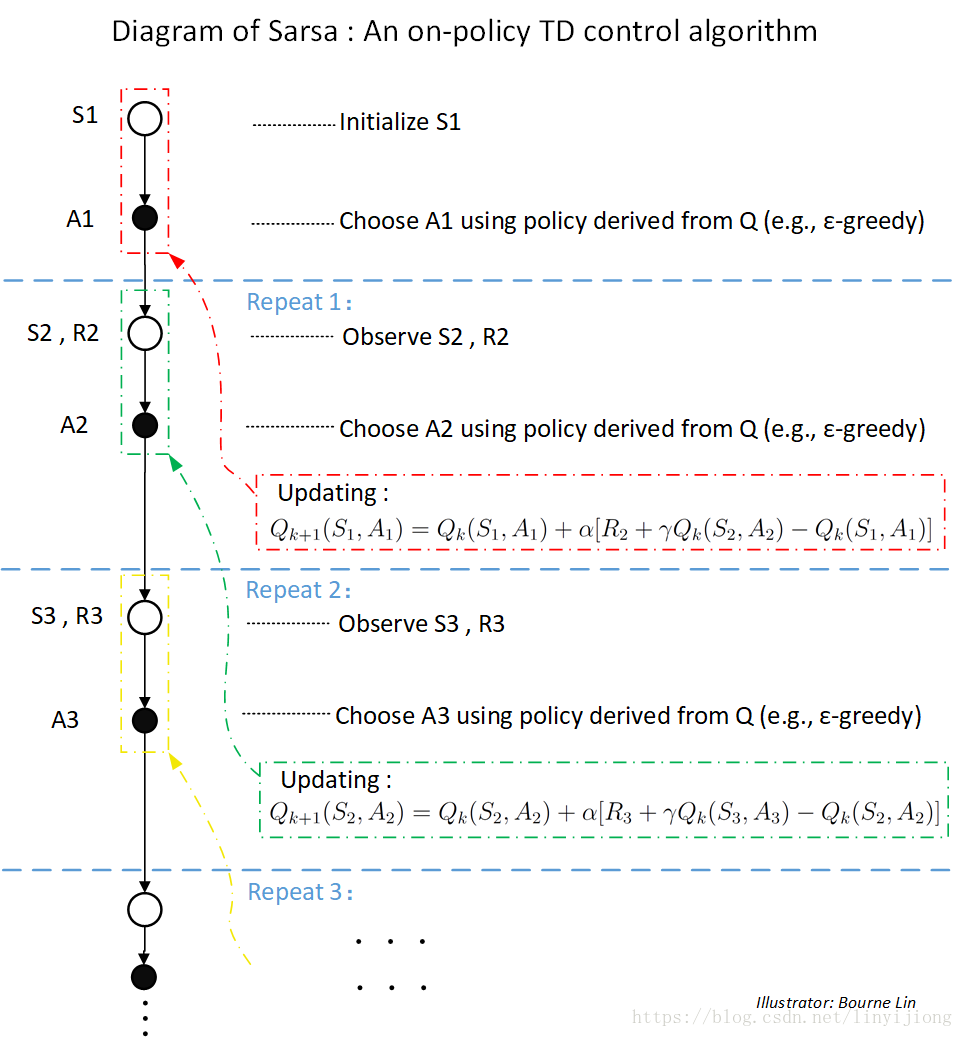
# Sarsa和Q-learning教程

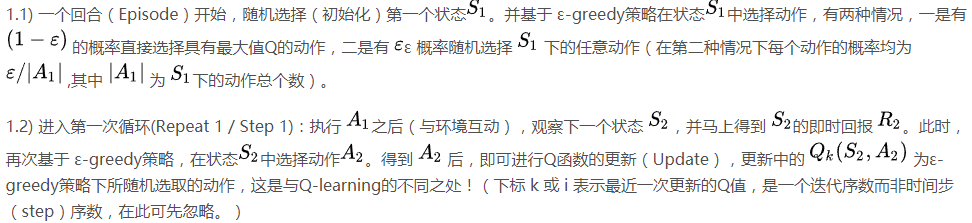
视频教程：<https://www.bilibili.com/video/av45015646/?p=7>

