强化学习部分定理证明

讲师：张伟楠　助教：王锡淮

2023 年 7 月 29 日

目录

1 ϵ-greedy 算法探索与利用的遗憾界

我们先给出 K-摇臂赌博机（bandit）的定义如下：

定义 1.1 (K-摇臂赌博机)**.** *K-摇臂赌博机可以表示成一个三元组 $< K, \mathcal{A}, T >$，其中 K 代表玩家可以拉动 K 个不同的拉杆，即玩家具有 K 个动作，$\mathcal{A}$ 则表示动作空间 $\{1, \ldots, K\}$，T 表示该问题的时间长度。在每一时刻 t：*

1. *玩家选择动作 $a_t \in \mathcal{A}$，即拉动杠杆 a_t；*
2. *玩家观测到奖励 $r_t \in [0, 1]$。*

关于 $\epsilon - greedy$ 算法，我们将只讨论随机 K-摇臂赌博机（stochastic bandit）问题设定下的奖励函数。在随机 K-摇臂赌博机问题中，对于拉杆 $a \in \mathcal{A}$，得到的奖励是独立同分布的，记该分布为 D_a，且该分布不能被玩家得知。在上述设定下，时刻 t 时，玩家选择拉杆 a_t，并观测到从分布 D_{a_t} 采样得到的奖励 r_t。

将拉杆 a 的期望奖励表示为 $\mu_a = \mathbb{E}[D_a]$,并将所有拉杆中最高的期望奖励记为 $\mu^* = \max_{a \in \mathcal{A}} \mu_a$。则玩家在 K-摇臂赌博机问题中的目标为最大化在时间长度 T 内收集到的奖励和，即 $\sum_{i=1}^{T} r_t$。于是最优的奖励和可以由每个时刻选择最优的拉杆得到。我们引入遗憾（regret）这一概念，来衡量与最优奖励和的差距，记为：

$$R(t) = \mu^* t - \sum_{i=1}^{t} \mu_{a_i} \ 。$$

于是最大化奖励和 $\sum_{i=1}^{T} r_t$ 等价于最小化在时间长度 T 内的遗憾 $R(T)$。同时需要注意的是 $R(T)$ 也是一个依赖于每一时刻选择的动作以及可能受奖励分布和算法的随机性影响的随机变量，所以在分析算法的遗憾界（regret bound）时要考虑的是遗憾的期望值，即 $\mathbb{E}[R(T)]$。

下面给出 ϵ-greedy 算法的伪代码。

在算法 1中，时刻 t 时，对于拉杆 $a \in \mathcal{A}$，$\hat{\mu}_a$ 和 c_a 分别记录了截至时刻 t 拉杆 a 获得的平均奖励和被选择的次数。

关于平均奖励 $\hat{\mu}_a$，我们可以利用霍夫丁不等式的一种变式得到平均奖励 $\hat{\mu}_a$ 与期望奖励 μ_a 之间的关系。

引理 1.1 (霍夫丁不等式)**.** *令 $X_1, \ldots, X_n$ 为独立的随机变量，且 $X_i \in [a_i, b_i]$, $\forall i = 1, \ldots, n$。则这些随机变量的均值可以表示为 $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$。霍夫丁不等式的一种形式可为：*

$$\forall \sigma > 0, P(|\bar{X} - \mathbb{E}[\bar{X}]| \geq \sigma) \leq 2e^{-\frac{2n^2\sigma^2}{\sum_{i=1}^{n}(b_i - a_i)^2}} \ 。$$

霍夫丁不等式的证明请见[1]。

我们分析时刻 t，在前 t 个时刻的探索阶段中共有 $\sum_{i=1}^{t} \epsilon_i$ 次探索行为出现，其中每个拉杆被选中的平均次数为 $\frac{\sum_{i=1}^{t} \epsilon_i}{K}$。有霍夫丁不等式可以得到平均奖励 $\hat{\mu}_a$ 与期望奖励 μ_a 之间的关系：

$$\forall \sigma > 0, P(|\hat{\mu}_a - \mu_a| \leq \sigma) \geq 1 - 2e^{-2\sigma^2(\frac{\sum_{i=1}^{t} \epsilon_i}{K})}$$

[1] http://cs229.stanford.edu/extra-notes/hoeffding.pdf

算法 - 1: ϵ-greedy 算法

输入: 摇臂数 K，时间长度 T。

输出: 奖励和 R。

for 时刻 $t = 1, \ldots, T$ **do**
  $R = 0$;
  初始化 $\hat{\mu}_a = 0, c_a = 0 \; \forall a = 1, \ldots, K$;
  if *rand()* $< \epsilon$ **then**
    探索：从 K 个拉杆中以均匀分布随机选取拉杆 k;
  else
    利用：$k = \arg\max_a \hat{\mu}_a$;
  end
  观测到当前奖励 r_t;
  $R = R + r_t$;
  $\hat{\mu}_a = \frac{\hat{\mu}_a * c_k + r_t}{c_k + 1}$;
  $c_k = c_k + 1$;
end

选择 $\sigma = \sqrt{\frac{K \log(2t)}{2 \sum_{i=1}^{t} \epsilon_i}}$，则

$$P(|\hat{\mu}_a - \mu_a| \leq \sqrt{\frac{K \log(2t)}{2 \sum_{i=1}^{t} \epsilon_i}}) \geq 1 - \frac{1}{t} \text{ 。} \tag{1.1}$$

当时间长度 T 较大时，公式 (1.1) 中的事件会以较大概率发生，我们下面考虑对所有拉杆 $a \in \mathcal{A}$，公式 (1.1) 中的事件都发生的场景，记为事件 E_1。

令 a^* 为具有期望奖励为 μ^* 的拉杆，可能存在其他拉杆 $a \neq a^*$，在探索阶段得到的平均奖励 $\hat{\mu}_a > \hat{\mu}_{a^*}$，因而拉杆 a 在利用阶段被选择。由公式 (1.1) 可得：

$$\mu_a + \sqrt{\frac{K \log(2t)}{2 \sum_{i=1}^{t} \epsilon_i}} \geq \hat{\mu}_a > \hat{\mu}_{a^*} \geq \mu_{a^*} - \sqrt{\frac{K \log(2t)}{2 \sum_{i=1}^{t} \epsilon_i}} \text{ 。}$$

于是，$\mu_{a^*} - \mu_a \leq 2\sqrt{\frac{K \log(2t)}{2 \sum_{i=1}^{t} \epsilon_i}}$。所以时刻 t 的期望遗憾界可以表示为：

$$\begin{aligned} & P(\text{探索}) \times 1 + P(\text{利用}) \times (\mu_{a^*} - \mu_a) \\ = \; & \epsilon_t + (1 - \epsilon_t) \times 2\sqrt{\frac{K \log(2t)}{2 \sum_{i=1}^{t} \epsilon_i}} \\ \leq \; & \epsilon_t + 2\sqrt{\frac{K \log(2t)}{2 \sum_{i=1}^{t} \epsilon_i}} \text{ ,} \end{aligned}$$

其中，探索中随机采样一个拉杆产生的遗憾界是 1，而在利用阶段，我们只考虑事件 E_1 发生的情况下的遗憾，因为事件 E_1 不发生的概率满足 $O(\frac{1}{t})$ 而且事件 E_1 不发生的遗憾界是 1，不会影响整体遗憾界的计算。

整体的遗憾则可以表示为:

$$\mathbb{E}[R(t)] = \mathbb{E}\left[\sum_{i=1}^{t}\left(\epsilon_i + 2\sqrt{\frac{K\log(2t)}{2\sum_{i=1}^{t}\epsilon_i}}\right)\right] = \sum_{i=1}^{t}\epsilon_i + 2t\sqrt{\frac{K\log(2t)}{2\sum_{i=1}^{t}\epsilon_i}} \text{ 。} \tag{1.2}$$

于是我们得到下面的定理。

定理 1.1 (ϵ-greedy 算法的遗憾界)**.** *对于在时刻 i 以 ϵ_i 概率进行探索的 ϵ-greedy 算法，其在前 t 时刻的遗憾界为 $\sum_{i=1}^{t}\epsilon_i + 2t\sqrt{\frac{K\log(2t)}{2\sum_{i=1}^{t}\epsilon_i}}$。*

