第 13 次课笔记

第 13 组

2025 年 12 月 3 日

1　乘性权重更新法

定义 1 (在线学习)**.** 在线学习（*Online Learning*）是一种机器学习范式，其中模型在数据逐步到达时进行训练，而不是在整个数据集可用时进行批量训练。在线学习算法在每次接收到新数据点时更新其模型，从而能够适应动态变化的数据分布。

定义 2 (乘性权重更新法)**.** 乘性权重更新法（*Multiplicative Weights Update Method*）是一种迭代算法，用于在多个选项或专家之间分配权重，以最小化损失或最大化收益。该方法通过根据每个选项的表现调整其权重，通常是通过将权重乘以一个因子来实现。表现较好的选项的权重会增加，而表现较差的选项的权重会减少，从而使得算法能够逐渐集中在最优选项上。

定义 3 (遗憾)**.** 遗憾（*Regret*）用于衡量一个在线算法在一系列决策中相对于最佳固定决策的表现差距。具体来说，遗憾定义为在线算法的累计损失与在相同时间段内选择最佳固定决策所产生的累计损失之间的差值。数学上，遗憾 R 可以表示为：

$$R = L_{online} - L_{best} \tag{1}$$

其中，L_{online} 是在线算法的累计损失，L_{best} 是最佳固定决策的累计损失。遗憾越小，表示在线算法的表现越接近于最佳固定决策。

定义 4 (专家)**.** 专家（*Experts*）指的是一组预定义的预测模型或策略，这些模型或策略可以为每个时间步提供建议或预测。每个专家都有其独特的预测方法和性能表现，在线学习算法通过结合这些专家的建议来做出决策。专家的目标是通过比较不同专家的表现，选择出最优的预测策略，从而最小化整体的损失或遗憾。

在在线学习的时候，有多个专家的预测序列和每天的实际结果，希望能够根据这些数据建立一个预测模型，使得预测误差最小化（效果接近于最好的专家）。

一个最简单的算法就是“随机算法”（randomized algorithm），其基本思想是根据专家的历史表现动态调整专家的权重，并在每个时间步随机选择一个专家进行预测。具体步骤如下：

设有 N 个专家，每个专家在每个时间步 t 给出一个预测。算法在每个时间步 t 根据专家的权重分布选择一个专家进行预测。假设在时间步 t，专家 i 的权重为 $w_i(t)$，则算法选择专家 i 的概率为：

$$P(i,t) = \frac{w_i(t)}{\sum_{j=1}^{N} w_j(t)} \tag{2}$$

在每个时间步 t，算法根据选择的专家进行预测，并根据实际结果计算损失。设专家 i 在时间步 t 的损失为 $l_i(t)$，则算法在时间步 t 的预测损失为：

$$L(t) = \sum_{i=1}^{N} P(i,t) \cdot \ell_i(t) \tag{3}$$

算法在每个时间步 t 结束后更新专家的权重。对于每个专家 i，如果其在时间步 t 的损失为 $\ell_i(t)$，则其权重更新规则为：

$$w_i(t+1) = w_i(t) \cdot (1 - \eta \cdot \ell_i(t)) \tag{4}$$

这里假设 $\ell_i(t) \in [0,1]$，或需要做适当的归一化处理。$\ell_i(t) > 1$ 时，可能会导致权重变为负数，这在实际应用中需要避免。η 是一个小的正数，称为学习率。通过这种方式，表现较差的专家的权重会减少，而表现较好的专家的权重会增加。

经过 T 个时间步后，算法的总损失为：

$$L_{total} = \sum_{t=1}^{T} L(t) \tag{5}$$

而最佳专家的总损失为：

$$L_{best} = \min_{i} \sum_{t=1}^{T} \ell_i(t) \tag{6}$$

因而我们实际上得到了遗憾的定义（这个更加明确）：

$$R_T = \sum_{t=1}^{T} L(t) - \min_{i \in [N]} \sum_{t=1}^{T} \ell_i(t) \tag{7}$$

其中，$L(t)$ 是在线算法在时间步 t 的损失，$\ell_i(t)$ 是专家 i 在时间步 t 的损失，N 是专家的数量，T 是总的时间步数。这个和 定义 3 实际上是等价的。

定理 1 (随机算法的遗憾界)**.** 随机算法能够达到遗憾界（*regret bound*）$R \leqslant O(\sqrt{T \log N})$ 的总损失。其中 T 为时间步数，N 为专家数量。

