人工智能综合作业 3——MountainCar

陈昭熹 2017011552

2020 年 1 月 16 日

目录

1 引言

本次综合作业选择任务 2，解决离散行动空间与连续行动空间下的小车上山问题 (MountainCar)。对于两个问题一些共同的约定将在本节予以阐述。

1.1 探索空间

Agent 可以感知的环境变量有小车当前位置以及小车速度，两个变量的探索空间定义如下:

Observation	最小值	最大值
Position	-1.2	0.6
Velocity	-0.07	0.07

1.2 初始状态

在本环境中，初始状态采用随机策略，取 $[-0.6, -0.4]$ 的随机位置初始化 agent 状态，且速度为 0 。

1.3 终止状态

终止状态即小车到达预定山顶或迭代步数超过 200，即:

$$\textbf{Position} = 0.5 \ \textit{or} \ \textbf{iteration} > 200 \rightarrow \textbf{terminate} \tag{1}$$

2 任务一 MountainCar-v0

由于该问题属于无模型问题，采用基于行动价值，迭代更新 Q 表的训练方式比较合理。结合课内所学，采用 Q-Learning, Sarsa, 期望 Sarsa 来解决这一问题。

2.1 行动空间

本任务的行动空间是离散的，仅有三个可选行动来控制小车的” 油门”: 向左加速, 向右加速与不加速。

2.2 回报定义

采用较为简单的前进代价式定义，即每走一步给予 agent 一个-1 的回报，直到终止状态。这样的定义与 gym 库中定义相同，因此可以直接使用。

$$reward \leftarrow reward - 1,\ for\ each\ step \tag{2}$$

2.3 ϵ 衰减

为了权衡探索与利用，本文方法对于 $\epsilon - greedy$ 策略中的 ϵ 采取随着训练轮次衰减的策略，从 1 开始逐渐衰减到 0，而非一开始就是一个极小值。这样的做法是为了让训练初期给予 agent 充分探索环境的机会 (ϵ 越大则越可能执行随机策略)，这让训练后期回报的收敛提供必要的条件，避免局部最优解的产生。

2.4 Q-Learning

Q-Learning 属于离线学习的控制方法，其行动策略遵循 $\epsilon - greedy$ 原则，而目标策略则使用贪心策略进行选择。由此有其行动价值递推式:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma max_{a \in A} Q(S_{t+1}, a) - Q(S_t, A_t)) \tag{3}$$

上式中的行动空间 A 与状态空间 S 定义及取值已经在前面的章节给出。在训练过程中，使用上式的方法更新行动价值，并反复迭代直至收敛，将最终的 Q 表存下来，用于回放训练结果。将前 10000 轮训练的平均收益、最大收益、最小收益与片段序号作折线图，得到训练过程的回报变化如下所示:

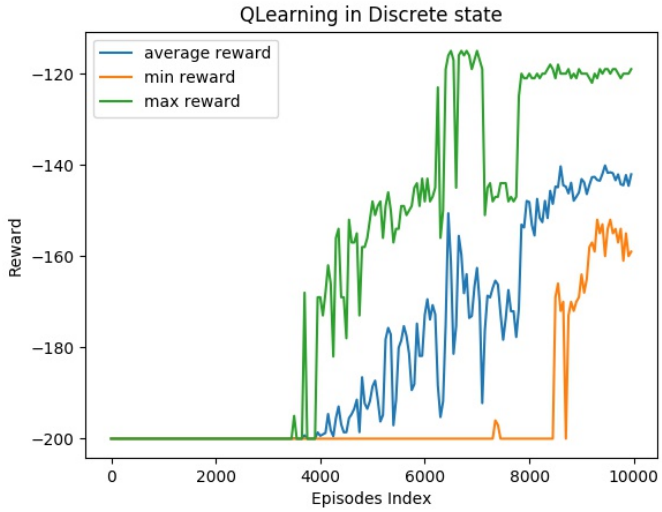

图 1: 任务一 Q-Learning 回报收敛过程

2.5 Sarsa

Sarsa 属于在线学习的控制方法，其行动策略与目标策略均遵循 $\epsilon - greedy$ 原则，且行动策略与目标策略相同，由此产生行动价值递推式:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma Q(S_{t+1}, a) - Q(S_t, A_t)) \tag{4}$$