<https://www.bilibili.com/video/av45015646/?p=8>

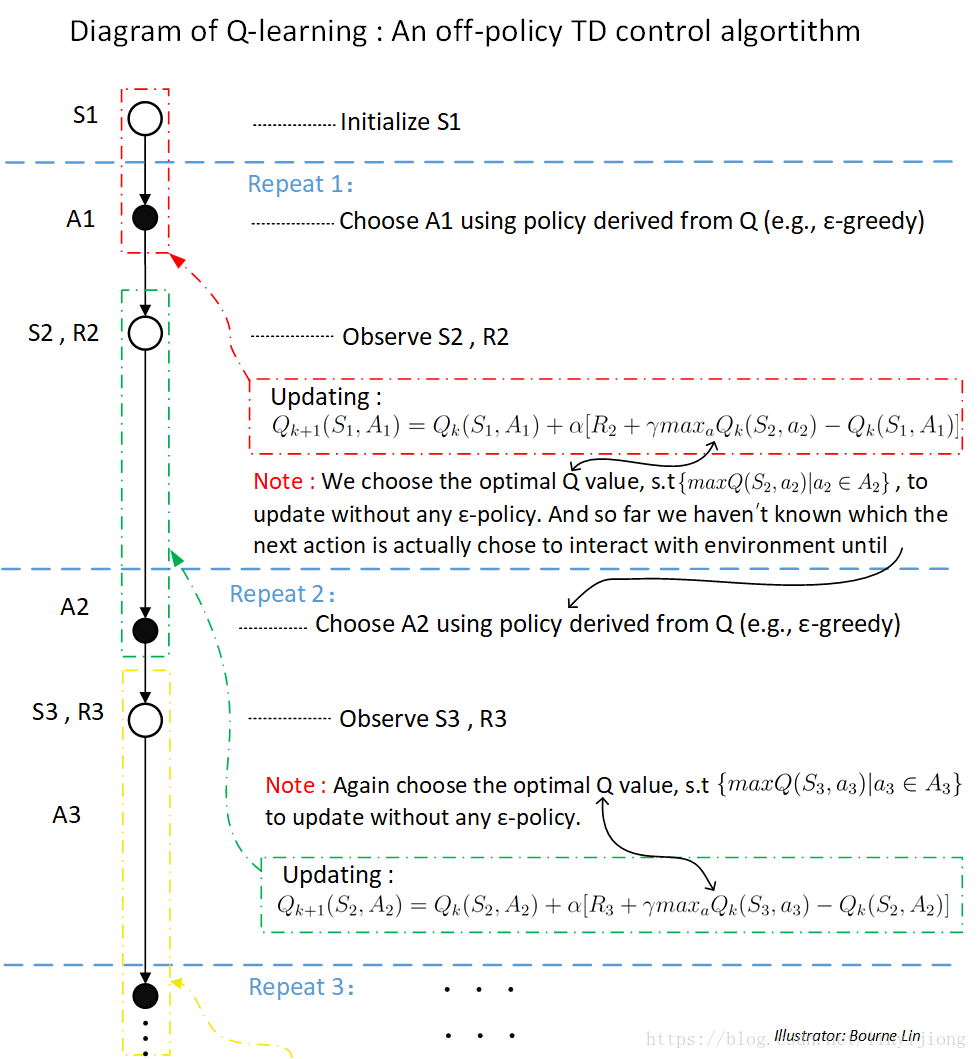
<https://www.bilibili.com/video/av45015646/?p=9>

