整数），则式（5.5）变为

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots + \gamma^{n-t-1} R_n \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots + \gamma^{n-t-2} R_n) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{5.6}$$

即 $G_t = R_{t+1} + \gamma G_{t+1}$。其中，$R_{t+1}$ 表示智能体在时间步 t 采取动作 A_t 后获得的即时奖励。因为每个回合总共有 n+1 个状态，所以，第 n+1 个状态处的回报 G_n 为 0。

在图 5.12 的基础上，引入折扣因子 γ，根据经验数据存储，从最后一个状态位置开始计算回报，反向迭代计算过程如图 5.13 所示。

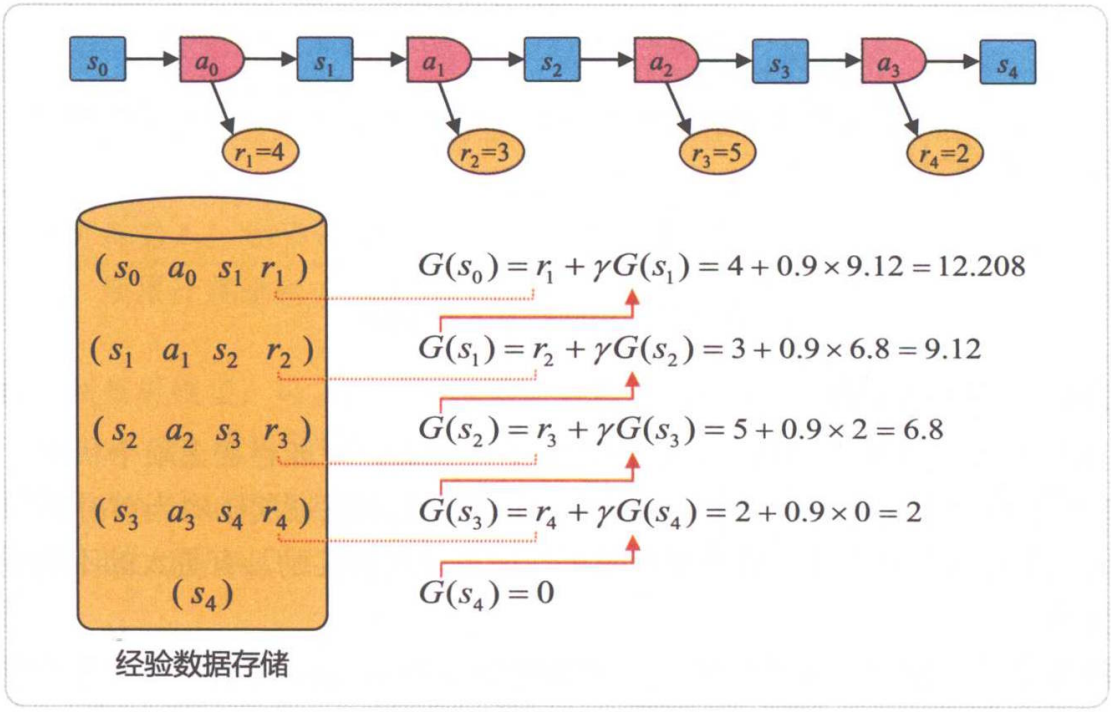

图 5.13　反向迭代计算过程

5.2.3　价值函数

在强化学习中，**价值函数**用于衡量智能体在特定状态 s 下，或在采取特定动作 a 后所能获得的回报期望 $\mathbb{E}_\pi[G_t]$。价值函数在指导智能体决策的过程中起着至关重要的作用。例如，当智能体处于状态 s 时，应该执行动作 a_0 还是 a_1 呢？通过价值函数可以预估这两个动作对应的未来回报，从而选择回报最高的动作。

本节将介绍四种价值函数：动作价值函数 $Q_\pi(s,a)$、状态价值函数 $V_\pi(s)$、最优状态价值函数 $V_*(s)$、最优动作价值函数 $Q_*(s,a)$，这四种价值函数**均表示未来回报**，但其条件和应用场景略有不同[32]。

1. 动作价值函数

动作价值函数（Action-Value Function）衡量在策略 π 下，智能体在状态 s 时采取动作 a 后所能获得的回报的期望值，该期望值有时也被泛称为 **Q 值**（**Q-value**），通常记为 $Q_\pi(s,a)$，其定义为