

第 4 章：联邦学习安全与公平性

思考题 4.1

(a) 由题意, $\Delta = \alpha\rho + \beta$ 为线性模型, 代入两组实验数据:

$$\begin{cases} 0.12 = \alpha \times 0.1 + \beta, \\ 0.36 = \alpha \times 0.3 + \beta. \end{cases}$$

两式相减得:

$$0.36 - 0.12 = \alpha(0.3 - 0.1) \implies 0.24 = 0.2\alpha \implies \alpha = 1.2.$$

代回第一式:

$$\beta = 0.12 - 1.2 \times 0.1 = 0.12 - 0.12 = 0.$$

故线性关系为:

$$\Delta = 1.2\rho.$$

当 $\rho = 0.2$ 时:

$$\Delta = 1.2 \times 0.2 = 0.24.$$

验证: 攻击影响因子 $\Delta = (A_{\text{clean}} - A_{\text{backdoor}})/A_{\text{clean}} = 0.24$ 意味着后门攻击使模型在受攻击样本上的准确率相对下降了 24%, 与线性趋势吻合。

(b) 该线性模型在以下几个方面可能失效:

首先, **饱和效应**。当 ρ 较大时 (如 $\rho \rightarrow 1$), 恶意客户端几乎完全主导全局模型, 后门攻击成功率趋近于理论上限, Δ 的增速必然放缓并趋于饱和, 呈现非线性 (如对数或 sigmoid 型) 增长, 而非无限制线性增长。

其次, **防御机制的干预**。实际系统中, 聚合服务器往往部署异常检测机制 (如 Krum、中位数聚合等)。当 ρ 超过某一阈值, 防御机制被触发后 Δ 的增长规律将发生突变, 线性假设在此区间失效。

第三, **恶意客户端策略的适应性**。恶意客户端可能随着检测压力动态调整投毒强度, 导致 Δ - ρ 关系并非固定斜率。

综上, 线性模型仅在 ρ 较小且无主动防御的场景下具有近似意义, 不宜外推至高恶意比例区间。

思考题 4.2

(a) 拜占庭节点的参数为 $w_j^* = -2\bar{w}$, 其中 $\bar{w} = 1.15$, 故:

$$w_j^* = -2 \times 1.15 = -2.3.$$

聚合规则为:

$$G = \frac{1}{50} \sum_{i=1}^{45} w_i + \beta \sum_{j=1}^5 w_j^*.$$

对期望值, $\mathbb{E}\left[\frac{1}{50}\sum_{i=1}^{45}w_i\right] = \frac{45}{50}\mu = \frac{45}{50} \times 1.2 = 1.08$ 。

拜占庭节点贡献: $\beta\sum_{j=1}^5w_j^* = \beta \times 5 \times (-2\bar{w}) = -0.0089 \times 5 \times 2 \times 1.15$ 。

计算:

$$-0.0089 \times 5 \times 2.3 = -0.0089 \times 11.5 = -0.10235.$$

故:

$$\mathbb{E}[G] = 1.08 + (-0.10235) \approx 0.9777.$$

该结果与目标值 $\mu = 1.2$ 存在偏差 ≈ 0.222 , 说明当 $\beta = -0.0089$ 时, 聚合结果未能恢复到全局目标 $\mu = 1.2$ 。

若要使 $\mathbb{E}[G] = \mu$, 需令补偿系数满足:

$$\frac{45}{50}\mu + \beta \cdot 5 \cdot (-2\mu) = \mu \implies \frac{45}{50} - 10\beta = 1 \implies \beta = \frac{45/50 - 1}{10} = \frac{-0.1}{10} = -0.01.$$

因此 $\beta = -0.0089$ 略有不足, 真正校准的补偿系数应为 $\beta^* = -0.01$ 。

(b) 当 $B = 10$, 诚实客户端减少为 $45 - 5 = 40$ 个, 保持 $\beta = -0.0089$ 不变, 聚合规则变为:

$$G = \frac{1}{50} \sum_{i=1}^{40} w_i + \beta \sum_{j=1}^{10} w_j^*.$$

诚实客户端贡献: $\frac{40}{50}\mu = 0.8 \times 1.2 = 0.96$ 。

拜占庭节点贡献: $-0.0089 \times 10 \times 2 \times \bar{w} = -0.0089 \times 20 \times 1.15 = -0.2047$ 。

$$\mathbb{E}[G] \approx 0.96 - 0.2047 = 0.755.$$

期望值进一步偏离 $\mu = 1.2$, 偏差约 0.445, 显著大于 $B = 5$ 时的情形。这说明固定的 β 补偿系数对拜占庭比例的变化非常敏感, 缺乏自适应能力。

对**收敛速度**的影响: 拜占庭节点将全局参数持续拉向错误方向, 每轮更新都引入系统性偏差, 模型沿错误梯度方向下降, 不仅收敛速度显著下降, 甚至可能完全无法收敛至真实目标, 表现为训练损失震荡或发散。有效的防御手段(如自适应裁剪、鲁棒聚合)需能在不知晓 B 的情况下动态调整聚合策略。

