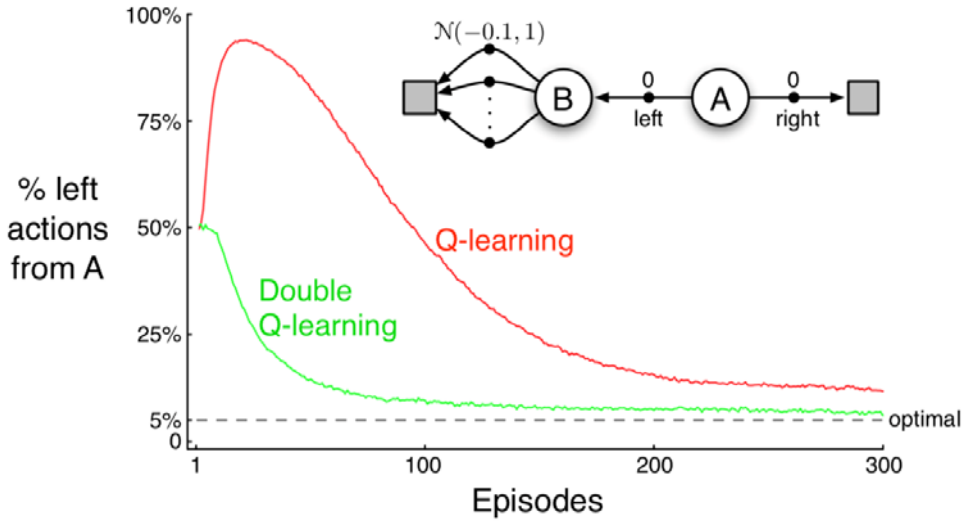
1. （最大化偏差与双 Q 学习，例 6.7）图 6.5 中显示的小型 MDP 提供了一个简单的示例，说明最大化偏差如何影响 TD 控制算法的性能。MDP 有两个非终端状态 A 和 B，回合总是从 A 开始，可以选择“左”、“右”两个动作。“右”动作立即转换到终止状态，奖励和返回为零。“左”动作转换为 B，同时奖励为零，之后有许多可能的动作，所有动作都会导致立即终止，奖励来自正态分布，平均值为 −0.1，方差为 1.0。 因此，从“左”开始的任何轨迹的预期回报是 −0.1，因此在状态 A 中向“左”移动总是错误的。 然而，我们的控制方法可能有利于“左”动作，因为最大化偏差使 B 看起来具有正值。图 6.5 显示，带有 ε -贪婪动作选择的 Q-learning 最初学会强烈支持“左”动作。即使在渐近线上，Q-learning 也比我们的参数设置 （ε=0.1，α=0.1 和 γ=0.1） 中的“左”动作大约多 5%。



问题：（1）复现图中曲线，提供代码。（2）从理论上解释（或计算）在上述参数设定下最优的向左运动的概率仍然存在 5%的原因。（3）从理论上解释（或计算）Q-learning 的“左”动作的概率大约多 5%的原因。（4）对算法进行适当改进使得向左运动的概率趋于 0（实际的最优值）。