1. 解答如下：

(a) 需要，Q 学习中之所以不需要是因为重要性采样率中上下都只包含 $P(S_{t+1},R_{t+1}|S_t,A_t)$，相除之后值正好等于 1。在 n 步 TD 中，则需要用到目标策略和行为策略的采样概率，所以不为 1，需要计算重要性采样。

(b) 策略如下：

1. **repeat**（对于每一个片段）

2. **repeat** 对于片段中的每一步

3. 根据 $\mu(S_t)$ 选择动作 A_t

4. 执行动作 A_t，观察到 R_{t+1}，并将其存储

5. **if** $\tau=t-n+1\geq 0$ **then**

6. $G\leftarrow \sum_{i=\tau+1}^{min(\tau+n,T)} \gamma^{i-\tau-1}R_i$

7. $W\leftarrow \prod_{i=\tau+1}^{min(\tau+n-1,T-1)} \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}$

8. **if** $\tau+n<T$, **then** $G\leftarrow G+\gamma^n Q(S_{\tau+n},A)$

9. $Q(S_\tau,A_\tau)\leftarrow Q(S_\tau,A_\tau)+W\alpha(G-Q(S_\tau,A_\tau))$

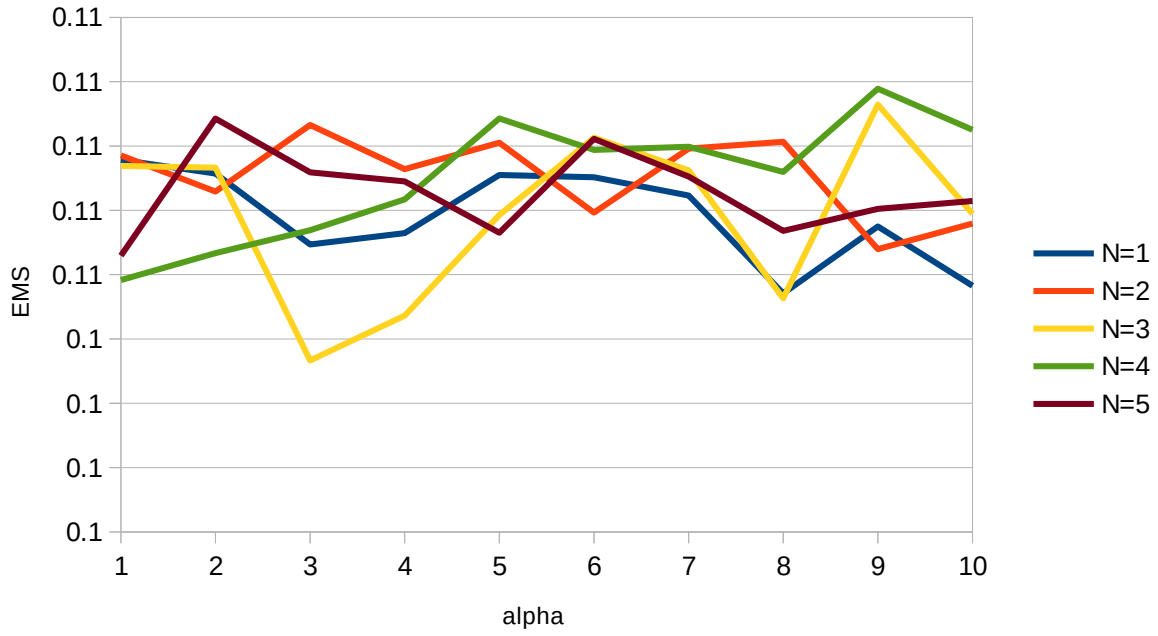
10. **end if**

11. **until** 直到终止状态

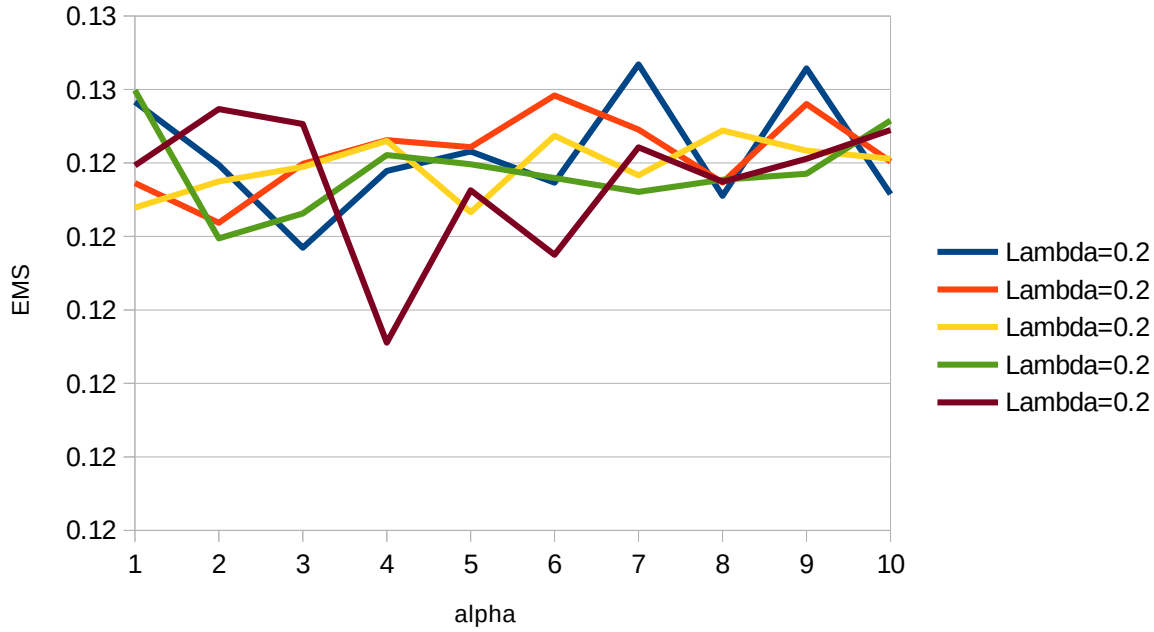
12. **until** 收敛

2. 解答如下：

(a) n 步 TD 的 RMS-alpha 曲线如下：



(b) TD（lambda）的 RMS-alpha 曲线如下：



(c) SARSA（lambda），最佳策略即为所有状态均向右，策略得到的 Q 值如下：

位置	向左	0
1	0.405	0.502607964769372
2	0.495	0.526258025115648
3	0.500017924529375	0.624052510683208
4	0.527129616308887	0.769263282721232
5	0.534122606428263	0.924952682351501