注意到懊悔界中 $\sum_{i=1}^{t}\epsilon_i$ 与第一项正相关,与第二项负相关,可以令 $\sum_{i=1}^{t}\epsilon_i = 2^{\frac{1}{3}}t^{\frac{2}{3}}(K\log(2t))^{\frac{1}{3}}$, 得到最小化的懊悔界为 $2^{\frac{4}{3}}t^{\frac{2}{3}}(K\log(2t))^{\frac{1}{3}}$，也可以写为 $\tilde{O}(t^{\frac{2}{3}}K^{\frac{1}{3}})$。$\tilde{O}$ 是忽略了 log 因子的粗略复杂度估计方式[2]。

[2]详细证明可以参考 Slivkins Aleksandrs, Introduction to multi-armed bandits, `http://arxiv.org/abs/1904.07272`。

2 占用度量与策略的一一对应

定义 2.1 (马尔可夫决策过程). 首先给出马尔可夫决策过程 (*Markov Decision Process*, *MDP*) 的定义。一个 *MDP* 可由五元组 $< \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma >$ 表示，其中：

- $\mathcal{S}$ 代表状态空间；
- $\mathcal{A}$ 代表状态空间；
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ 是状态转移函数；
- $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ 是奖励函数；
- $\gamma \in (0, 1)$ 是折扣因子。

为了方便证明[3]，我们考虑离散情况下的占用度量 (occupancy measure)。记策略 $\pi : \mathcal{S} \times \mathcal{A} \mapsto \Omega(\mathcal{A})$ 下得到的占用度量为 ρ^π，则

$$\rho^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}(s_t = s, a_t = a)\right], \forall s \in \mathcal{S}, a \in \mathcal{A} \tag{2.1}$$

那么一个策略的优化目标可以写成

$$\max_\pi \mathbb{E}_\pi[r(s, a)] = \max_\pi \sum_{s,a} r(s, a)\rho^\pi(s, a) \ ,$$

其中对任意 $s \in \mathcal{S}, a \in \mathcal{A}$，满足下列条件：

$$\sum_a \rho^\pi(s, a) = P(s_0 = s) + \gamma \sum_{s',a} \rho^\pi(s', a)\mathcal{T}(s', a, s) \ , \tag{2.2}$$

$$\rho^\pi(s, a) \geq 0 \ 。 \tag{2.3}$$

公式 (2.2) 和公式 (2.3) 构成了贝尔曼流约束（Bellman Flow Constraints)。

下面的定理表明了占用度量和策略一一对应的关系。

定理 2.1 (占用度量和策略一一对应). 令 ρ 为满足贝尔曼流约束的占用度量，π 为一个策略，其中 $\pi(a|s) = \frac{\rho(s,a)}{\sum_a \rho(s,a)}, \forall s \in \mathcal{S}, a \in \mathcal{A}$，则 ρ 是 π 的占用度量；反过来，如果 π 是一个策略且 ρ 是对应的占用度量，那么 π 满足 $\pi(a|s) = \frac{\rho(s,a)}{\sum_a \rho(s,a)}, \forall s \in \mathcal{S}, a \in \mathcal{A}$，且 ρ 满足贝尔曼流约束。

在证明定理 2.1之前，我们需要证明两个引理。先引入一个新的约束，特定策略的贝尔曼流约束（π-specific Bellman Flow Constraints)：对任意 $s \in \mathcal{S}, a \in \mathcal{A}$，占用度量 ρ 满足下列条件：

$$\rho(s, a) = \pi(a|s)P(s_0 = s) + \pi(a|s)\gamma \sum_{s',a'} \rho(s', a')\mathcal{T}(s|s', a') \tag{2.4}$$

$$\rho(s, a) \geq 0 \tag{2.5}$$

[3]详细证明可以参考 Syed, Umar, Michael Bowling, and Robert E. Schapire, Apprenticeship learning using linear programming 定理 2 的证明或者 Feinberg, Eugene A., and Adam Shwartz, eds, Handbook of Markov decision processes: methods and applications, p. 178。

引理 2.1. 对于任意策略 π，其对应的占用度量 ρ^π 满足 π 对应的特定策略的贝尔曼流约束。

证明. 公式 (2.5) 显然满足。对于公式 (2.4)，使用占用度量的定义公式 (2.1)，可以得到：

$$\begin{aligned}
\rho^\pi(s,a) &= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{I}(s_t=s,a_t=a)\right] \\
&= \sum_{t=0}^{\infty}\gamma^t P(s_t=s,a_t=a) \\
&= \pi(a|s)P(s_0=s)+\sum_{t=0}^{\infty}\gamma^{t+1}P(s_{t+1}=s,a_{t+1}=a) \\
&= \pi(a|s)P(s_0=s)+\sum_{t=0}^{\infty}\gamma^{t+1}\sum_{s',a'}P(s_t=s',a_t=a',s_{t+1}=s,a_{t+1}=a) \\
&= \pi(a|s)P(s_0=s)+\sum_{t=0}^{\infty}\gamma^{t+1}\sum_{s',a'}P(s_t=s',a_t=a')\mathcal{T}(s|s',a')\pi(a|s) \\
&= \pi(a|s)P(s_0=s)+\gamma\sum_{s',a'}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{I}(s_t=s,a_t=a)\right]\mathcal{T}(s|s',a')\pi(a|s) \\
&= \pi(a|s)P(s_0=s)+\pi(a|s)\gamma\sum_{s',a'}\rho^\pi(s',a')\mathcal{T}(s|s',a')
\end{aligned}$$

□

下面我们证明满足特定策略的贝尔曼流约束的占用度量是唯一的。

引理 2.2. 任意策略 π 所对应的特定策略的贝尔曼流约束有至多一个解。

证明. 定义大小为 $|\mathcal{SA}|\times|\mathcal{SA}|$ 的矩阵 A，其中每一项为:

$$A[(s,a),(s',a')] = \begin{cases} 1 - \gamma\mathcal{T}(s|s',a')\pi(a|s) & \text{if } (s,a)=(s',a') \\ - \gamma\mathcal{T}(s|s',a')\pi(a|s) & \text{otherwise} \end{cases} 。$$

定义大小为 $|\mathcal{SA}|$ 的向量 b，其中每一项为 $b[(s,a)]=\pi(a|s)P(s_0=s)$，以及大小为 $|\mathcal{SA}|$ 的向量 x，其中每一项为 $x[(s,a)]=\rho(s,a)$，注意 A，b，x 均是被状态-动作对索引的矩阵或者向量。那么特定策略的贝尔曼流约束可以等价表示为

$$Ax = b \tag{2.6}$$

$$x \geq 0 \tag{2.7}$$

因为 $\sum_{s\in\mathcal{S}}\mathcal{T}(s|s',a')=1$，$\sum_a\pi(a|s)=1$ 且我们只考虑 $\gamma<1$ 的情况，所以对于任意 s',a'，有

$$\begin{array}{lll}
\sum_{s,a}\gamma\mathcal{T}(s|s',a')\pi(a|s)=\gamma & < & 1 \\
\Rightarrow \quad 1-\gamma\mathcal{T}(s'|s',a') & > & \sum_{(s,a)\neq(s',a')}\gamma\mathcal{T}(s|s',a')\pi(a|s) \\
\Rightarrow \quad |A_{(s',a'),(s',a')}| & > & \sum_{(s,a)\neq(s',a')}|A_{(s,a),(s',a')}|
\end{array}$$

可见矩阵 A 是对角占优矩阵，故其为非奇异矩阵，于是由公式 (2.6) 和公式 (2.7) 构成的线性系统至多有一个解。所以满足策略 π 所对应约束的占用度量至多有一个。 □

注 2.1. 引理 *2.1*表明任意 π 对应的策略特定的贝尔曼流约束有解，同时引理 *2.2*表明任意 π 对应的策略特定的贝尔曼流约束解唯一，故占用度量 ρ 满足 π 对应的策略特定的贝尔曼流约束是 ρ 为 π 对应的占用度量的充要条件。

现在我们开始证明定理 2.1。

证明. 对于证明 π 能确定一个占用度量 ρ，首先有

$$\pi(a|s) = \frac{\rho(s,a)}{\sum_a \rho(s,a)} = \frac{\rho(s,a)}{P(s_0 = s) + \gamma \sum_{s',a} \rho(s',a)\mathcal{T}(s',a,s)} \tag{2.8}$$