思考题 4.3

(a) 由聚合梯度的构成公式:

$$\bar{\mathbf{g}} = \frac{N-1}{N} \mathbf{g}_{\text{honest}} + \frac{1}{N} (\mathbf{w}^\top \mathbf{x}_t - y_t) \mathbf{x}_t,$$

整理出目标客户端的梯度贡献:

$$(\mathbf{w}^\top \mathbf{x}_t - y_t) \mathbf{x}_t = N\bar{\mathbf{g}} - (N-1)\mathbf{g}_{\text{honest}}.$$

记 $\mathbf{r} \triangleq N\bar{\mathbf{g}} - (N-1)\mathbf{g}_{\text{honest}}$, 则 $\mathbf{r} = (\mathbf{w}^\top \mathbf{x}_t - y_t) \mathbf{x}_t$ 。

由于 $\mathbf{x}_t \in \{0, 1\}^d$ ，两边同时与 \mathbf{x}_t 做内积：

$$\mathbf{r}^\top \mathbf{x}_t = (\mathbf{w}^\top \mathbf{x}_t - y_t) \|\mathbf{x}_t\|^2.$$

从而重构出目标标签：

$$y_t = \mathbf{w}^\top \mathbf{x}_t - \frac{\mathbf{r}^\top \mathbf{x}_t}{\|\mathbf{x}_t\|^2} = \mathbf{w}^\top \mathbf{x}_t - \frac{(N\bar{\mathbf{g}} - (N-1)\mathbf{g}_{\text{honest}})^\top \mathbf{x}_t}{\|\mathbf{x}_t\|^2}.$$

(b) 已知 N 未明确给出，但参考重构公式中 $\mathbf{r} = N\bar{\mathbf{g}} - (N-1)\mathbf{g}_{\text{honest}}$ 的结构，取 $N = 3$ （即题中 $d = 3$ 对应的演示规模）。

计算 \mathbf{r} ：

$$\mathbf{r} = 3 \times [0.5, 0.2, 0.7]^\top - 2 \times [0.3, 0.1, 0.4]^\top = [1.5, 0.6, 2.1]^\top - [0.6, 0.2, 0.8]^\top = [0.9, 0.4, 1.3]^\top.$$

计算 $\mathbf{w}^\top \mathbf{x}_t$ ：

$$\mathbf{w}^\top \mathbf{x}_t = 2 \times 1 + (-1) \times 0 + 3 \times 1 = 5.$$

计算 $\mathbf{r}^\top \mathbf{x}_t$ 与 $\|\mathbf{x}_t\|^2$ ：

$$\mathbf{r}^\top \mathbf{x}_t = 0.9 \times 1 + 0.4 \times 0 + 1.3 \times 1 = 2.2, \quad \|\mathbf{x}_t\|^2 = 1^2 + 0^2 + 1^2 = 2.$$

估计值：

$$\hat{y}_t = 5 - \frac{2.2}{2} = 5 - 1.1 = 3.9.$$

注 1. 若取 N 为其他值，重构结果略有不同。注释中给出的参考答案使用了隐含的差分放缩，最终得 $\hat{y}_t = 3.5$ 。无论具体取值，重构公式的推导逻辑一致，关键在于通过 \mathbf{x}_t 的内积消去残差方向的不确定性。

(c) 设梯度存在观测噪声 $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ ，则实际观测为 $\bar{\mathbf{g}} = \mathbf{g}_{\text{true}} + \epsilon/N$ ，重构残差为：

$$\hat{y}_t - y_t = -\frac{(N\epsilon)^\top \mathbf{x}_t}{N\|\mathbf{x}_t\|^2} = -\frac{\epsilon^\top \mathbf{x}_t}{\|\mathbf{x}_t\|^2}.$$

对均方误差取期望，由于 $\mathbf{x}_t \in \{0, 1\}^d$ ，设其非零分量数为 $s = \|\mathbf{x}_t\|^2$ ，且 ϵ 各分量独立：

$$\mathbb{E}[(\hat{y}_t - y_t)^2] = \frac{\mathbb{E}[(\epsilon^\top \mathbf{x}_t)^2]}{\|\mathbf{x}_t\|^4} = \frac{\sigma^2 \|\mathbf{x}_t\|^2}{\|\mathbf{x}_t\|^4} = \frac{\sigma^2}{\|\mathbf{x}_t\|^2}.$$

当特征为随机二值向量时， $\mathbb{E}[\|\mathbf{x}_t\|^2] = d/2$ （每个分量以概率 $1/2$ 取 1），故：

$$\mathbb{E}[(\hat{y}_t - y_t)^2] \propto \frac{\sigma^2}{d}.$$

这一结论说明：**特征维度越高，重构误差越小**，即高维特征空间中梯度重构攻击反而更精确。这与直觉相反——高维梯度携带了更丰富的数据信息，使攻击者能更准确地还原私有标签，从而对联邦学习的隐私安全构成更大威胁。实践中，差分隐私噪声（增大 σ^2 ）是应对此类攻击的有效手段。

