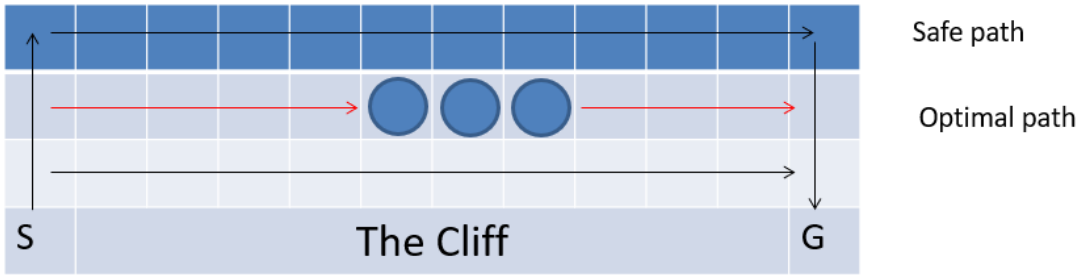

这是一个走悬崖的问题。强化学习中的主体从 S 出发走到 G 处一个回合结束，除了在边缘以外都有上下左右四个行动，如果主体走入悬崖区域，回报为-100，走入中间三个圆圈中的任一个，会得到-1 的奖励，走入其他所有的位置，回报都为-5。

问题：用 Q-learning 来使 agent 学习最优的策略。

12*4 格，S(0,3),G(11,3) ，(5,1)、(6,1)、(7,1)位置奖励-1，(1,3)—(10,3)为悬崖，