由公式 (2.8) 可知 ρ 满足 π 对应的策略特定贝尔曼流约束，故由引理 2.1和引理 2.2（注 2.1）可知 ρ 是 π 对应的占用度量。

对于证明占用度量 ρ 能确定一个策略 π，由引理 2.1可知 ρ 满足 π 对应的特定策略的贝尔曼流约束，故满足公式 (2.4)，因而可以得到公式 (2.8)，证得 $\pi(a|s) = \frac{\rho(s,a)}{\sum_a \rho(s,a)}$。此外对于公式 (2.4)，两边对 $\forall a \in \mathcal{A}$ 求和可得公式 (2.2)，故 ρ 也满足贝尔曼流约束。 □

3 贝尔曼最优性算子的不动点

继续使用 MDP 的定义 2.1，则贝尔曼最优性方程（Bellman Optimality Equation）可以表示为：

$$V^*(s) = \max_{a\in\mathcal{A}} r(s,a) + \gamma \mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V^*(s')\right] \text{ 。} \tag{3.1}$$

于是我们将贝尔曼最优性算子定义成：

$$BV(s) = \max_{a\in\mathcal{A}} r(s,a) + \gamma \mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V(s')\right] \text{ 。} \tag{3.2}$$

则贝尔曼最优性算子的不动点 V^* 满足 $BV^*(s) = V^*(s)$。

下面我们将用到压缩映射定理（contraction mapping theorem）来分析贝尔曼最优性算子[4]。首先我们介绍需要使用的数学工具：度量空间（metric space）和压缩映射（contraction mapping）。一个度量空间可以表示为一个二元组 (X,d)，其中 X 为一个集合，d 为定义在 X 上的度量。而在度量空间 (X,d) 上的一个压缩映射 $f: X\mapsto X$ 满足 $d(f(x),f(y)) \le \theta d(x,y)$，其中 $\theta\in(0,1)$。

对于分析贝尔曼最优性算子，为了便于理解，我们考虑离散的状态和动作空间。对于每一个 $s\in\mathcal{S}$，贝尔曼最优性算子 B 将 $V(s)\in\mathbb{R}$ 映射到 $BV(s)\in\mathbb{R}$。故对于贝尔曼最优性算子 $B: \mathbb{R}^{|\mathcal{S}|}\mapsto\mathbb{R}^{|\mathcal{S}|}$，我们选择 $X=\mathbb{R}^{|\mathcal{S}|}$，选择 $d(X) = \|X\| = \|X\|_\infty = max_{i\in[0,|X|]}|X_i|$ 为无穷范数。在该定义下，(X,d) 中的任意柯西列的极限仍属于 (X,d)，该度量空间完备。

下面证明贝尔曼最优性算子为 (X,d) 上的压缩映射。

引理 3.1 (贝尔曼最优性算子为压缩映射)**.** *对于任意两个值函数 $V_1, V_2\in\mathbb{R}^{|\mathcal{S}|}$，有*

$$\|BV_1 - BV_2\| \le \gamma\|V_1 - V_2\| \text{ 。}$$

证明.

$$\begin{aligned}
\|BV_1 - BV_2\| &= \max_{s\in\mathcal{S}}\left|\max_{a\in\mathcal{A}}\left(r(s,a)+\gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s'|s,a)V_1(s')\right) - \max_{b\in\mathcal{A}}\left(r(s,b)+\gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s'|s,b)V_2(s')\right)\right| \\
&\le \max_{s\in\mathcal{S},a\in\mathcal{A}}\left|\left(r(s,a)+\gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s'|s,a)V_1(s')\right) - \left(r(s,a)+\gamma\sum_{s'\in\mathcal{S}}\mathcal{T}(s'|s,a)V_2(s')\right)\right| \\
&= \gamma\max_{s\in\mathcal{S},a\in\mathcal{A}}\left|\sum_{s'\in\mathcal{S}}\mathcal{T}(s'|s,a)\left(V_1(s') - V_2(s')\right)\right| \\
&\le \gamma\max_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\mathcal{T}(s'|s,a)\left|\left(V_1(s') - V_2(s')\right)\right| \\
&\le \gamma\max_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\mathcal{T}(s'|s,a)\max_{s''\in\mathcal{S}}\left|\left(V_1(s'') - V_2(s'')\right)\right| \\
&= \gamma\max_{s''}\left|\left(V_1(s'') - V_2(s'')\right)\right| = \gamma\|V_1 - V_2\|
\end{aligned}$$

其中第二行外提了一个对动作求最大值的操作来进行放大，第六行利用了 $\sum_{s'\in S}\mathcal{T}(s'|s,a) = 1$。

□

[4]关于压缩映射定理的证明和其他细节可以参考https://web.stanford.edu/class/math51h/contraction.pdf。

下面介绍压缩映射定理。

定理 3.1 (压缩映射定理)**.** *假设 (X,d) 是一个完备度量空间，且 $f:X\mapsto X$ 是一个压缩映射，则 f 具有唯一的不动点 $x\in X$。*

由上文的分析可知，当度量空间定义为 $(\mathbb{R}^{|S|},\|\cdot\|_\infty)$ 时，该度量空间为完备度量空间，且贝尔曼最优性算子 B 是压缩映射，故由压缩映射定理可知对于一个 MDP，存在不动点，即最优值函数 V^*。

4 策略提升定理

定理 4.1 (策略提升定理)**.** 考虑任意两个策略 π，π'，定义

$$Q^{\pi}(s,\pi') = \mathbb{E}_{a\sim\pi'(a|s)}\left[Q^{\pi}(s,a)\right] \text{ 。}$$

如果对于 $\forall s \in \mathcal{S}$，都有 $Q^{\pi}(s,\pi') \geq V^{\pi}(s)$，那么 $V^{\pi'}(s) \geq V^{\pi}(s)$。

证明.

$$\begin{aligned}
V^{\pi}(s) &\leq Q^{\pi}\left(s,\pi'\right) \\
&= \mathbb{E}_{a\sim\pi'(a|s),s'\sim\mathcal{T}(s'|s,a)}\left[r(s,a)+\gamma V^{\pi}\left(s'\right)\right] \\
&\leq \mathbb{E}_{a\sim\pi'(a|s),s'\sim\mathcal{T}(s'|s,a)}\left[r(s,a)+\gamma Q^{\pi}\left(s',\pi'\right)\right] \\
&= \mathbb{E}_{a,a'\sim\pi'}\left[r(s,a)+\gamma r\left(s',a'\right)+\gamma^{2}V^{\pi}\left(s''\right)\right] \\
&\quad\vdots \\
&\leq \mathbb{E}_{a,a',a''\ldots\sim\pi'}\left[r(s,a)+\gamma r\left(s',a'\right)+\gamma^{2}r\left(s'',a''\right)+\ldots\right] \\
&= V^{\pi'}(s)
\end{aligned}$$

□

注 4.1. 定理 *4.1*指出了一种能让价值函数单调提升的方法,即使得对于 $\forall s \in \mathcal{S}$,都有 $Q^{\pi}(s,\pi') \geq V^{\pi}(s)$，一个贪心的方法是令 $Q^{\pi}(s,\pi') = \max_{a\in\mathcal{A}} Q^{\pi}(s,a) \geq \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi}(s,a)\right] = V^{\pi}(s)$，也即令 $\pi'(\cdot|s) = \arg\max_{a\in\mathcal{A}} Q^{\pi}(s,a)$。

我们下面证明 ϵ-greedy 策略也能使得价值函数单调提升。

推论 4.1 (ϵ-greedy 策略的单调提升)**.** 对于 *ϵ-greedy* 的策略 π 满足:

$$\pi(a|s) = \begin{cases} \frac{\epsilon}{|\mathcal{A}|} + 1 - \epsilon & \text{if } a = \arg\max_{a'} Q^{\pi}(s,a') \\ \frac{\epsilon}{|\mathcal{A}|} & \text{otherwise} \end{cases},$$

而且 $Q^{\pi}(s,\pi') = \sum_{a\in\mathcal{A}} \pi'(a|s) Q^{\pi}(s,a)$，则 $V^{\pi'}(s) \geq V^{\pi}(s)$。

