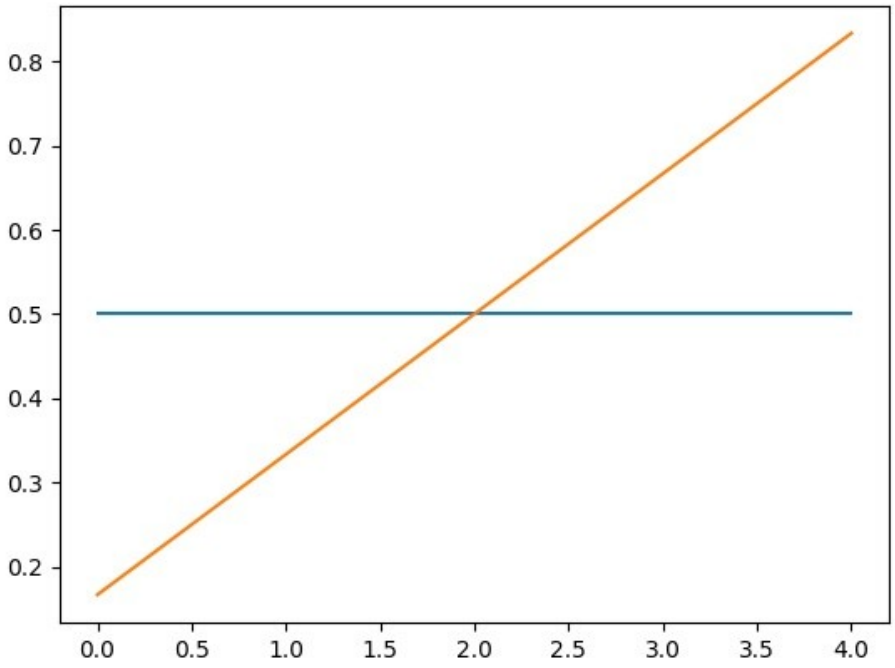
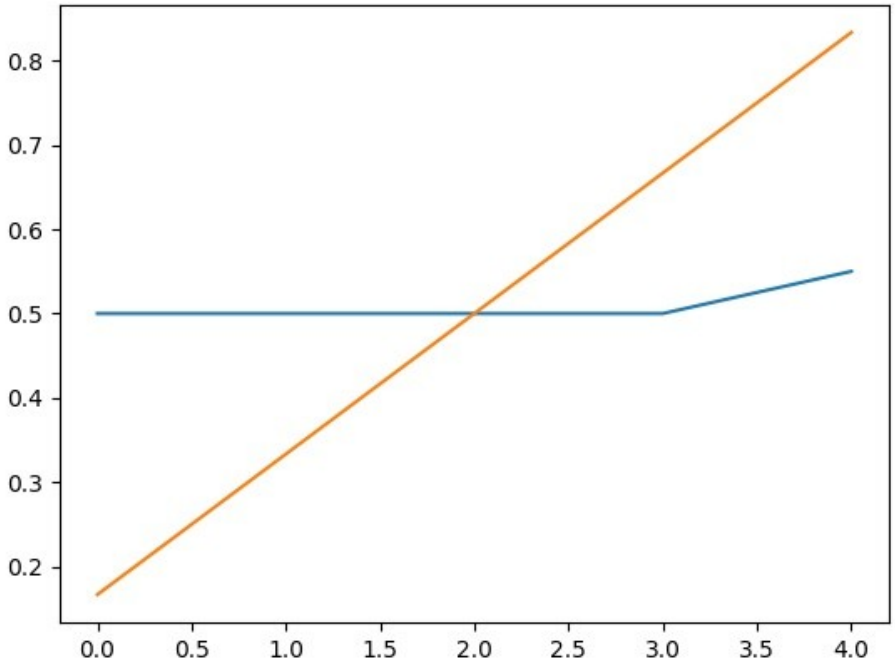
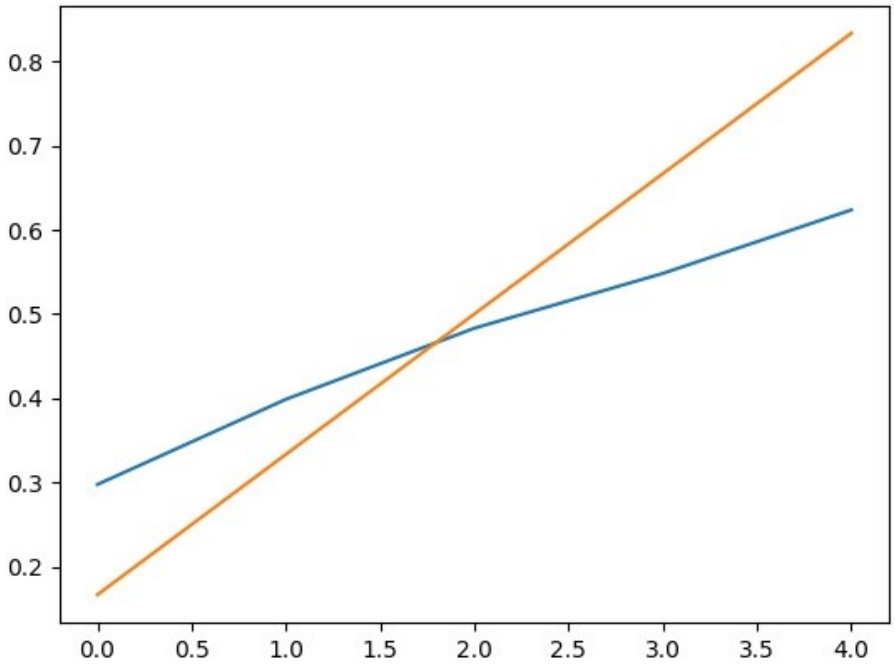
1. 解答如下：
   a) alpah 为 0.1 时，值函数图像如下：
      1. 片段为 0：