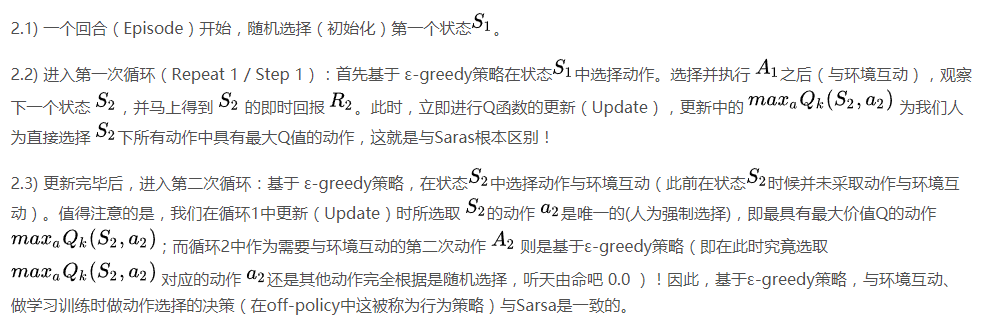
## Sarsa

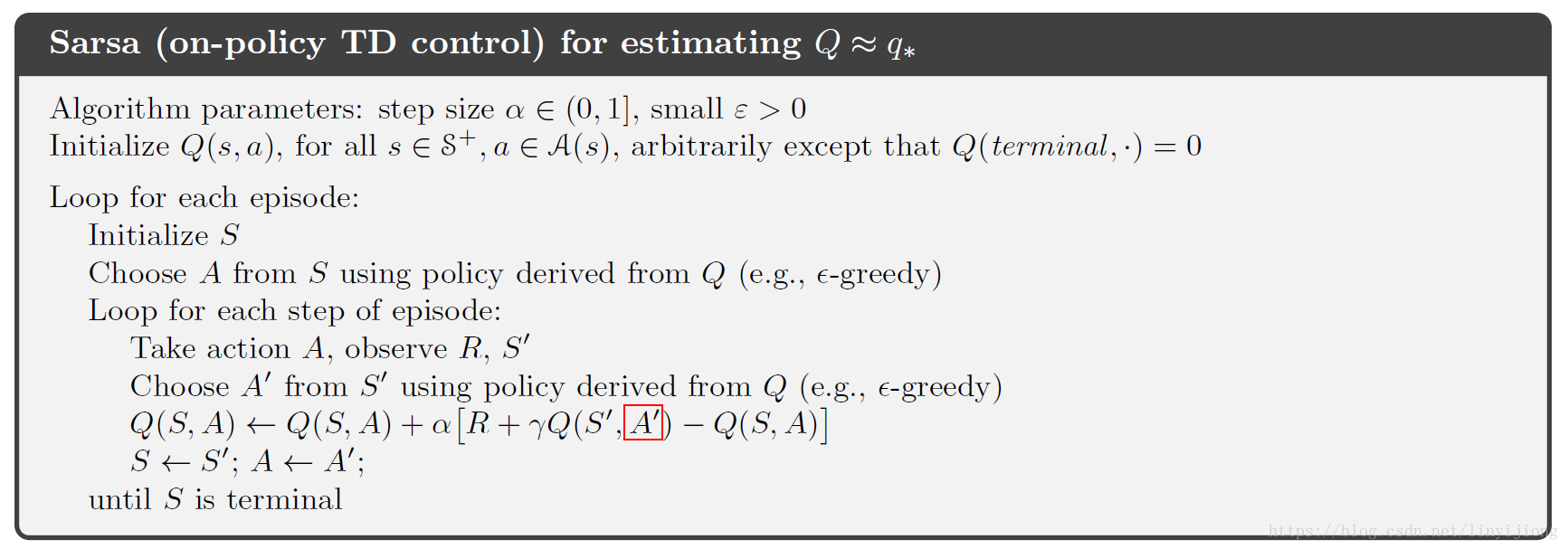


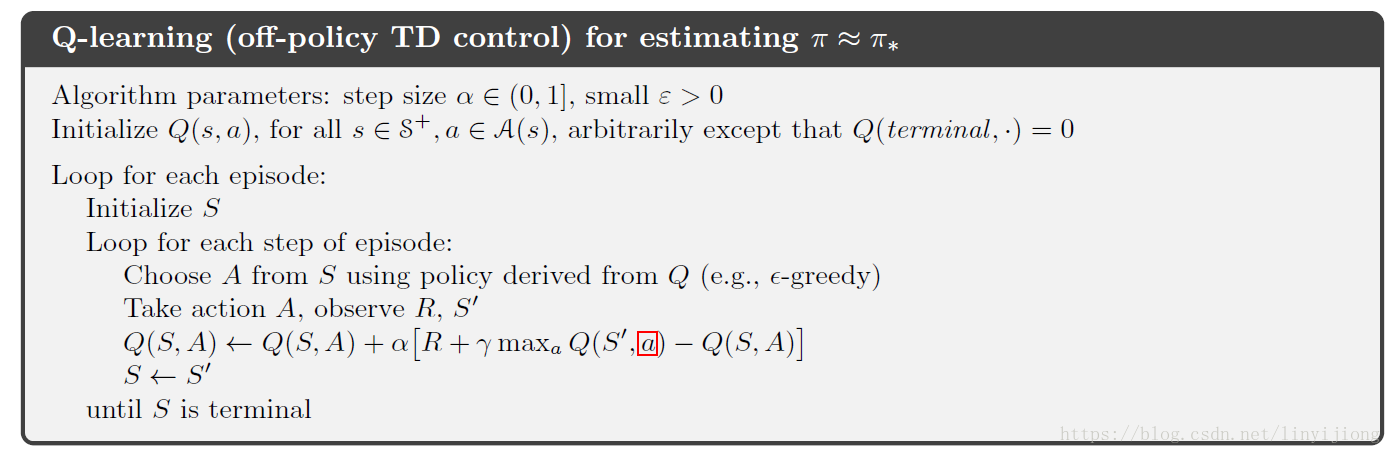


## Learning





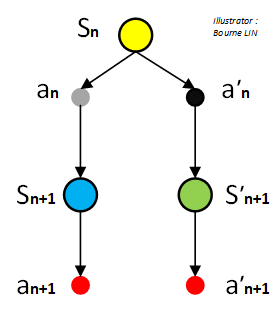


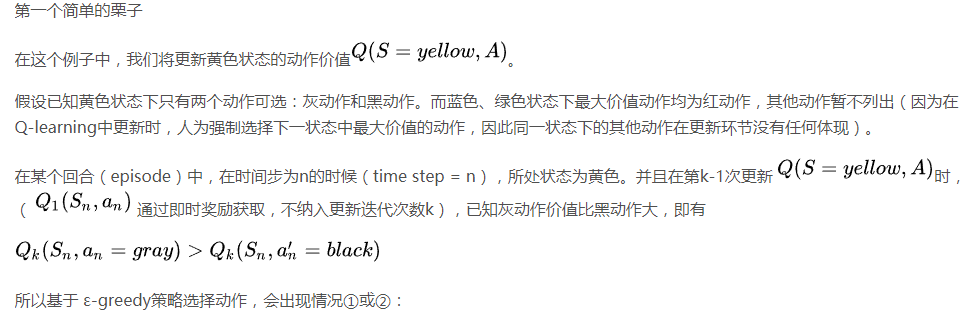