证明.

$$\begin{aligned}
Q^{\pi}(s,\pi') &= \sum_{a\in\mathcal{A}} \pi'(a|s) Q^{\pi}(s,a) \\
&= \frac{\epsilon}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon)\max_{a\in\mathcal{A}} Q^{\pi}(s,a) \\
&= \frac{\epsilon}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon)\max_{a\in\mathcal{A}} Q^{\pi}(s,a)\frac{1-\epsilon}{1-\epsilon} \\
&= \frac{\epsilon}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon)\max_{a\in\mathcal{A}} Q^{\pi}(s,a)\frac{\sum_{a\in\mathcal{A}}\pi(a|s) - \sum_{a\in\mathcal{A}}\frac{\epsilon}{|\mathcal{A}|}}{1-\epsilon}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\epsilon}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon)\max_{a\in\mathcal{A}} Q^{\pi}(s,a)\sum_{a\in\mathcal{A}}\frac{\pi(a|s)-\frac{\epsilon}{|\mathcal{A}|}}{1-\epsilon} \\
&\geq \frac{\epsilon}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} Q^{\pi}(s,a) + (1-\epsilon)\sum_{a\in\mathcal{A}} Q^{\pi}(s,a)\frac{\pi(a|s)-\frac{\epsilon}{|\mathcal{A}|}}{1-\epsilon} \\
&= \sum_{a\in\mathcal{A}}\pi(a|s)Q^{\pi}(s,a) = V^{\pi}(s)
\end{aligned}
$$

于是 $Q^{\pi}(s,\pi') \geq V^{\pi}(s)$，由策略提升定理可知 $V^{\pi'}(s) \geq V^{\pi}(s)$。 □

5 Q-learning 收敛性

学界对 Q-learning 的收敛性有严谨的证明[5]，我们这里给出一个较为简洁和精确的证明。在开始证明前，我们对 MDP 的性质做出一些假设。

假设 5.1. *我们对 MDP 作出下列必要假设：*

- *MDP 的状态空间 $\mathcal{S}$ 和动作空间 $\mathcal{A}$ 有限；*
- *任意状态空间对 $(s,a)\ \forall s\in\mathcal{S}, a\in\mathcal{A}$ 都被无限次数访问；*
- *奖励函数与下一时刻的状态无关*[6]*。*

然后我们分析对于 Q 函数的贝尔曼算子的性质，此处我们稍微滥用符号 $\mathcal{B}$，记作用于 Q 函数的贝尔曼算子为 $\mathcal{B}$，满足：

$$\mathcal{B}Q(s,a)=r(s,a)+\gamma\mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[\max_{a'}Q(s',a')\right]$$

使用与章节 3类似的技巧，我们可以知道作用于 Q 函数的算子 $\mathcal{B}$ 也是一个 γ-压缩映射，且类似的由压缩映射定理，我们知道存在不动点 Q^*，满足 $\mathcal{B}Q^*=Q^*$。

至此我们证明了作用于 Q 函数的贝尔曼算子能收敛到最优 Q 函数，然而 Q-learning 并不是该贝尔曼算子的直接应用，因为在 Q-learning 中 Q 函数在各时刻被随机采样得到的样本更新。于是我们需要用到随机逼近（stochastic approximation）对 Q-learning 的收敛性进行分析。我们先介绍一下需要用到的随机逼近相关的术语和概念。从随机逼近的角度来看，Q-learning 在寻找方程 $\mathcal{B}Q-Q=0$ 的解，即 Q^*。问题中时刻 t 时存在的随机性来源于状态转移函数以及当前时刻奖励的随机性[7]。于是该随机逼近的更新规则变为：

$$Q_{t+1}(s,a)=Q_t(s,a)+a_t(s,a)\left(r_t(s,a)+\gamma\mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[\max_{a'}Q_t(s',a')\right]-Q_t(s,a)\right)$$

关于类似结构的随机逼近更新规则，有下面的收敛性质：

引理 5.1. *给定随机逼近更新规则的结构为：*

$$\Delta_{t+1}(x)=\Delta_t(x)-a_t(x)\left(\Delta_t(x)-F_t(x)\right)=(1-a_t(x))\Delta_t(x)+a_t(x)F_t(x)\ ,$$

其中 x 为要估计的变量，$\Delta_t(x)$ 和 $F_t(x)$ 为时刻 t 时 x 对应的两个函数值，$a_t(x)$ 为时刻 t 下 x 对应的学习率。同时给定下列假设：

1. 对于任意变量 x 和时刻 t，学习率满足：$\sum_t a_t(x)=\infty$，$\sum_t a_t(x)^2<\infty$ 且 $0\le a_t(x)\le 1$；

2. $\|\mathbb{E}\left[F_t\right]\|_\infty\le\gamma\|\Delta_t\|_\infty$，其中 $\gamma\in(0,1)$；

[5]较为经典的证明包括 Tsitsiklis J N. Asynchronous stochastic approximation and Q-learning，Jaakkola T, Jordan M, Singh S. Convergence of stochastic iterative dynamic programming algorithms 和 Melo F S. Convergence of Q-learning: A simple proof。

[6]如果奖励函数与下一时刻的状态相关，证明的过程会有一些改变，见 Tsitsiklis J N. Asynchronous stochastic approximation and Q-learning。

[7]为了便于理解，我们不涉及随机过程和测度论的知识，完整的使用随机过程和测度论的证明可见上文提过的严谨证明。

3. 对于任意变量 x 和时刻 t，$Var\left(F_t(x)\right) \leq C\left(1+\|\Delta_t\|_\infty^2\right)$，*for* $C>0$。

则 $\Delta_t(x)$ 以 1 概率收敛到 0。

引理 5.1的证明可见 Tsitsiklis J N. Asynchronous stochastic approximation and Q-learning。

对于随机逼近 Q-learning，我们定义

$$\Delta_t(s,a) = Q_t(s,a) - Q^*(s,a)$$

以及

$$F_t(s,a) = r(s,a) + \gamma \max_{b\in\mathcal{A}} Q_t(s',b) - Q^*(s,a)\ ,$$

其中 Q^* 的存在性在上文中提到，由贝尔曼算子的压缩性质和压缩映射定理保证。于是我们得到更新规则为

$$\Delta_{t+1}(s,a) = (1-a_t(s,a))\,\Delta_t(s,a) + a_t(s,a)F_t(s,a)\ 。 \tag{5.1}$$

利用引理 5.1，我们将证明下面关于 Q-learning 算法收敛性的定理。

定理 5.1 (Q-learning 收敛)**.** 假设状态空间 $\mathcal{S}$ 和动作空间 $\mathcal{A}$ 有限且离散，同时假设学习率满足 $\sum_t a_t(s,a)=\infty$，$\sum_t a_t(s,a)^2<\infty, \forall (s,a)\in\mathcal{S}\times\mathcal{A}$。此外，假设 $r(s,a)>0, Q_0(s,a)>0, \forall s,a$。在此假设下，$Q_t(s,a)\geq 0, \forall s,a\ t$。于是对于公式 (5.1)，$Q$ 函数以概率 1 收敛到最优 Q 函数。

证明. 引理 5.1中的假设 1 显然成立，现在只需证明公式 (5.1) 满足假设 2 和 3。前面提到时刻 t 时 Q-learning 过程的随机性来自于对时刻 t 奖励的未知和时刻 $t+1$ 状态依赖于状态转移函数的随机性，故对 $\forall s,a$，

$$\begin{aligned}
\mathbb{E}\left[F_t(s,a)\right] &= \mathbb{E}\left[r(s,a) + \gamma \max_{b\in\mathcal{A}} Q_t(s',b) - Q^*(s,a)\right] \\
&= \mathbb{E}\left[r(s,a) + \gamma \max_{b\in\mathcal{A}} Q_t(s',b)\right] - Q^*(s,a) \\
&= \mathcal{B}Q_t(s,a) - Q^*(s,a) \\
&= \mathcal{B}Q_t(s,a) - \mathcal{B}Q^*(s,a)
\end{aligned}$$