在训练过程中，使用上式的方法更新行动价值，并反复迭代直至收敛，将最终的 Q 表存下来，用于回放训练结果。将前 10000 轮训练的平均收益、最大收益、最小收益与片段序号作折线图，得到训练过程的回报变化如下所示:

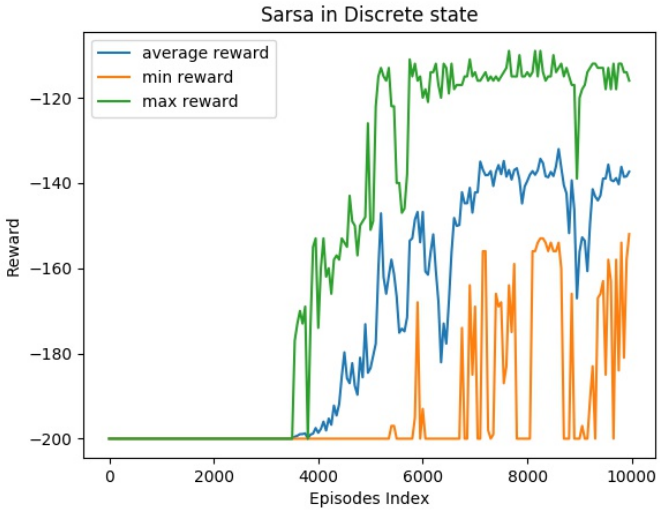

图 2: 任务一 Sarsa 回报收敛过程

对比 Q-Learning 的回报收敛过程，可以发现 Sarsa 在平均回报上并无差别，但在最大回报上明显高于 Q-Learning，这表征了 Sarsa 是一个更加保守的算法，相比于 Q-Learning 的用于探索，Sarsa 更乐于保证状态的”安全”，这一点在任务二3.4.2中更换不同的回报定义后会有明显的反映。同时也可以看到，Sarsa 在最小回报的收敛上要明显早于 Q-Learning，这也说明了 Sarsa 更保守，而 Q-Learning 则更倾向于探索。

2.6 期望 Sarsa

期望 Sarsa 也属于离线学习的控制方法，其行动策略遵循 $\epsilon - greedy$ 原则，而目标策略则取期望意义上的估计值。可以理解为 Sarsa 是单点采样，而期望 Sarsa 则是均匀采样并计算期望。由此有其行动价值递推式:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \sum_{a \in A} \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t)) \quad (5)$$

在训练过程中，使用上式的方法更新行动价值，并反复迭代直至收敛，将最终的 Q 表存下来，用于回放训练结果。将前 10000 轮训练的平均收益、最大收益、最小收益与片段序号作折线图，得到训练过程的回报变化如下所示:

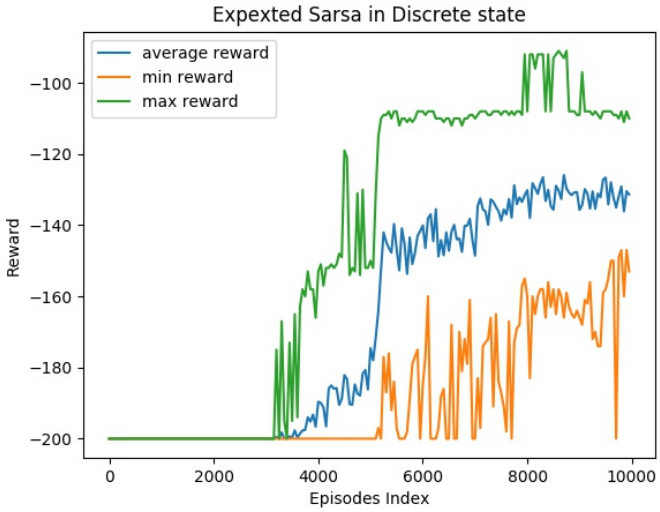

图 3: 任务一 Expected Sarsa 回报收敛过程

从上面的折线图中可以看出，与上面两种方法对比，无论从收敛速度上，从结果上，期望 Sarsa 均优于前面的两种方法。

3 任务二 MountainCarContinuous-v0

3.1 行动空间

本任务的行动空间是连续的，不仅可以决定加速的方向，还可以决定加速的大小，若用 A_i 表示状态 i 下的行动，则其取值应当是全体实数，用其符号区分加速方向，用其绝对值代表加速大小：

$$A_i \in R, \begin{cases} push\ left, A_i < 0 \\ push\ right, A_i > 0 \\ no\ push, A_i = 0 \end{cases} \tag{6}$$