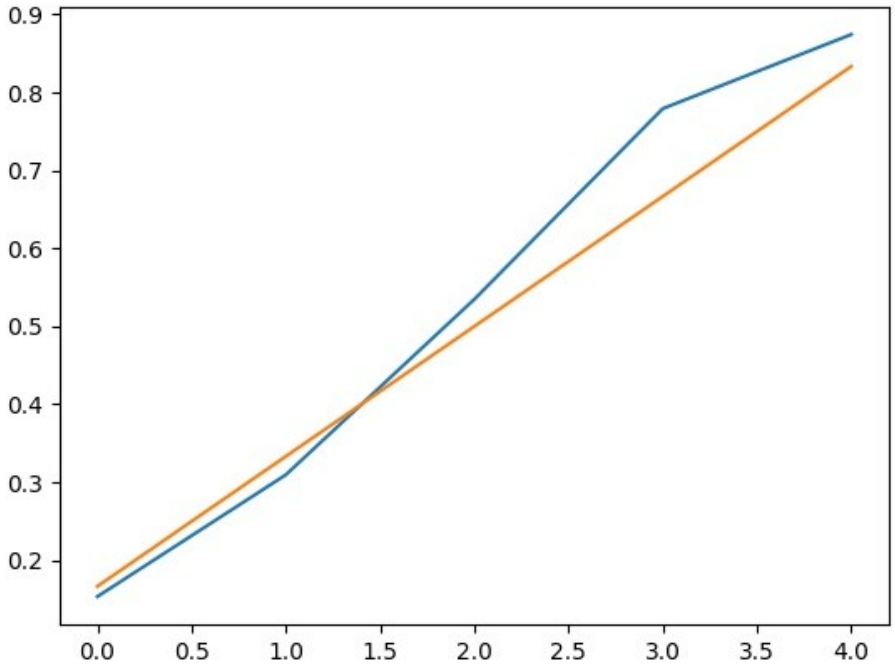
      2. 片段为 1：



      3. 片段为 10：

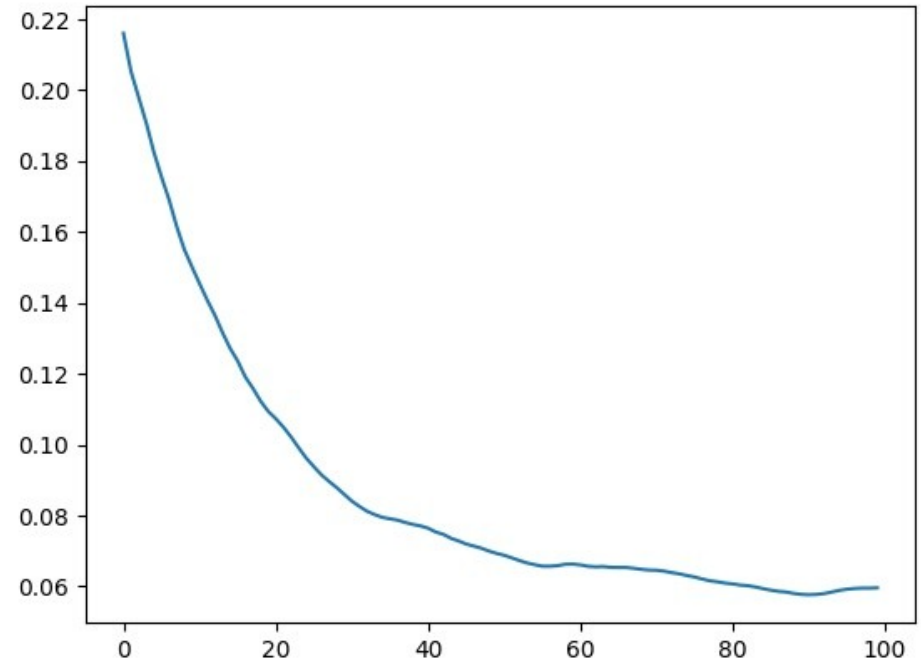


4. 片段为 100：

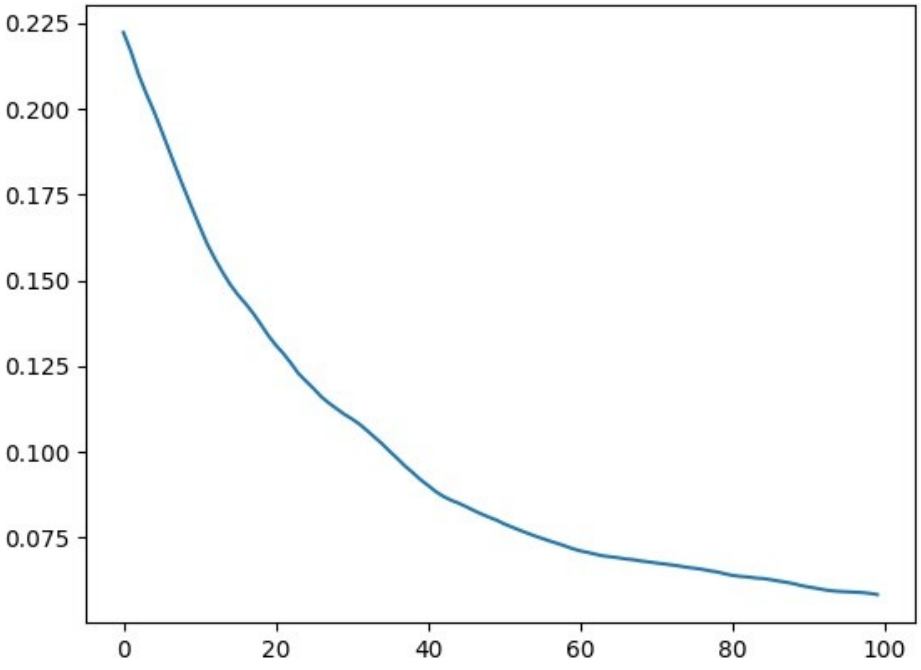


b) TD 算法的 RMS-error 曲线