如上文所述，作用于 Q 函数的贝尔曼算子 $\mathcal{B}$ 是一个 γ-压缩映射，故

$$\begin{aligned}
\|\mathbb{E}\left[F_t\right]\|_\infty &= \max_{s,a} |\mathcal{B}Q_t(s,a) - \mathcal{B}Q^*(s,a)| \\
&= \|\mathcal{B}Q_t - \mathcal{B}Q^*\|_\infty \\
&\leq \gamma \|Q_t - Q^*\|_\infty \\
&= \gamma \|\Delta_t\|_\infty
\end{aligned}$$

至此我们证明了 Δ_t 的更新规则满足假设 2。下面证明其满足假设 3：对 $\forall s\in\mathcal{S}, a\in\mathcal{A}$，

$$\begin{aligned}
Var(F_t(s,a)) &= \mathbb{E}\left[\left(F_t(s,a) - \mathbb{E}\left[F_t(s,a)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(r(s,a) + \gamma \max_{b\in\mathcal{A}} Q_t(s',b) - Q^*(s,a) - (\mathcal{B}Q_t(s,a) - Q^*(s,a))\right)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(r(s,a)+\gamma\max_{b\in\mathcal{A}}Q_t(s',b)-\left(r(s,a)+\gamma\mathbb{E}_{x\sim\mathcal{T}(\cdot|s,a)}\left[\max_b Q_t(x,b)\right]\right)\right)^2\right]\\
&= \gamma\mathbb{E}\left[\left(\max_{b\in\mathcal{A}}Q_t(s',b)-\mathbb{E}_{x\sim\mathcal{T}(\cdot|s,a)}\left[\max_b Q_t(x,b)\right]\right)^2\right]\\
&= \gamma^2 Var(\max_{b\in\mathcal{A}}Q_t(s',b))\\
&= \gamma^2\mathbb{E}\left[\left(\max_{b\in\mathcal{A}}Q_t(s',b)\right)^2\right]-\gamma^2\left(\mathbb{E}_{x\in\mathcal{S}}\left[\max_{b\in\mathcal{A}}Q_t(x,b)\right]\right)^2\\
&\le \gamma^2\mathbb{E}\left[\left(\max_{b\in\mathcal{A}}Q_t(s',b)\right)^2\right]\\
&\le \gamma^2\max_{x\in\mathcal{S}}\max_{b\in\mathcal{A}}\left(Q_t(x,b)\right)^2\\
&\le \gamma^2\|\Delta_t+Q^*\|_\infty^2\\
&= \gamma^2\|\Delta_t\|_\infty^2+2\gamma^2\|\Delta_t\|_\infty\|Q^*\|_\infty+\gamma^2\|Q^*\|_\infty^2
\end{aligned}
$$

注意到奖励函数有界，则 $\|Q^*\|_\infty$ 非负且有界。下面分类讨论：

- $\|\Delta_t\|_\infty \le 1$，则

$$\gamma^2\|\Delta_t\|_\infty^2+2\gamma^2\|\Delta_t\|_\infty\|Q^*\|_\infty+\gamma^2\|Q^*\|_\infty^2 \le \gamma^2\|\Delta_t\|_\infty^2+2\gamma^2\|Q^*\|_\infty+\gamma^2\|Q^*\|_\infty^2\ ;$$

- $\|\Delta_t\|_\infty > 1$，那么 $\|\Delta_t\|_\infty \le \|\Delta_t\|_\infty^2$，则

$$\gamma^2\|\Delta_t\|_\infty^2+2\gamma^2\|\Delta_t\|_\infty\|Q^*\|_\infty+\gamma^2\|Q^*\|_\infty^2 \le \gamma^2\left(1+2\|Q^*\|_\infty\right)\|\Delta_t\|_\infty^2+\gamma^2\|Q^*\|_\infty^2\ 。$$

选择 $C = max\{\gamma^2(1+2\|Q^*\|_\infty), 2\gamma^2\|Q^*\|_\infty+\gamma^2\|Q^*\|_\infty^2\}$，则

$$
\begin{aligned}
Var(F_t(s,a)) &= \gamma^2\|\Delta_t\|_\infty^2+2\gamma^2\|\Delta_t\|_\infty\|Q^*\|_\infty+\gamma^2\|Q^*\|_\infty^2\\
&\le C\left(1+\|\Delta_t\|_\infty^2\right)\ 。
\end{aligned}
$$

至此我们证明了假设 3 成立。于是由引理 5.1可知 Δ_t 以概率 1 收敛到 0，即 Q_t 以概率 1 收敛到 Q^*。 □

6 策略梯度定理

在本章节我们讨论参数化且处处可微的策略 π_θ，其中 θ 为对应的参数。

定理 6.1 (策略梯度定理)**.** 对于任意 *MDP*，定义策略学习的目标为 $\mathcal{J}(\theta)=\mathbb{E}_{s_0}\left[V^{\pi_\theta}(s_0)\right]$，定义状态访问度量 $\rho^{\pi_\theta}(s)=\sum_{t=0}^{\infty}\gamma^t P^{\pi_\theta}(s_t=s)=\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{s_0}\left[P^{\pi_\theta}(s_t=s|s_0)\right]$，则

$$\nabla_{\pi_\theta}\mathcal{J}(\theta)=\sum_{s\in\mathcal{S}}\rho^{\pi_\theta}(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(s,a)\ 。$$

证明.

$$\begin{aligned}
&\nabla_\theta V^{\pi_\theta}(s)\\
=\ &\nabla_\theta\left(\textstyle\sum_{a\in\mathcal{A}}\pi_\theta(a|s)Q^{\pi_\theta}(s,a)\right)\\
=\ &\textstyle\sum_{a\in\mathcal{A}}\left(\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\pi_\theta(a|s)\nabla_\theta Q^{\pi_\theta}(s,a)\right)\\
=\ &\textstyle\sum_{a\in\mathcal{A}}\left(\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\pi_\theta(a|s)\nabla_\theta\sum_{s'\in\mathcal{S}}p(s'|s,a)\left(r(s,a)+\gamma V^{\pi_\theta}(s')\right)\right)\\
=\ &\textstyle\sum_{a\in\mathcal{A}}\left(\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\gamma\pi_\theta(a|s)\sum_{s'\in\mathcal{S}}p(s'|s,a)\nabla_\theta V^{\pi_\theta}(s')\right) &\text{(a)}\\
=\ &\textstyle\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\gamma\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\pi_\theta(a|s)p(s'|s,a)\nabla_\theta V^{\pi_\theta}(s')\\
=\ &\textstyle\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(s,a)+\gamma\sum_{s'\in\mathcal{S}}\rho^{\pi_\theta}(s\to s',1)\nabla_\theta V^{\pi_\theta}(s') &\text{(b)}\\
=\ &\textstyle\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(s,a)\\
&\textstyle+\gamma\sum_{s'\in\mathcal{S}}\rho^{\pi_\theta}(s\to s',1)\left(\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s')Q^{\pi_\theta}(s',a)+\gamma\sum_{s''\in\mathcal{S}}\rho^{\pi_\theta}(s'\to s'',1)\nabla_\theta V^{\pi_\theta}(s'')\right) &\text{(c)}\\
=\ &\phi(s)+\gamma\textstyle\sum_{s'\in\mathcal{S}}\rho^{\pi_\theta}(s\to s',1)\phi(s')+\gamma^2\sum_{s'\in\mathcal{S}}\rho^{\pi_\theta}(s\to s',1)\sum_{s''\in\mathcal{S}}\rho^{\pi_\theta}(s'\to s'',1)\nabla_\theta V^{\pi_\theta}(s'') &\text{(d)}\\
=\ &\phi(s)+\gamma\textstyle\sum_{s'\in\mathcal{S}}\rho^{\pi_\theta}(s\to s',1)\phi(s')+\gamma^2\sum_{s''\in\mathcal{S}}\left(\sum_{s'\in\mathcal{S}}\rho^{\pi_\theta}(s\to s',1)\right)\rho^{\pi_\theta}(s'\to s'',1)\nabla_\theta V^{\pi_\theta}(s'')\\
=\ &\phi(s)+\gamma\textstyle\sum_{s'\in\mathcal{S}}\rho^{\pi_\theta}(s\to s',1)\phi(s')+\gamma^2\sum_{s''\in\mathcal{S}}\rho^{\pi_\theta}(s\to s'',2)\nabla_\theta V^{\pi_\theta}(s'')\\
&\cdots\\
=\ &\textstyle\sum_{k=0}^{\infty}\gamma^k\sum_{s'}\rho^{\pi_\theta}(s\to s',k)\phi(s')\\
=\ &\textstyle\sum_{s'}\sum_{k=0}^{\infty}\gamma^k\rho^{\pi_\theta}(s\to s',k)\phi(s')
\end{aligned}$$