3.2 回报定义

与任务一的回报定义不同，此处到达终点将给予 agent 回报 100，并且计算此路径上花费的代价，因此累计回报可能是正值。

$$reward = \begin{cases} 100 - \sum_{n}^{i=1} {A_i}^2, \ where\ state\ is\ terminated \\ -\sum_{n}^{i=1} {A_i}^2, \ otherwise \end{cases} \tag{7}$$

由于本任务中行动空间是连续的，因此不能直接用上一节所述的基于 Q 表的方法，或者说应当做一定处理才可以使用基于 Q 表的方法。本文最终给出基于离散行动空间的方法，求解时仍用与上一任务类似的训练算法，但是需要将行动空间进行合适的离散化。

3.3 行动空间离散化

为了仍使用基于 Q 表的诸多经典方法，在本问题中可以将行动空间离散化。只要行动空间的分割间隔合适，就不会影响连续行动空间下的小车运动，从而将连续问题转化为离散问题予以解决。经过多次实验验证，最终将本任务中的行动空间离散化为下面的形式:

$$A_i \in \{-2.0, -1.6 - 1.2, -0.8, -0.4, 0, 0.4, 0.8, 1.2, 1.6, 2.0\} \tag{8}$$

这样的处理可以大幅减少计算量，同时让上一章节中的算法可以继续使用。值得注意的是，这样的离散化考虑了物理约束条件，即小车坐标的范围在-1.2 到 0.6 之间，因此行动空间内的最大值与最小值没必要过大，即小车不可能一次加速就冲出地图。

3.4 行动空间离散化后的训练算法

下面的三种算法在原理上与上一章节类似，因此只阐述在本任务中实现上的细节问题。下面的三种方法最终平均回报均收敛在 90 到 100 之间，符合官方 wiki 对于成功解决问题的定义。

3.4.1 Q-Learning

原理见2.4

用类似的方法将回报与片段序列绘制折线图如下所示:

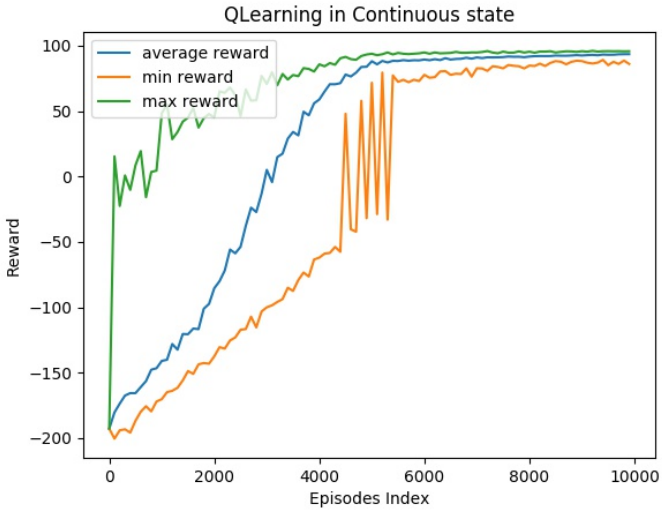

图 4: 任务二 Q-Learning 回报收敛过程

首先可以看出，本任务中回报的定义与上一任务不同，因此最终得到的回报结果会出现正值。而这样的回报定义虽然有些不寻常，但是从其收敛过程上可以看到，它能够帮助 agent 在短时间内更快的对环境建立正向认知，并朝着目标方向稳步努力 (在前 4000 轮中平均回报近乎线性上升)。这是由于目标点的大额回报值能够更好的传递到前面的状态中，让智能体更好的感知到任务目标。从其收敛后期可以看到，最小最大和平均收益几乎收敛到相同的值，而不像上一任务中仍存在较大差异。

3.4.2 Sarsa

原理见2.5 上文提到过，Sarsa 是一个保守的方法，而在本任务的回报定义下，显然对于不善于探索的方法是不友好的，极有可能让这类方法陷入局部最优解，而无法及时发现目标状态。在最初的实验中，使用 Sarsa 方法就遇到了类似的问题，迭代 10000 轮后最终智能体认为停留在原点仿佛是回报最高的选择，得到其回报收敛过程如下:

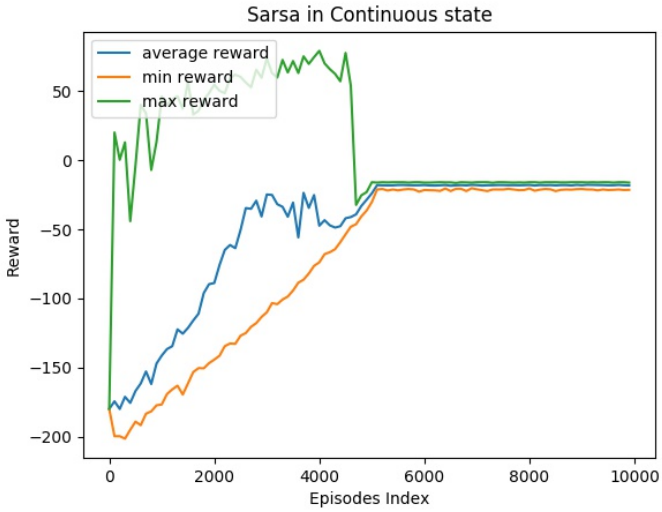

图 5: 任务二 憨憨的 Sarsa 回报收敛过程

而相同的参数对于前一节的 Q-Learning 则不会出现陷入局部最优的问题。为此对于本任务中 Sarsa 的训练过程进行参数调整，增强其” 探索性”，将行动空间的范围扩大，同时缩减 ϵ 衰减系数，让确定策略出现的更晚，并获得了如下结果:

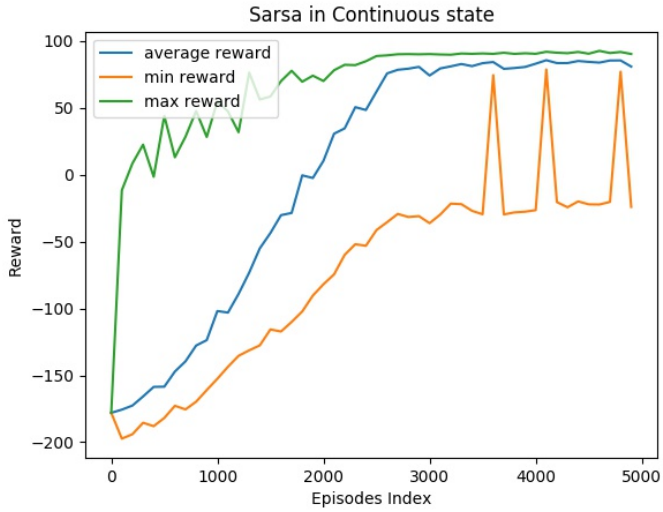

图 6: 任务二 Sarsa 回报收敛过程

从这样的对比试验中，更能看出 Sarsa 与 Q-Learning 的区别，同时也凸显了回报的定义对于强化学习训练过程的重要影响。

3.4.3 期望 Sarsa

原理见2.6

用类似的方法将回报与片段序列绘制折线图如下所示:

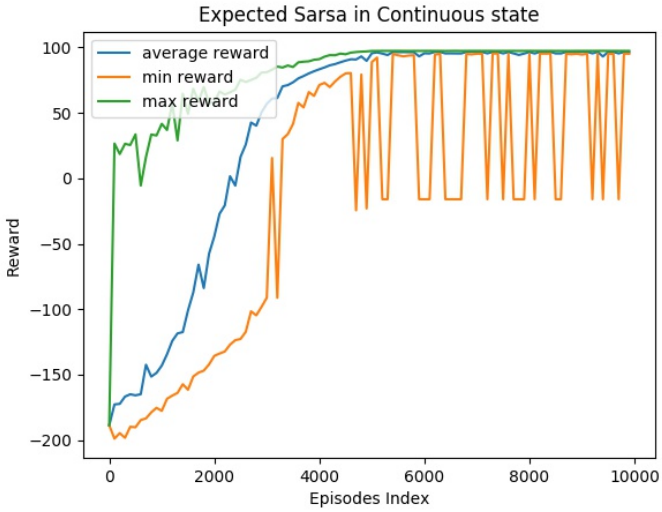

图 7: 任务二 Expected Sarsa 回报收敛过程

从收敛的后期过程可以看出，期望 Sarsa 能够做到平均期望与最大期望基本一致，也就是说几乎每一个片段的选择都是最优的。而这样的性能是前面 Q-Leanring 和 Sarsa 所做不到的。在本任务中三种方法的收敛速度相同，而 Expected Sarsa 的性能显然更优。