证明. 下面试图证明 定理 1 。

定义在 t 时刻的权重和为:

$$W_t = \sum_{i=1}^{N} w_i(t) \tag{8}$$

我们关心权重和的变化，并和最佳专家的性能比较。我们对两个相邻时间步的权重和相除，得到:

$$\frac{W_{t+1}}{W_t} = \frac{\sum_{i=1}^{N} w_i(t+1)}{\sum_{i=1}^{N} w_i(t)} = \frac{\sum_{i=1}^{N} w_i(t)(1-\eta\ell_i(t))}{\sum_{i=1}^{N} w_i(t)} = 1 - \eta \sum_{i=1}^{N} \frac{w_i(t)}{W_t}\ell_i(t) \tag{9}$$

注意到 $\sum_{i=1}^{N} \frac{w_i(t)}{W_t}\ell_i(t)$ 实际上就是算法在时间步 t 的损失 $L(t)$，因此我们可以将上式改写为:

$$\frac{W_{t+1}}{W_t} = 1 - \eta L(t) \tag{10}$$

对上式取对数，并进行简单的放缩:

$$\log \frac{W_{t+1}}{W_t} = \log(1-\eta L(t)) \leqslant -\eta L(t) \tag{11}$$

将上式对 t 从 1 到 T 求和，得到:

$$\log \frac{W_{T+1}}{W_1} = sum_{t=1}^{T} \log \frac{W_{t+1}}{W_t} \leqslant -\eta \sum_{t=1}^{T} L(t) \tag{12}$$

注意到初始时刻的权重和为 $W_1 = N$（每个专家的初始权重为 1），因此我们有:

$$\log W_{T+1} - \log N \leqslant -\eta \sum_{t=1}^{T} L(t) \tag{13}$$

接下来，我们需要对 W_{T+1} 进行下界估计。不妨设最佳专家为专家 j，则有:

$$W_{T+1} = \sum_{i=1}^{N} w_i(T+1) \geqslant w_j(T+1) = \prod_{t=1}^{T}(1-\eta\ell_j(t)) \tag{14}$$

因此，我们可以得到:

$$\log W_{T+1} \geqslant \sum_{t=1}^{T} \log(1-\eta\ell_j(t)) \geqslant -\eta \sum_{t=1}^{T} \ell_j(t) - \eta^2 \sum_{t=1}^{T} \ell_j(t)^2 \geqslant -\eta \sum_{t=1}^{T} \ell_j(t) - \eta^2 T \tag{15}$$

把这个下界代入之前的式子，得到:

$$-\eta \sum_{t=1}^{T} \ell_j(t) - \eta^2 T - \log N \leqslant -\eta \sum_{t=1}^{T} L(t) \tag{16}$$

发现 $\sum_{t=1}^{T} \ell_j(t)$ 就是最佳专家的总损失 L_{best}，因此我们可以将上式改写为：

$$\sum_{t=1}^{T} L(t) - L_{best} \leqslant \frac{\log N}{\eta} + \eta T \tag{17}$$

上式左边实际上就是遗憾 R_T，因此我们有：

$$R_T \leqslant \frac{\log N}{\eta} + \eta T \tag{18}$$

右边是一个非常经典的、可以使用均值不等式的形式。通过选择 $\eta = \sqrt{\frac{\log N}{T}}$，我们可以最小化右边的表达式，从而得到：

$$R_T \leqslant 2\sqrt{T \log N} \tag{19}$$

上式实际上就是定理 1所要证明的遗憾界。 □

引理 2 (随机算法的平均损失界)**.** 在时间充分大的情况下，随机算法的效果可以接近于最好的专家。具体来说，随机算法的平均损失与最好的专家的平均损失之差满足以下不等式：

$$\frac{L_{random}}{T} - \frac{L_{best}}{T} \leqslant O(\sqrt{\frac{\log N}{T}}) \tag{20}$$

其中 L_{random} 为随机算法的总损失，L_{best} 为最好的专家的总损失，T 为时间步数，N 为专家数量。

证明. 根据随机算法的遗憾界，我们有：

$$R = L_{random} - L_{best} \leqslant O(\sqrt{T \log N}) \tag{21}$$

将上式两边除以时间步数 T，得到：

$$\frac{L_{random}}{T} - \frac{L_{best}}{T} \leqslant O\left(\frac{\sqrt{T \log N}}{T}\right) = O\left(\sqrt{\frac{\log N}{T}}\right) \tag{22}$$