在上面的推导中：

(a) ： 因为 $\nabla_\theta r(s,a)=0$;

(b) ： 定义 $\rho^{\pi_\theta}(x\to y,k)$ 为策略 π_θ 从状态 x 出发 k 步到状态 y 的概率，同时令 $\forall x,y\in\mathcal{S},y\neq x,\rho^{\pi_\theta}(x\to y,0)=0$ 且 $\rho^{\pi_\theta}(x\to x,0)=1$;

(c) ： 继续展开 $\nabla_\theta V^{\pi_\theta}(s')$;

(d) ： 定义 $\phi(s)=\sum_{a\in\mathcal{A}}\nabla_\theta\pi_\theta(a|s)Q^{\pi_\theta}(s,a)$。

故有：

$$\begin{aligned}
\nabla_\theta\mathcal{J}(\pi_\theta)&=\nabla_\theta\mathbb{E}_{s_0}\left[V^{\pi_\theta}(s_0)\right]\\
&=\mathbb{E}_{s_0}\left[\sum_{s\in\mathcal{S}}\sum_{t=0}^{\infty}\gamma^t\rho^{\pi_\theta}(s_0\to s,t)\phi(s)\right]\\
&=\mathbb{E}_{s_0}\left[\sum_{s\in\mathcal{S}}\sum_{t=0}^{\infty}\gamma^t P^{\pi_\theta}(s_t=s|s_0)\phi(s)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_0} \left[P^{\pi_\theta}(s_t = s | s_0) \right] \phi(s) \\
&= \sum_{s \in \mathcal{S}} \rho^{\pi_\theta}(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s, a)
\end{aligned}
$$

□

7 TRPO 的单调性提升

7.1 朴素策略梯度存在的问题

我们已经证明了对于 $\mathcal{J}(\pi_\theta) = \mathbb{E}_{\tau\sim(\mu,\pi_\theta)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$，其中 μ 为初始状态 s_0 的分布，即 $s_0 \sim \mu(\cdot)$，策略梯度为

$$\nabla_\theta \mathcal{J}(\pi_\theta) = \sum_{s\in\mathcal{S}} \rho^{\pi_\theta}(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s,a) \ ,$$

其中 $\rho^{\pi_\theta}(s) = \sum_{t=0}^{\infty}\gamma^t P^{\pi_\theta}(s_t = s)$ 为状态访问度量。那么参数 θ 的更新可以表示为 $\theta_{t+1} = \theta_t + \alpha_t \nabla_\theta \mathcal{J}(\pi_{\theta_t})$。

朴素的策略梯度存在以下的问题：

- **步长难以确定**。a）因为策略的更新依赖于用于采样的策略得到的状态访问度量，于是策略梯度的步长不仅影响当下更新得到的策略，也会极大地影响后续更新，如果过大的步长导致更新得到的策略表现较差，后续的更新依赖于该较差策略采样得到的数据，将增大获得更优的后续策略的难度；b）步长难以确定的另一个原因是策略梯度只关注了参数 θ 所在空间的距离，而没有关注策略空间的距离，在参数空间的极小变化可以会引起策略的巨大改变，这令策略梯度的步长更难确定。

- **样本效率低**。由于策略更新后将对应不同的状态和奖励分布，用旧策略收集到的数据相对于更新后的策略已经过时，用这些数据计算得到的梯度与更新后的策略不匹配，于是每更新策略一次都要重新收集一次数据，导致了较低的样本效率。如果不重新收集数据，简单地使用重要性采样（Importance Sampling）会给策略梯度带来较大的方差。

针对这两个朴素策略梯度的缺点，一个更好的策略优化算法应该满足以下条件：

- 能使用策略之间的距离而不是参数之间的距离来决定步长，并能够选取合适的步长使得新策略的表现优于旧策略的表现。
- 能重复利用采样得到的数据，提高样本效率；

7.2 衡量策略间的差异

基于以上的两个要求，我们需要衡量两个策略之间的差异以研究策略间的关系。首先我们分析两个策略在表现上的差异，下面我们证明策略差分引理（Policy Difference Lemma）。

引理 7.1 (策略差分)**.** 对于任意两个策略 π，$\bar{\pi}$，下面的等式成立：

$$\mathcal{J}(\bar{\pi}) - \mathcal{J}(\pi) = \mathbb{E}_{(s,a)\sim(\rho^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s,a)\right] \ , \tag{7.1}$$

其中优势函数 $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V^\pi(s')\right] - V^\pi(s)$。

证明.

$$\mathcal{J}(\bar{\pi}) - \mathcal{J}(\pi) = \mathcal{J}(\bar{\pi}) - \mathbb{E}_{s_0\sim\mu}\left[V^\pi(s_0)\right]$$

$$
\begin{aligned}
&= \mathcal{J}(\bar{\pi}) - \mathbb{E}_{\tau\sim(\mu,\bar{\pi})}\left[V^{\pi}(s_0)\right] && \text{(a)}\\
&= \mathcal{J}(\bar{\pi}) - \mathbb{E}_{\tau\sim(\mu,\bar{\pi})}\left[\sum_{t=0}^{\infty}\gamma^t V^{\pi}(s_t) - \sum_{t=1}^{\infty}\gamma^t V^{\pi}(s_t)\right]\\
&= \mathcal{J}(\bar{\pi}) - \mathbb{E}_{\tau\sim(\mu,\bar{\pi})}\left[\sum_{t=0}^{\infty}\gamma^t V^{\pi}(s_t) - \sum_{t=0}^{\infty}\gamma^{t+1} V^{\pi}(s_{t+1})\right]\\
&= \mathbb{E}_{\tau\sim(\mu,\bar{\pi})}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] - \mathbb{E}_{\tau\sim(\mu,\bar{\pi})}\left[\sum_{t=0}^{\infty}\gamma^t V^{\pi}(s_t) - \sum_{t=0}^{\infty}\gamma^{t+1} V^{\pi}(s_{t+1})\right]\\
&= \mathbb{E}_{\tau\sim(\mu,\bar{\pi})}\left[\gamma^t\left(r(s_t,a_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)\right)\right]\\
&= \mathbb{E}_{\tau\sim(\mu,\bar{\pi})}\left[\gamma^t A^{\pi}(s_t,a_t)\right]\\
&= \mathbb{E}_{(s,a)\sim(\rho^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right] \ ,
\end{aligned}
$$

其中 (a) 处注意到 π 和 $\bar{\pi}$ 不会影响初始状态 s_0 的分布，且 $V^{\pi}(s_0)$ 与 τ 中除 s_0 外的状态跟动作无关。 □

注 7.1 (策略差分引理的作用)**.** 公式 (7.1) 给出了间接更新策略 $\bar{\pi}$ 的方法，或者说给出了寻找替代更新目标的方法：将 $\bar{\pi}$ 视作更新后得到的策略，π 视作当前策略/旧策略，则公式 (7.1) 将更新后策略的表现拆解出了一项不可优化的 $\mathcal{J}(\pi)$，而优化另一项 $\mathbb{E}_{(s,a)\sim(\rho^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right]$ 则能间接优化 $\mathcal{J}(\bar{\pi})$。$\mathbb{E}_{(s,a)\sim(\rho^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right]$ 目前也是不可优化的，下面我们关注于如何利用该项寻找能间接更新 $\mathcal{J}(\bar{\pi})$ 的替代更新目标。

从引理 7.1看出，我们现在可以使用旧策略的优势函数来估计更新后策略的表现，然而该估计来作为更新目标不可行，因为该估计仍然依赖于使用更新后策略采样的数据。要做到高效率地使用已有数据，还需要对进一步设计更新目标。可以使用重要性采样将更新目标变为：

$$
\begin{aligned}
\mathcal{J}(\bar{\pi}) - \mathcal{J}(\pi) &= \mathbb{E}_{(s,a)\sim(\rho^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right]\\
&= \mathbb{E}_{(s,a)\sim(\rho^{\bar{\pi}},\pi)}\left[\frac{\bar{\pi}(a_t|s_t)}{\pi(a_t|s_t)}A^{\pi}(s_t,a_t)\right]
\end{aligned}
$$