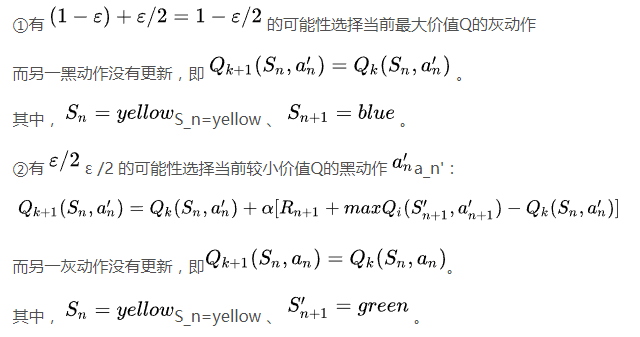
大写的A表示集合，比如 则表示 下的所有动作，而  则表示具体的一个动作，它们之间的关系为：。回到流程图中，可以发现出现a都在Q-learning的update公式中，这是因为我们在更新时，人为指定选择具有最大值Q的a，这是具有确定性的事件（Deterministic）。而在Q-learning中与环境互动的环节、在Sarsa中更新Q值的环节与环境互动的环节时，动作的选择是随机的（ ε-greedy），因此所有动作都有可能被选中，只不过是具有最大值Q的动作被选中的概率大。