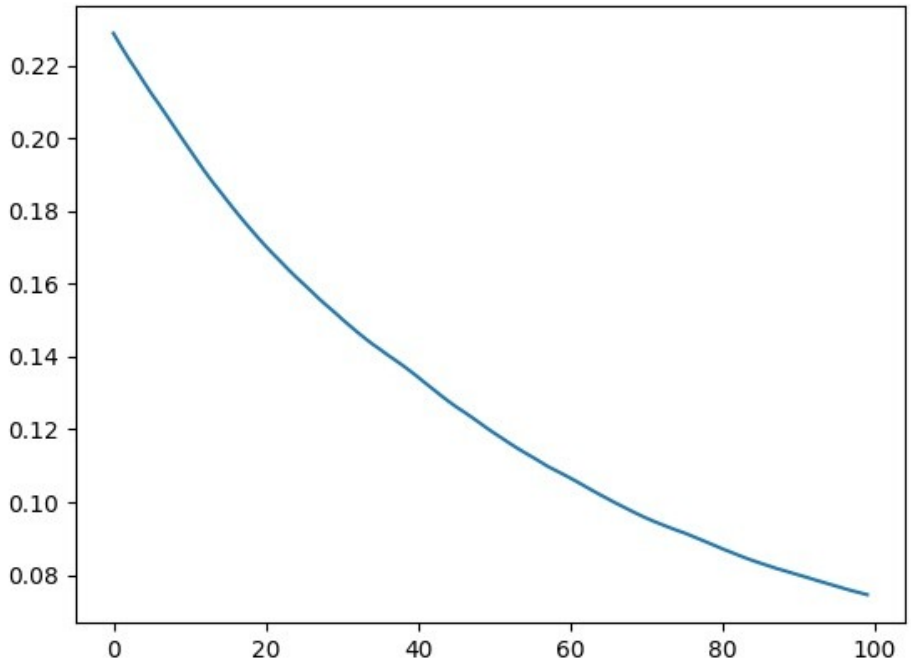
1. alpha 为 0.15：



2. alpha 为 0.1：

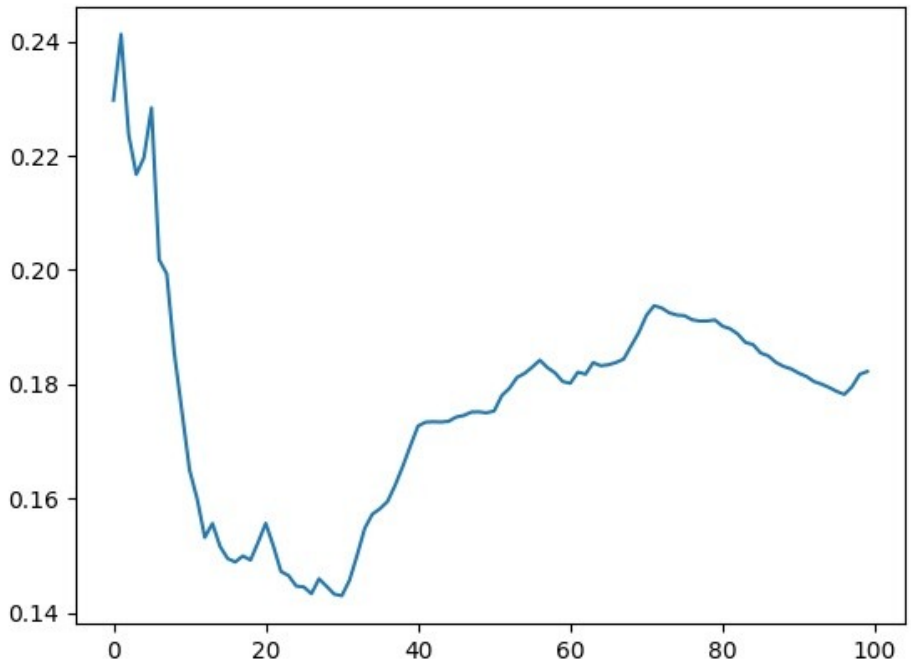


3. alpha 为 0.05：

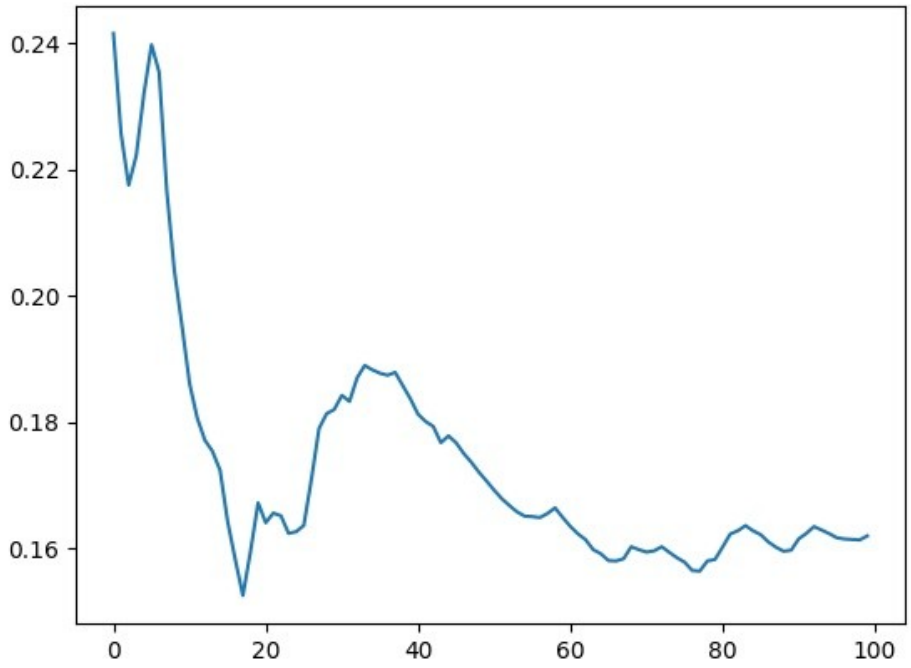