然而，该更新目标仍然依赖于更新后策略的状态访问度量。那么一个直接的想法是使用 ρ^{π} 近似 $\rho^{\bar{\pi}}$，那么更新目标变成了

$$
\mathcal{L}_{\pi}(\bar{\pi}) = \mathbb{E}_{(s,a)\sim(\rho^{\pi},\pi)}\left[\frac{\bar{\pi}(a_t|s_t)}{\pi(a_t|s_t)}A^{\pi}(s_t,a_t)\right] = \mathbb{E}_{(s,a)\sim(\rho^{\pi},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right] \ 。
$$

剩下的问题变成了如何控制 ρ^{π} 和 $\rho^{\bar{\pi}}$ 的相似程度/度量使得替代更新目标 $\mathcal{L}_{\pi}(\bar{\pi})$ 能够有效地优化 $\mathcal{J}(\bar{\pi})$。TRPO 通过证明 $\mathcal{J}(\pi) + \mathcal{L}_{\pi}(\bar{\pi})$ 能够构成 $\mathcal{J}(\bar{\pi})$ 的一个下界来回答这个问题。

为了证明这一点，我们需要衡量两个策略本身的差异，下面介绍要用到的数学工具。

全变差距离 (Total variation distance)。两个在同一空间 Ω 的分布 u 和 v 的全变差距离为

$$
D_{TV}(u\|v) = \|u - v\|_{TV} = \frac{1}{2}\sum_{x\in\Omega}|u(x) - v(x)|
$$

耦合 (Coupling)。耦合是用来分析两个分布概率上相似程度的有力工具。

定义 7.1 (耦合)**.** 现有空间 Ω 上的随机变量 X 和 Y，分别服从于分布 u 和 v。则一个在 $\Omega \times \Omega$ 上的耦合 w 满足：

$$\forall x \in \Omega, \quad \sum_{y \in \Omega} w(x, y) = u(x)$$
$$\forall y \in \Omega, \quad \sum_{x \in \Omega} w(x, y) = v(y)$$

耦合有如下性质[8]：

引理 7.2. 对于任意耦合 w，有 $D_{TV}(u\|v) \leq P(X \neq Y)$。

引理 7.3. 存在一个耦合 w，满足 $D_{TV}(u\|v) = P(X \neq Y)$。

基于耦合的概念，我们定义 α-耦合的策略对。

定义 7.2 (α-耦合的策略对)**.** 对于策略空间上的两个策略 π 和 $\bar{\pi}$，一个 α-耦合的策略对 $(\pi, \bar{\pi})$ 满足

$$\forall s \in \mathcal{S}, P(a \neq \bar{a}|s) \leq \alpha \ ,$$

其中 $a \sim \pi(\cdot|s), \bar{a} \sim \bar{\pi}(\cdot|s)$。

注意到，我们可以选择合适的 α 使得对任意策略对 π 和 $\bar{\pi}$，α-耦合的策略对 $(\pi, \bar{\pi})$ 始终存在。定义 $D_{TV}^{max}(\pi\|\bar{\pi}) = \max_{s \in \mathcal{S}} D_{TV}(\pi(\cdot|s)\|\bar{\pi}(\cdot|s))$，有如下推论：

推论 7.1. 对于任意策略对 π 和 $\bar{\pi}$，$D_{TV}^{max}(\pi\|\bar{\pi})$-耦合的策略对始终存在。

证明. 从引理 7.3可知，对任意 $s \in \mathcal{S}, a \sim \pi(\cdot|s), \bar{a} \sim \bar{\pi}(\cdot|s)$，存在耦合 $(\pi(\cdot|s), \bar{\pi}(\cdot|s))$，满足

$$P(a \neq \bar{a}|s) = D_{TV}(\pi(\cdot|s)\|\bar{\pi}(\cdot|s)) \leq D_{TV}^{max}(\pi\|\bar{\pi}) \ 。$$

上式对任意 $s \in \mathcal{S}$ 都成立，于是集合所有 s 以及对应的耦合，可知存在策略耦合 $(\pi, \bar{\pi})$，满足

$$\forall s \in \mathcal{S}, P(a \neq \bar{a}|s) \leq D_{TV}^{max}(\pi\|\bar{\pi}) \ 。$$

□

至此我们得到了衡量两个策略本身差异程度的方法。

更进一步，基于耦合策略对的性质，我们可以利用 π 和 $\bar{\pi}$ 的关系得到 $|\mathbb{E}_{a \sim \bar{\pi}}[A^{\pi}(s, a)]|$ 的范围，证明如下：

引理 7.4. 给定任意策略 π 和 $\bar{\pi}$，如果 $(\pi, \bar{\pi})$ 是 α-耦合策略对，下面的不等式成立：

$$|\mathbb{E}_{a \sim \bar{\pi}}[A^{\pi}(s, a)]| \leq 2\epsilon\alpha \ ,$$

其中 $\alpha = D_{TV}^{max}(\bar{\pi}\|\pi)$，$\epsilon = \max_{s,a} |A^{\pi}(s, a)|$。

证明. 注意到 $\mathbb{E}_{a \sim \pi}[A^{\pi}(s, a)] = 0$，故

$$|\mathbb{E}_{a \sim \bar{\pi}}[A^{\pi}(s, a)]| = |\mathbb{E}_{\bar{a} \sim \pi}[A^{\pi}(s, \bar{a})] - \mathbb{E}_{a \sim \pi}[A^{\pi}(s, a)]|$$

[8] 关于耦合性质的证明可见https://courses.cs.duke.edu/spring13/compsci590.2/slides/lec5.pdf。

$$
\begin{aligned}
&= \left|\mathbb{E}_{(\bar{a},a)\sim(\bar{\pi},\pi)}\left[A^{\pi}(s,\bar{a}) - A^{\pi}(s,a)\right]\right| \\
&= \left|P(\bar{a}\neq a|s)\mathbb{E}_{(\bar{a},a)\sim(\bar{\pi},\pi)}\left[A^{\pi}(s,\bar{a}) - A^{\pi}(s,a)\right]\right| \qquad \text{(a)} \\
&\leq \alpha\mathbb{E}_{(\bar{a},a)\sim(\bar{\pi},\pi)}\left[|A^{\pi}(s,\bar{a}) - A^{\pi}(s,a)|\right] \\
&\leq 2\epsilon\alpha \ ,
\end{aligned}
$$

其中 (a) 处注意到可以分为 $\bar{a}=a$ 和 $\bar{a}\neq a$ 两项，$\bar{a}=a$ 的项为 0。 □

更进一步地，将状态访问度量考虑在内，我们得到如下引理：

引理 7.5. 给定任意策略 π 和 $\bar{\pi}$，如果 $(\pi,\bar{\pi})$ 是 α-耦合策略对，则

$$
\left|\mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right] - \mathbb{E}_{(s_t,a_t)\sim(P^{\pi},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right]\right| \leq 4\epsilon\alpha\left(1-(1-\alpha)^t\right) \ ,
$$

其中 $\alpha = D_{TV}^{max}(\bar{\pi}\|\pi)$，$\epsilon = \max_{s,a}|A^{\pi}(s,a)|$。

证明. 记时刻 t 之前（包括时刻 t）$\bar{\pi}$ 和 π 产生不一致的动作的次数为 n_t。

$$
\begin{aligned}
&\left|\mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right] - \mathbb{E}_{(s_t,a_t)\sim(P^{\pi},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right]\right| \\
=&P(n_t>0)\cdot\left|\mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})|n_t>0}\left[A^{\pi}(s_t,a_t)\right] - \mathbb{E}_{(s_t,a_t)\sim(P^{\pi},\bar{\pi})|n_t>0}\left[A^{\pi}(s_t,a_t)\right]\right| \qquad \text{(a)} \\
=&(1-P(n_t=0))\cdot E \qquad \text{(b)} \\
\leq&\left(1-\prod_{k=1}^{t}P\left(\bar{a}_k=a_k|\bar{a}_k\sim\bar{\pi}(\cdot|s_k),a_k\sim\pi(\cdot|s_k)\right)\right)\cdot E \\
\leq&\left(1-\prod_{k=1}^{t}\left(1-P\left(\bar{a}_k\neq a_k|\bar{a}_k\sim\bar{\pi}(\cdot|s_k),a_k\sim\pi(\cdot|s_k)\right)\right)\right)\cdot E \\
\leq&\left(1-(1-\alpha)^t\right)\cdot E \qquad \text{(c)} \\
\leq&4\epsilon\alpha\left(1-(1-\alpha)^t\right) \ , \qquad \text{(d)}
\end{aligned}
$$