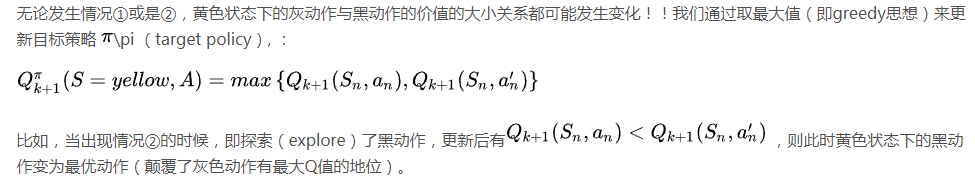
## Q-learning如何实现更加有效的探索？

清楚整个流程之后，我们来具体看看，Q-learning到底是怎么实现有意义的探索，如何在环境中发掘出更有价值的动作？（即一个当前估值（evaluate）不高但潜力巨大的动作的逆袭之路）





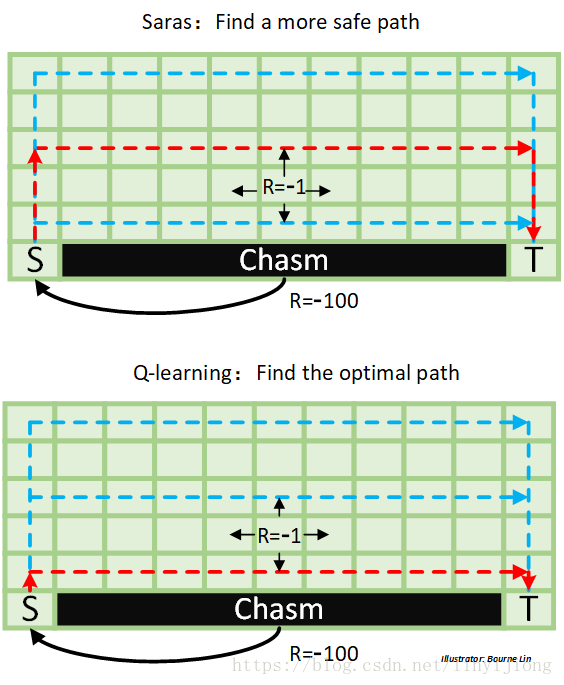




## 一个例子

在此举一个非常直观的例子来帮助我们认识一下Q-learning和Sarsa的实际应用效果的区别。

在下面栅格化的小世界中，绿色区域为草地，在上面每移动一格子就会扣1分，而踏入黑色区域的悬崖（chasm），会扣一百分，并且回到起始点S (Start)。我们希望能学习到一条得分最高的路径到达终点T (Terminal)。分别使用Sarsa和Q-learning进行学习。结果如图所示，红色为相应算法的最优路径。



可以看到，Q-learning寻找到一条全局最优的路径，因为虽然Q-learning的行为策略（behavior）是基于 ε-greedy策略，但其目标策略（target policy）只考虑最优行为；而Sarsa只能找到一条次优路径，这条路径在直观上更加安全，这是因为Sarsa（其目标策略和行为策略为同一策略）考虑了所有动作的可能性（ ε-greedy），当靠近悬崖时，由于会有一定概率选择往悬崖走一步，从而使得这些悬崖边路的价值更低。

## 总结

Q-learning虽然具有学习到全局最优的能力，但是其收敛慢；而Sarsa虽然学习效果不如Q-learning，但是其收敛快，直观简单。因此，对于不同的问题，我们需要有所斟酌。

参考文档：<https://blog.csdn.net/linyijiong/article/details/81607691>