因此，随机算法的平均损失与最好的专家的平均损失之差满足不等式：

$$\frac{L_{random}}{T} - \frac{L_{best}}{T} \leqslant O\left(\sqrt{\frac{\log N}{T}}\right) \tag{23}$$

这表明随着时间步数 T 的增加，随机算法的平均损失将越来越接近于最好的专家的平均损失。 □

引理 3 (随机算法的渐近最优性)**.** 随着时间步数 T 的增加，随机算法的平均损失将趋近于最好的专家的平均损失。具体来说，当 $T \to \infty$ 时，有：

$$\lim_{T \to \infty} \left(\frac{L_{random}}{T} - \frac{L_{best}}{T}\right) = 0 \tag{24}$$

证明. 根据随机算法的平均损失界，我们有：

$$\frac{L_{random}}{T} - \frac{L_{best}}{T} \leqslant O\left(\sqrt{\frac{\log N}{T}}\right) \tag{25}$$

当时间步数 T 趋近于无穷大时，$\sqrt{\frac{\log N}{T}}$ 趋近于 0。因此，我们可以得出上述结论。 □

2 矩阵博弈以及其和乘性权重更新法的关系

定义 5 (矩阵博弈)**.** 矩阵博弈（*Matrix Game*）是一种零和博弈，其中两个玩家通过选择策略来最大化自己的收益，同时最小化对手的收益。其中，矩阵的行表示一个玩家的策略，列表示另一个玩家的策略，矩阵中的元素表示对应策略组合下的收益或损失。

这一博弈可以表示为：

$$\min_{p}\max_{q} p^T\mathbf{M}q = \max_{q}\min_{p} p^T\boldsymbol{M}q, \quad M_{i,j} \in [0,1] \tag{26}$$

其中，p 和 q 分别表示两个玩家的混合策略，M 为收益矩阵；行玩家（*row player*）视为在线学习者，列玩家（*column player*）视为环境（对手）。矩阵 $\mathbf{M}$ 的每一行视为一个专家。这一定理也被称为 ***min-max* 定理**（*Minimax Theorem*）。

下面，我们将会证明 左边 $\leqslant$ 右边。

证明. 给出以下设定：

- 矩阵 $\mathbf{M}_{i,j} \in [0,1]$，表示一个双人零和博弈的支付矩阵。
- 行玩家（row player）选择混合策略 p，列玩家（column player）选择混合策略 q。
- 行玩家希望最小化 $p^T\mathbf{M}q$，而列玩家希望最大化该值。

我们将会使用在线学习框架来证明。将行玩家视为在线学习者，列玩家视为环境（对手）。矩阵 $\mathbf{M}$ 的每一行视为一个专家。

使用乘性权重更新：第 i 个专家在 t 轮的损失是 $\ell_i(t) = \mathbf{M}_i^T q_t$，其中 q_t 是列玩家在第 t 轮选择的混合策略。行玩家使用乘性权重更新法来选择专家。其更新规则是

$$p_{t+1}(i) = \frac{p_t(i)\beta^{\ell_i(t)}}{\sum_j p_t(j)\beta^{\ell_j(t)}} \tag{27}$$

其中 $\beta \in (0,1)$ 是一个参数。

从而行玩家的平均损失是

$$\frac{1}{T}\sum_{t=1}^{T} p_t^T \mathbf{M} q_t \tag{28}$$

最佳玩家的损失是

$$\min_i \frac{1}{T}\sum_{t=1}^{T} \mathbf{M}_i^T q_t = \min_p p^T \mathbf{M} \bar{q} \tag{29}$$

上述式中，$\bar{q} = \frac{1}{T}\sum_{t=1}^{T} q_t$。

假设每一轮列玩家选择的策略是最优的，即

$$q_t = \arg\max_q p_t^T \mathbf{M} q \tag{30}$$

则根据乘性权重更新法的遗憾界，我们有

$$\frac{1}{T}\sum_{t=1}^{T} p_t^T \mathbf{M} q_t \leqslant \min_p p^T \mathbf{M} \bar{q} + O\left(\sqrt{\frac{\log N}{T}}\right) \tag{31}$$

当 $T \to \infty$ 时，右边的误差项趋近于 0，因此我们得到

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} p_t^T \mathbf{M} q_t \leqslant \min_p p^T \mathbf{M} \bar{q} \tag{32}$$