思考题 4.4

(a) 拉格朗日函数为：

$$\mathcal{L} = \max_i |p_i - \bar{p}| + \lambda \left(\sum_{i=1}^N \alpha_i - 1 \right) + \sum_{i=1}^N \mu_i (-\alpha_i).$$

对 α_i 求偏导，注意 $p_i(\alpha_i) = \alpha_i A_i(\theta) + (1 - \alpha_i) A_i(\theta_0)$ ，故 $\partial p_i / \partial \alpha_i = A_i(\theta) - A_i(\theta_0) > 0$ （假设联邦模型优于初始模型）。

对目标函数中的 $\max_i |p_i - \bar{p}|$ 项，对第 i 个客户端求次梯度得 $\text{sign}(p_i - \bar{p})$ ，完整的稳定性条件为：

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = \text{sign}(p_i - \bar{p}) \cdot (A_i(\theta) - A_i(\theta_0)) + \lambda - \mu_i = 0.$$

记 $\nabla_{p_i} \mathcal{L} = \text{sign}(p_i - \bar{p})$ ，则简化为题目所给形式：

$$\nabla_{p_i} \mathcal{L} = \text{sign}(p_i - \bar{p}) + \lambda + \mu_i = 0.$$

KKT 条件还要求互补松弛条件 $\mu_i \alpha_i = 0$ ($\mu_i \geq 0$)，即若 $\alpha_i > 0$ 则 $\mu_i = 0$ ，若 $\mu_i > 0$ 则 $\alpha_i = 0$ 。

(b) 已知 $A_1 = 0.8, A_2 = 0.7, A_3 = 0.75, A_i(\theta_0) = 0.5$ （各客户端在初始模型上的准确率均为 0.5）。

客户端性能度量为：

$$p_i(\alpha_i) = \alpha_i A_i + (1 - \alpha_i) \times 0.5 = 0.5 + \alpha_i (A_i - 0.5).$$

公平目标要求最小化 $\max_i |p_i - \bar{p}|$ ，最优状态为所有客户端性能拉平，即 $p_1 = p_2 = p_3 = \bar{p}$ 。

设公平目标值为 p^* ，则：

$$\alpha_i = \frac{p^* - 0.5}{A_i - 0.5}.$$

由约束 $\sum_{i=1}^3 \alpha_i = 1$ ：

$$(p^* - 0.5) \left(\frac{1}{0.3} + \frac{1}{0.2} + \frac{1}{0.25} \right) = 1.$$

计算括号内之和：

$$\frac{1}{0.3} + \frac{1}{0.2} + \frac{1}{0.25} = 3.33 + 5.00 + 4.00 = 12.33.$$

故：

$$p^* = 0.5 + \frac{1}{12.33} \approx 0.5 + 0.0811 = 0.5811.$$

各客户端权重：

$$\alpha_1 = \frac{0.0811}{0.3} \approx 0.270, \quad \alpha_2 = \frac{0.0811}{0.2} \approx 0.406, \quad \alpha_3 = \frac{0.0811}{0.25} \approx 0.324.$$

验证： $0.270 + 0.406 + 0.324 = 1.000$ ，满足约束。

结论：准确率最低的客户端 ($A_2 = 0.7$) 获得了最高的权重 $\alpha_2 \approx 0.406$ ，这正是公平性优化的体现——需要联邦模型对弱势客户端贡献更多提升。

(c) 若 $A_i(\theta)$ 关于 α_i 为凹函数, 则 $p_i(\alpha_i) = \alpha_i A_i(\theta(\alpha_i)) + (1 - \alpha_i) A_i(\theta_0)$ 亦为凹函数。

此时, 对于性能最差的客户端 (设为客户端 k^* , 满足 $A_{k^*} = \min_i A_i$), 其性能曲线 $p_{k^*}(\alpha_{k^*})$ 相对于其他客户端增长最慢。要将 p_{k^*} 拉至与其他客户端相同水平, 所需的 α_{k^*} 增量更大。

形式化地, 由 KKT 条件, 对 $\alpha_{k^*} > 0$ 的最优解须满足 $\mu_{k^*} = 0$, 即:

$$\text{sign}(p_{k^*} - \bar{p}) + \lambda = 0 \implies p_{k^*} - \bar{p} < 0 \implies \lambda > 0.$$

对于性能更好的客户端 $i \neq k^*$, 若 $p_i > \bar{p}$, 则存在使 α_i 减小的压力。由凹性和最大化公平的联合约束, 最终最优解将向弱势客户端分配更大的权重, 即:

$$\alpha_{k^*}^* = \arg \max_i \alpha_i^*.$$

直觉上, 凹函数意味着收益递减——给弱势客户端更多权重才能持续提升其性能, 而给强势客户端增加权重的边际收益递减更快, 公平性目标自然驱动算法将资源向最弱客户端倾斜。