其中

(a) 产生动作不一致次数为 0 时 $\left|\mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right] - \mathbb{E}_{(s_t,a_t)\sim(P^{\pi},\bar{\pi})}\left[A^{\pi}(s_t,a_t)\right]\right|$ 产生的差异为 0；

(b) 令 $E = \left|\mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})|n_t>0}\left[A^{\pi}(s_t,a_t)\right] - \mathbb{E}_{(s_t,a_t)\sim(P^{\pi},\bar{\pi})|n_t>0}\left[A^{\pi}(s_t,a_t)\right]\right|$；

(c) 由耦合策略对的性质可知，$\forall k, P\left(\bar{a}_k\neq a_k|\bar{a}_k\sim\bar{\pi}(\cdot|s_k),a_k\sim\pi(\cdot|s_k)\right)\leq\alpha$；

(d) 分析 E，有

$$
\begin{aligned}
&\left|\mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})|n_t>0}\left[A^{\pi}(s_t,a_t)\right] - \mathbb{E}_{(s_t,a_t)\sim(P^{\pi},\bar{\pi})|n_t>0}\left[A^{\pi}(s_t,a_t)\right]\right| \\
\leq&\left|\mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})|n_t>0}\left[A^{\pi}(s_t,a_t)\right]\right| + \left|\mathbb{E}_{(s_t,a_t)\sim(P^{\pi},\bar{\pi})|n_t>0}\left[A^{\pi}(s_t,a_t)\right]\right| \\
\leq&2\cdot 2\epsilon\alpha = 4\epsilon\alpha \ 。
\end{aligned}
$$

□

7.3 单调性提升

在本小节中我们将证明 $\mathcal{J}(\pi)+\mathcal{L}_\pi(\bar{\pi})$ 是 $\mathcal{J}(\bar{\pi})$ 的一个下界，进而证明 TRPO 的单调性提升性质。

注 7.2. 如果我们证明了 $\mathcal{J}(\pi)+\mathcal{L}_\pi(\bar{\pi})$ 能够构成 $\mathcal{J}(\bar{\pi})$ 的一个下界，不妨表示为 $\mathcal{J}(\pi)+\mathcal{L}_\pi(\bar{\pi})-\beta$，则

$$\begin{aligned} \mathcal{J}(\bar{\pi}) &\geq \mathcal{J}(\pi)+\mathcal{L}_\pi(\bar{\pi})-\beta \\ \Rightarrow\quad \mathcal{J}(\bar{\pi})-\mathcal{J}(\pi) &\geq \mathcal{L}_\pi(\bar{\pi})-\beta \ 。 \end{aligned}$$

那么，如果我们优化 $\mathcal{L}_\pi(\bar{\pi})$ 时能在维持 β 较小的同时增大 $\mathcal{L}_\pi(\bar{\pi})$ 使得 $\mathcal{L}_\pi(\bar{\pi})-\beta\geq 0$，那么 $\mathcal{J}(\bar{\pi})-\mathcal{J}(\pi)\geq\mathcal{L}_\pi(\bar{\pi})-\beta\geq 0$，即更新后的策略相较于旧策略具有更好的表现，可以得到一个在表现上单调性递增的策略序列。

定理 7.1. 给定任意策略 π 和 $\bar{\pi}$，如果 $(\pi,\bar{\pi})$ 是 α-耦合策略对，则

$$|\mathcal{J}(\bar{\pi})-\mathcal{J}(\pi)-\mathcal{L}_\pi(\bar{\pi})|\leq\frac{4\gamma\alpha^2\epsilon}{(1-\gamma)^2}\ ,$$

其中 $\mathcal{L}_\pi(\bar{\pi})=\mathbb{E}_{(s,a)\sim(\rho^\pi,\bar{\pi})}\left[A^\pi(s_t,a_t)\right]$，$\alpha=D_{TV}^{max}\left(\bar{\pi}\|\pi\right)$，$\epsilon=\max_{s,a}|A^\pi(s,a)|$。

证明.

$$|\mathcal{J}(\bar{\pi})-\mathcal{J}(\pi)-\mathcal{L}_\pi(\bar{\pi})|=\left|\mathbb{E}_{(s,a)\sim(\rho^{\bar{\pi}},\bar{\pi})}\left[A^\pi(s,a)\right]-\mathbb{E}_{(s,a)\sim(\rho^\pi,\bar{\pi})}\left[A^\pi(s,a)\right]\right|$$

对于状态访问度量，我们证明下面的定理来将其展开。

引理 7.6. 对于任意策略 π^1 和 π^2，以及函数 $f:\mathcal{S}\times\mathcal{A}\mapsto\mathbb{R}$，下面的等式成立：

$$\mathbb{E}_{(s,a)\sim(\rho^{\pi^1},\pi^2)}\left[f(s,a)\right]=\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{(s_t,a_t)\sim(P^{\pi^1},\pi^2)}\left[f(s_t,a_t)\right]$$

证明.

$$\begin{aligned} \mathbb{E}_{(s,a)\sim(\rho^{\pi^1},\pi^2)}\left[f(s,a)\right] &= \sum_{s\in\mathcal{S}}\sum_{t=0}^{\infty}\gamma^t P(s_t=s|\pi^1)\sum_{a\in\mathcal{A}}\pi^2(a|s)f(s,a) \\ &= \sum_{t=0}^{\infty}\gamma^t\sum_{s\in\mathcal{S}}P(s_t=s|\pi^1)\sum_{a\in\mathcal{A}}\pi^2(a|s)f(s,a) \\ &= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{(s_t,a_t)\sim(P^{\pi^1},\pi^2)}\left[f(s_t,a_t)\right] \end{aligned}$$

□

基于此，有

$$\begin{aligned} |\mathcal{J}(\bar{\pi})-\mathcal{J}(\pi)-\mathcal{L}_\pi(\bar{\pi})| &= \left|\mathbb{E}_{(s,a)\sim(\rho^{\bar{\pi}},\bar{\pi})}\left[A^\pi(s,a)\right]-\mathbb{E}_{(s,a)\sim(\rho^\pi,\bar{\pi})}\left[A^\pi(s,a)\right]\right| \\ &= \left|\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})}\left[A^\pi(s_t,a_t)\right]-\sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{(s_t,a_t)\sim(P^\pi,\bar{\pi})}\left[A^\pi(s_t,a_t)\right]\right| \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{t=0}^{\infty} \gamma^t \left| \mathbb{E}_{(s_t,a_t)\sim(P^{\bar{\pi}},\bar{\pi})} \left[A^{\pi}(s_t,a_t) \right] - \mathbb{E}_{(s_t,a_t)\sim(P^{\pi},\bar{\pi})} \left[A^{\pi}(s_t,a_t) \right] \right| \\
&\leq \sum_{t=0}^{\infty} \gamma^t \cdot 4\epsilon\alpha \left(1 - (1-\alpha)^t \right) \qquad \text{(a)} \\
&= 4\epsilon\alpha \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)} \right) \\
&= 4\epsilon\alpha \frac{\gamma\alpha}{(1-\gamma)(1-\gamma(1-\alpha))} \\
&\leq \frac{4\gamma\epsilon\alpha^2}{(1-\gamma)^2}
\end{aligned}
$$

其中 (a) 处使用了引理 7.5。 □

结合注 7.2和定理 7.1，我们知道 $\beta = \frac{4\gamma\epsilon\alpha^2}{(1-\gamma)^2}$，分析 β 中的各组成成分，一个实现单调性策略优化的可行方式通过维持 α 较小同时优化 $\mathcal{L}_{\pi}(\bar{\pi})$，即在限制更新前后的策略差异较小的同时进行策略优化，这也就是 TRPO 算法的目的。