c) MC 算法的 RMS-error 曲线：

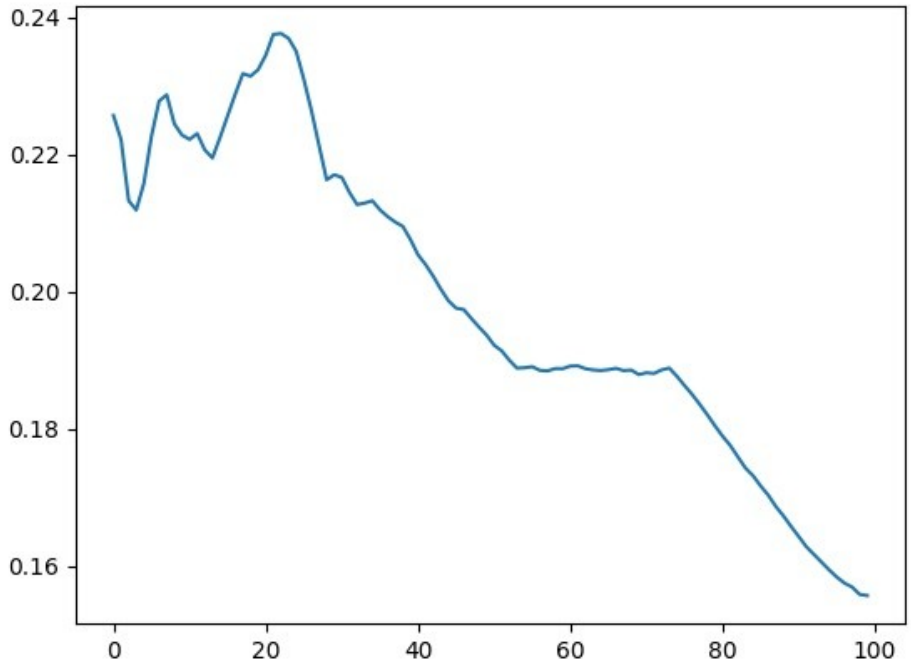
1. alpha 为 0.15：



2. alpha 为 0.1：

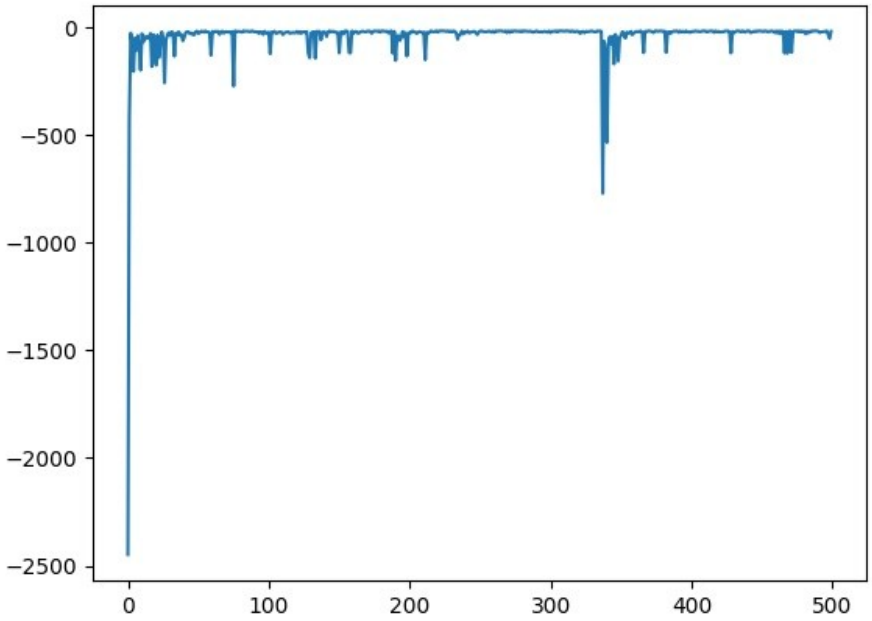


3. alpha 为 0.05：



d) TD 算法收敛速度稍慢，但是收敛稳定。MC 收敛伴随着巨大的方差，alpha 越大结果越差。

