- 进一步把公式 (4.2) 中的 $Q_\star$ 近似成 $\widetilde{Q}$，得到

$$\widehat{y}_t \triangleq r_t + \gamma \cdot \max_{a \in \mathcal{A}} \widetilde{Q}(s_{t+1}, a).$$

把它称作 TD 目标。它是表格在 $t+1$ 时刻对 $Q_\star(s_t, a_t)$ 做出的估计。

$\widetilde{Q}(s_t, a_t)$ 和 $\widehat{y}_t$ 都是对最优动作价值 $Q_\star(s_t, a_t)$ 的估计。由于 $\widehat{y}_t$ 部分基于真实观测到的奖励 r_t，我们认为 $\widehat{y}_t$ 是更可靠的估计，所以鼓励 $\widetilde{Q}(s_t, a_t)$ 更接近 $\widehat{y}_t$。更新表格 $\widetilde{Q}$ 中 (s_t, a_t) 位置上的元素：

$$\widetilde{Q}(s_t, a_t) \leftarrow (1-\alpha) \cdot \widetilde{Q}(s_t, a_t) + \alpha \cdot \widehat{y}_t.$$

这样可以使得 $\widetilde{Q}(s_t, a_t)$ 更接近 $\widehat{y}_t$。Q 学习的目的是让 $\widetilde{Q}$ 逐渐趋近于 $Q_\star$。

收集训练数据： Q 学习更新 $\widetilde{Q}$ 的公式不依赖于具体的策略。我们可以用任意策略控制智能体，与环境交互，把得到的轨迹划分成 (s_t, a_t, r_t, s_{t+1}) 这样的四元组，存入经验回放数组。这个控制智能体的策略叫做行为策略 (Behavior Policy)，比较常用的行为策略是 ϵ-greedy:

$$a_t = \begin{cases} \text{argmax}_a \widetilde{Q}(s_t, a), & \text{以概率 } (1-\epsilon); \\ \text{均匀抽取 } \mathcal{A} \text{ 中的一个动作}, & \text{以概率 } \epsilon. \end{cases}$$

事后用经验回放更新表格 $\widetilde{Q}$，可以重复利用收集到的四元组。

经验回放更新表格 $\widetilde{Q}$： 随机从经验回放数组中抽取一个四元组，记作 (s_j, a_j, r_j, s_{j+1})。设当前表格为 $\widetilde{Q}_{\text{now}}$。更新表格中 (s_j, a_j) 位置上的元素，把更新之后的表格记作 $\widetilde{Q}_{\text{new}}$。

1. 把表格 $\widetilde{Q}_{\text{now}}$ 中第 (s_j, a_j) 位置上的元素记作：

$$\widehat{q}_j = \widetilde{Q}_{\text{now}}(s_j, a_j).$$

2. 查看表格 $\widetilde{Q}_{\text{now}}$ 的第 s_{j+1} 行，把该行的最大值记作：

$$\widehat{q}_{j+1} = \max_a \widetilde{Q}_{\text{now}}(s_{j+1}, a).$$

3. 计算 TD 目标和 TD 误差：

$$\widehat{y}_j = r_j + \gamma \cdot \widehat{q}_{j+1}, \qquad \delta_j = \widehat{q}_j - \widehat{y}_j.$$

4. 更新表格中 (s_j, a_j) 位置上的元素：

$$\widetilde{Q}_{\text{new}}(s_j, a_j) \leftarrow (1-\alpha) \cdot \widetilde{Q}_{\text{now}}(s_j, a_j) - \alpha \cdot \delta_j.$$

收集经验与更新表格 $\widetilde{Q}$ 可以同时进行。每当智能体执行一次动作，我们可以用经验回放对 $\widetilde{Q}$ 做几次更新。也可以当完成一局游戏，对 $\widetilde{Q}$ 做几次更新。