.5$，样本为 $\boldsymbol{x} = (1, -1)^{\top}$，$y = 3$。

(a) 计算 $\boldsymbol{z}$、$\boldsymbol{h}$、$\hat{y}$ 和 $\mathcal{L}$。

(b) 计算 $\partial\mathcal{L}/\partial\boldsymbol{w}_2$、$\partial\mathcal{L}/\partial b_2$、$\partial\mathcal{L}/\partial W_1$、$\partial\mathcal{L}/\partial\boldsymbol{b}_1$（与参数同形状）。哪些梯度为零？为什么？

(c) 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. **(a)** $\boldsymbol{z} = \begin{pmatrix} 2 - 1 + 0 \\ 1 - 3 + 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$，$\boldsymbol{h} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$，$\hat{y} = 1.5 \times 1 + (-2) \times 0 + 0.5 = 2$，$\mathcal{L} = \frac{1}{2}(2 - 3)^2 = 0.5$。

(b) $\dfrac{\partial\mathcal{L}}{\partial\hat{y}} = \hat{y} - y = -1$。于是

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}_2} = (\hat{y} - y)\,\boldsymbol{h} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \qquad \frac{\partial\mathcal{L}}{\partial b_2} = -1, \qquad \frac{\partial\mathcal{L}}{\partial\boldsymbol{h}} = (\hat{y} - y)\,\boldsymbol{w}_2 = \begin{pmatrix} -1.5 \\ 2 \end{pmatrix}.$$

经过 ReLU：$\dfrac{\partial\mathcal{L}}{\partial\boldsymbol{z}} = \dfrac{\partial\mathcal{L}}{\partial\boldsymbol{h}} \odot \mathrm{ReLU}'(\boldsymbol{z}) = \begin{pmatrix} -1.5 \\ 0 \end{pmatrix}$，所以

$$\frac{\partial\mathcal{L}}{\partial W_1} = \frac{\partial\mathcal{L}}{\partial\boldsymbol{z}}\,\boldsymbol{x}^{\top} = \begin{pmatrix} -1.5 & 1.5 \\ 0 & 0 \end{pmatrix}, \qquad \frac{\partial\mathcal{L}}{\partial\boldsymbol{b}_1} = \begin{pmatrix} -1.5 \\ 0 \end{pmatrix}.$$

为零的有：$\boldsymbol{w}_2$ 的第 2 个分量、W_1 的第 2 行、$\boldsymbol{b}_1$ 的第 2 个分量。原因是第 2 个隐藏单元 $z_2 = -1 < 0$ 未被激活：$h_2 = 0$，对输出没有贡献；且 $\mathrm{ReLU}'(z_2) = 0$，虽然 $\partial\mathcal{L}/\partial h_2 = 2 \neq 0$，梯度也传不回 W_1 的第 2 行（“死亡神经元”）。

(c) 更新后 $W_1 = \begin{pmatrix} 2.15 & 0.85 \\ 1 & 3 \end{pmatrix}$，$\boldsymbol{b}_1 = \begin{pmatrix} 0.15 \\ 1 \end{pmatrix}$，$\boldsymbol{w}_2 = \begin{pmatrix} 1.6 \\ -2 \end{pmatrix}$，$b_2 = 0.6$。重新前向：$\boldsymbol{z} = (1.45, -1)^{\top}$，$\boldsymbol{h} = (1.45, 0)^{\top}$，$\hat{y} = 1.6 \times 1.45 + 0.6 = 2.92$，$\mathcal{L} = \frac{1}{2}(2.92 - 3)^2 = 0.0032$，比原来的 0.5 小得多。

题目 5 (STE 与 REINFORCE，15 pts).

