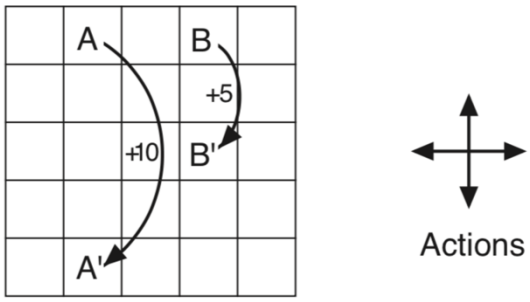
以下三个选题都改自 Sutton 和 Barto 的著作《强化学习（第 2 版）》中文版。

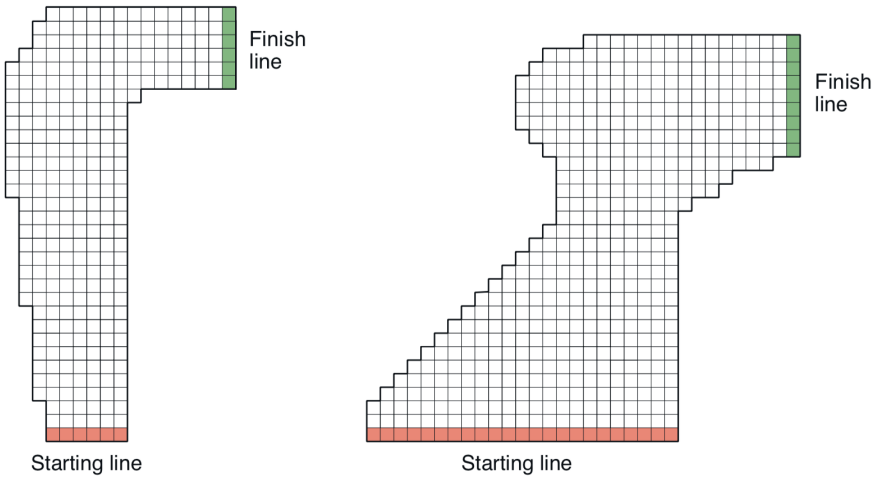
1. （网格世界，例 3.8）在每个单元格中，可以有四个动作：东，南，西，北。这明确让个体在网格上的相应方向上移动一个单元格。使个体离开网格的操作会使其位置保持不变，但会导致 -1 的奖励。除了将个体从特殊状态 A 和 B 移出的行为，其他行为奖励值为 0。在状态 A，所有四个动作都会产生 +10 的奖励，并将个体送到 A'；从状态 B，所有动作都会获得 +5 的奖励，并将个体转到 B'。



问题：分别设计 Q 学习和期望 SARSA 算法求解上述问题的最佳策略和值函数，比较两种算法在训练过程中的表现（在线性能）；提供源代码。

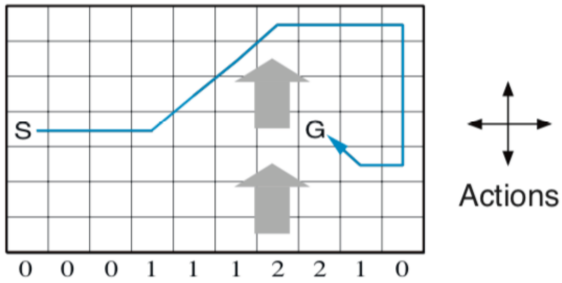
2. （赛道问题，练习 5.12）考虑驾驶赛车在像下图那样的赛道上拐弯。你想要尽可能的快，但是又不能冲出赛道。在我们简化版的赛车轨迹问题中，赛车在其中的一个离散的格子中。赛车的速度也是离散的，表示每个时间步长会在水平方向和竖直方向移动的格子数。动作是表示对速度的加速，每个时间步长增长量为 +1,−1,0，这样一共九种（3×3）动作。 所有的速度分量都是严格非负的，且不超过 5，除了起点，它们也不能同时为零。 每个回合开始时，选择一个随机的开始状态，速度分量均为零，当赛车跨过终点线时结束。在

结束之前的每一步，奖励为 −1。如果赛车碰到赛道的边界，又会从起点的随机位置重新开始，速度分量同时变为零，本回合继续。每个时间步长更新赛车的坐标之前，检查赛车的轨迹与赛道是否相交，如果相交在终点线，那么回合结束； 如果相交在其它，那么赛车碰到边界了，就得从起点开始。为了让问题更有挑战性，每个时间步长，速度有 0.1 的可能性保持原样。



问题：利用资格迹算法**或**策略梯度法求解上述问题的最佳策略和值函数，可仅针对一个赛道（可稍微简化赛道）测试算法在不同超参数下的表现；提供源代码。

3. （带障碍的、有风的网格世界，例 6.5）下图是一个标准的网格世界，有开始和目标状态，但有一个差异：在网格中间有一个向上运行的侧风。 动作是标准的四个—— 上，下，右 和 左，但在中间区域，结果的下一个状态向上移动一个“风”，其强度因列而异。在每列下方给出风的强度，向上移动的格子数量。例如，如果你是目标右侧的一个单元格，则左侧的操作会将你带到目标上方的单元格。这是一个没有折扣的回合任务，在达到目标状态之前回报恒定为 −1。



问题：自行增加网格规模（如 20*20），自行增加障碍（必须），重新设定风向和强度（可增加随机性），并设计两种强化学习算法（Q 学习，n 步自举，SARSA(λ)，策略梯度法等）求解该问题，比较不同算法，提供源代码。