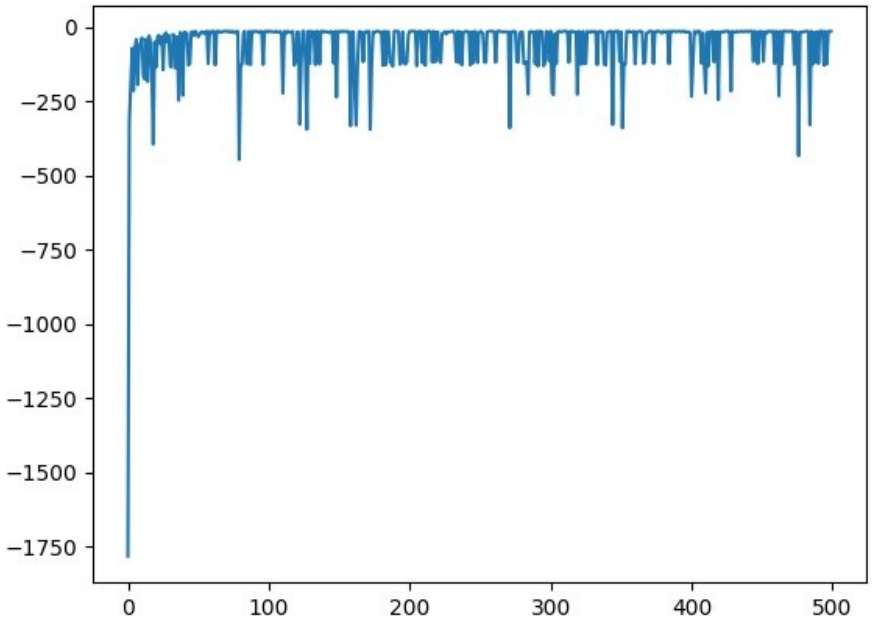
2. TD 优化：
   a) SARSA 回报值变化：



最终策略如下：

➡	➡	➡	⬆	➡	➡	➡	➡	➡	➡	➡	➡
➡	➡	⬆	➡	⬆	⬆	➡	⬆	➡	⬆	➡	⬇
⬆	⬅	➡	⬆	⬆	⬆	⬆	⬆	⬆	⬆	➡	⬇
⬆	Cliff										final

b) Q-learning 回报值变化：



最终策略如下

➡	⬅	➡	⬇	⬆	⬇	➡	⬇	⬇	➡	⬇	⬇
⬇	➡	⬇	➡	➡	➡	➡	➡	⬇	⬇	➡	⬇
➡	➡	➡	➡	➡	➡	➡	➡	➡	➡	➡	⬇
⬆	cliff										final

d) Q 学习回报值更大，并且收敛到了最优策略。因为 Q 学习的目标策略采用的是贪婪的策略函数，在训练过程中方差更小，减少了不必要的随机性，收敛性更好。