(a) 设 $\mathcal{L}(x) = \big(\mathrm{round}(x) - 1\big)^2$，当前 $x = 0.3$。求真实导数 $\mathrm{d}\mathcal{L}/\mathrm{d}x$ 和 STE（前向用 round，反向令 $\mathrm{d}\,\mathrm{round}(x)/\mathrm{d}x := 1$）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。

(b) 设 $z \in \{0, 1\}$，$P(z = 1) = p = \sigma(\theta)$，$\sigma$ 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 $f : \{0, 1\} \to \mathbb{R}$ 与常数 b（均与 θ、z 无关），$\hat{g}_b = \big(f(z) - b\big)(z - p)$ 都是 $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[f(z)]$ 的无偏估计。

(c) 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. **(a)** round 在 $x = 0.3$ 附近为常数，真实导数为 0，梯度下降停在原地。$\mathcal{L}(0.3) = 1$。STE 令 $\mathrm{d}\,\mathrm{round}(x)/\mathrm{d}x := 1$，梯度为 $2(\mathrm{round}(x) - 1) = -2$。更新 $x \leftarrow 0.3 - 0.15 \times (-2) = 0.6$，$\mathrm{round}(0.6) = 1$，损失降为 0。

(b) $\log p_\theta(z) = z \log \sigma(\theta) + (1 - z) \log(1 - \sigma(\theta))$。由 $\sigma' = \sigma(1 - \sigma)$，$\frac{\partial}{\partial \theta} \log p_\theta(z) = z(1 - p) - (1 - z)p = z - p$。于是

$$\mathbb{E}[\hat{g}_b] = p\big(f(1) - b\big)(1 - p) + (1 - p)\big(f(0) - b\big)(-p) = p(1 - p)\big(f(1) - f(0)\big),$$

与 b 无关。而 $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[f(z)] = \frac{\mathrm{d}}{\mathrm{d}\theta}\big[pf(1) + (1 - p)f(0)\big] = p(1 - p)\big(f(1) - f(0)\big)$，所以无偏。

(c) $p = 0.5$，真实梯度为 0.5。$b = 0$ 时 $\hat{g}_b$ 等概率取 $99 \times (-0.5) = -49.5$ 与 $101 \times 0.5 = 50.5$，方差为 2500；$b = 100$ 时两种情况都等于 0.5，方差为 0。

一般地，$\hat{g}_b$ 的均值与 b 无关，所以最小化方差等价于最小化 $\mathbb{E}[\hat{g}_b^2] = (1 - p)p^2\big(f(0) - b\big)^2 + p(1 - p)^2\big(f(1) - b\big)^2$。对 b 求导令其为零，得

$$b^* = (1 - p)\, f(1) + p\, f(0).$$

代入后 $\hat{g}_{b^*}$ 在 $z = 0$ 与 $z = 1$ 时都等于 $p(1 - p)\big(f(1) - f(0)\big)$，即真实梯度，方差为 0。$p = 0.5$ 时 $b^* = 100$，与上面一致。baseline 不改变期望，却能大幅降低方差。

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 $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta\, \boldsymbol{g}_t$，其中随机梯度无偏（$\mathbb{E}[\boldsymbol{g}_t \mid \boldsymbol{w}_t] = \nabla\mathcal{L}(\boldsymbol{w}_t)$），且满足

$$\mathbb{E}\big[\|\boldsymbol{g}_t - \nabla\mathcal{L}(\boldsymbol{w}_t)\|^2 \mid \boldsymbol{w}_t\big] \le \sigma^2.$$

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

(b) 设 $\eta \le 1/L$，证明

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

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

解答. **(a)** L-光滑意味着 $\mathcal{L}(\boldsymbol{w}') \le \mathcal{L}(\boldsymbol{w}) + \langle\nabla\mathcal{L}(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w}\rangle + \frac{L}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2$。取 $\boldsymbol{w} = \boldsymbol{w}_t$，$\boldsymbol{w}' = \boldsymbol{w}_t - \eta\boldsymbol{g}_t$：

$$\mathcal{L}(\boldsymbol{w}_{t+1}) \le \mathcal{L}(\boldsymbol{w}_t) - \eta\langle\nabla\mathcal{L}(\boldsymbol{w}_t), \boldsymbol{g}_t\rangle + \frac{L\eta^2}{2}\|\boldsymbol{g}_t\|^2.$$