注意到左边实际上是行玩家在长期博弈中的平均损失，而右边是行玩家在面对列玩家的平均策略 $\bar{q}$ 时的最小损失。因此，我们得出结论：

$$\min_p \max_q p^T \mathbf{M} q \leqslant \max_q \min_p p^T \mathbf{M} q \tag{33}$$

□

而通过类似的思路，我们也可以证明 左边 $\geqslant$ 右边（弱对偶性），从而完成整个 min-max 定理的证明。

证明. 另一种证明方式大概是这样的，从式 (30) 出发，显然对于所有的 $\varepsilon > 0$，都有

$$\frac{1}{T}\sum_t p_t^T M q_t \leq \min_p p^T M \bar{q} + \epsilon \tag{34}$$

$$\Rightarrow \frac{1}{T}\sum_t \max_q p_t^T M q \leq \max_q \min_p p^T M q + \epsilon \tag{35}$$

$$\Rightarrow \max_q \frac{1}{T}\sum_t p_t^T M q \leq \max_q \min_p p^T M q + \epsilon \tag{36}$$

$$\Rightarrow \max_q \bar{p}^T M q \leq \max_q \min_p p^T M q + \epsilon \tag{37}$$

$$\Rightarrow \min_p \max_q p^T M q \leq \max_q \min_p p^T M q + \epsilon \tag{38}$$

□

3 多臂老虎机问题

定义 6 (多臂老虎机问题)**.** 多臂老虎机问题（*Multi-Armed Bandit Problem*）是一种经典的强化学习问题，描述了一个决策者在面对多个选择（或“臂”）时，如何在探索（尝试不同选择以获取信息，*Exploration*）和利用（选择已知的最佳选择以最大化收益，*Exploitation*）之间进行权衡。每个臂都有一个未知的奖励分布，决策者的目标是通过一系列选择来最大化累积奖励。

该问题有以下设定：

- 有 k 个臂（选项），每个臂的期望损失是 $\mu(1), \mu(2), \ldots, \mu(k)$，但这些期望损失是未知的。
- 目标是：最小化在线学习者的总损失与选择最优臂的总损失之间的差距，即最小化遗憾（*Regret*）。
- 该问题是“暗箱”的，也就是说，在线学习者在每次选择一个臂后，只能观察到该臂的损失，而无法获得其他臂的信息。

在多臂老虎机中，遗憾可以由期望总损失 $E[\sum_{t=1}^{T} l(a_t)]$ 和最优臂的期望损失 $a^* = \arg\min_a \mu(a)$ 之间的差距来定义：

$$R = E\left[\sum_{t=1}^{T} l(a_t)\right] - T\mu(a^*) = \sum_{t=1}^{T}(E[l(a_t)] - \mu(a^*)) \tag{39}$$

命题 4 (乐观面对不确定性算法)**.** 乐观面对不确定性（*Optimism in the Face of Uncertainty, OFU*）算法是一种用于解决多臂老虎机问题的策略。该算法通过在每次选择臂时考虑当前估计的奖励和一个不确定性项，从而在探索和利用之间进行权衡。具体来说，*OFU* 算法在每个时间步 t 选择臂 a_t，使得以下表达式最大化：

$$a_t = \arg\max_a \left(\hat{\mu}_t(a) + c \cdot \sqrt{\frac{\log t}{N_t(a)}}\right) \tag{40}$$

上式中，c 是一个调节探索程度的常数，$\hat{\mu}_t(a)$ 是臂 a 在时间步 t 的当前估计奖励，$N_t(a)$ 是臂 a 在时间步 t 之前被选择的次数。

定理 5 (遗憾界定理)**.** 在多臂老虎机问题中，使用乐观面对不确定性（*OFU*）算法可以实现以下遗憾界：

$$R_T \leq \sum_{a:\Delta_a>0}\left(\frac{16\log T}{\Delta_a} + 2\Delta_a\right) \tag{41}$$

其中 $\Delta_a = \mu(a) - \mu(a^*)$。

该遗憾界表明，随着时间步数 T 的增加，遗憾 R_T 的增长速度是对数级别的，这意味着算法在长期内能够有效地平衡探索和利用，从而接近最优策略。但是，当非最优臂和最优臂之间的差距 Δ_a 较小时，遗憾界中的项会变得较大，该上界的刻画效果会变差。