对 $\boldsymbol{w}_t$ 取条件期望。由无偏性，$\mathbb{E}[\langle\nabla\mathcal{L}(\boldsymbol{w}_t), \boldsymbol{g}_t\rangle \mid \boldsymbol{w}_t] = \|\nabla\mathcal{L}(\boldsymbol{w}_t)\|^2$；将 $\boldsymbol{g}_t$ 写成 $\nabla\mathcal{L}(\boldsymbol{w}_t) + (\boldsymbol{g}_t - \nabla\mathcal{L}(\boldsymbol{w}_t))$ 展开，交叉项的条件期望为零，所以 $\mathbb{E}[\|\boldsymbol{g}_t\|^2 \mid \boldsymbol{w}_t] = \|\nabla\mathcal{L}(\boldsymbol{w}_t)\|^2 +$

$\mathbb{E}[\|\boldsymbol{g}_t - \nabla\mathcal{L}(\boldsymbol{w}_t)\|^2 \mid \boldsymbol{w}_t] \le \|\nabla\mathcal{L}(\boldsymbol{w}_t)\|^2 + \sigma^2$。代入整理即得。

(b) $\eta \le 1/L$ 时 $\eta(1 - \frac{L\eta}{2}) \ge \frac{\eta}{2}$。对 (a) 取全期望并移项：$\frac{\eta}{2}\mathbb{E}\,\|\nabla\mathcal{L}(\boldsymbol{w}_t)\|^2 \le \mathbb{E}\mathcal{L}(\boldsymbol{w}_t) - \mathbb{E}\mathcal{L}(\boldsymbol{w}_{t+1}) + \frac{L\eta^2\sigma^2}{2}$。对 $t = 0, \dots, T-1$ 求和，右边前两项相消，且 $\mathbb{E}\mathcal{L}(\boldsymbol{w}_T) \ge \mathcal{L}^*$：

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

两边乘以 $\frac{2}{\eta T}$ 即得。

(c) 固定 η、$T \to \infty$ 时，上界趋于 $L\eta\sigma^2$，不趋于零：SGD 最终在驻点附近一个大小 $\propto \eta$ 的区域内游走（课上的 “噪声球”），这也是需要衰减学习率的原因。取 $\eta = 1/\sqrt{T}$（$T \ge L^2$ 保证 $\eta \le 1/L$），上界为 $\big(2(\mathcal{L}(\boldsymbol{w}_0) - \mathcal{L}^*) + L\sigma^2\big)/\sqrt{T}$，收敛速度为 $O(1/\sqrt{T})$。

最小值不超过平均值，由 (b)（$\eta = 1/\sqrt{T}$）：

$$0 \le \min_{0\le t\le T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{w}_t^{(T)})\right\|^2 \le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{w}_t^{(T)})\right\|^2 \le \frac{2\big(\mathcal{L}(\boldsymbol{w}_0) - \mathcal{L}^*\big) + L\sigma^2}{\sqrt{T}} \xrightarrow{T\to\infty} 0.$$

注意步长必须随 T 减小：若固定 η 且 $\sigma > 0$，结论一般不成立。例如 $\mathcal{L}(w) = \frac{1}{2}w^2$，$g_t = w_t + \xi_t$，$\xi_t \sim \mathcal{N}(0, \sigma^2)$，则 $\mathbb{E}w_{t+1}^2 = (1-\eta)^2\mathbb{E}w_t^2 + \eta^2\sigma^2 \ge \eta^2\sigma^2$，所以对所有 $t \ge 1$ 都有 $\mathbb{E}\,\|\nabla\mathcal{L}(w_t)\|^2 \ge \eta^2\sigma^2 > 0$。注意这只说明梯度趋于零（收敛到驻点），不保证是全局最小；若再加上 PL 条件（题目 2），就能保证收敛到全局最小值。