# 6 Policy Gradient-Based Reinforcement Learning Method

## 6.1 基于直接策略搜索的强化学习方法

广义值函数的方法包括策略评估和策略改善两个步骤。当值函数最优时，策略是最优的。此时的最优策略是贪婪策略。贪婪策略是指 $\arg\max_a Q_\theta(s,a)$，即在状态为 $s$ 时，对应最大行为值函数的动作，它是一个状态空间向动作空间的迎神，该映射就是最优策略。利用这种方法得到的策略往往是状态空间向有限集动作空间的映射。

策略搜索是将策略参数化，即 $\pi_\theta(s)$：利用参数化的线性函数或者非线性函数（如神经网络）表示策略，寻找最优的参数，使强化学习的目标——累计回报的期望 $E[\sum_{t=0}^{H} R(s_t)|\pi_\theta]$ 最大。

在值函数的方法中，我们迭代计算的是值函数，再根据值函数改善策略；而在策略搜索方法中，我们直接对策略进行迭代计算，也就是迭代更新策略的参数值，直到累积回报的期望最大，此时的参数所对应的策略为最优策略。

值函数方法和直接策略搜索方法的优缺点对比总结如下。

（1）直接策略搜索方法是对策略 $\pi$ 进行参数化表示，与值函数方法中对值函数进行参数化表示相比，策略参数化更简单，有更好的收敛性。

（2）利用值函数方法求解最优策略时，策略改善需要求解 $\arg\max_a Q_\theta(s,a)$，当要解决的问题动作空间很大或者动作为连续集时，该式无法有效求解。

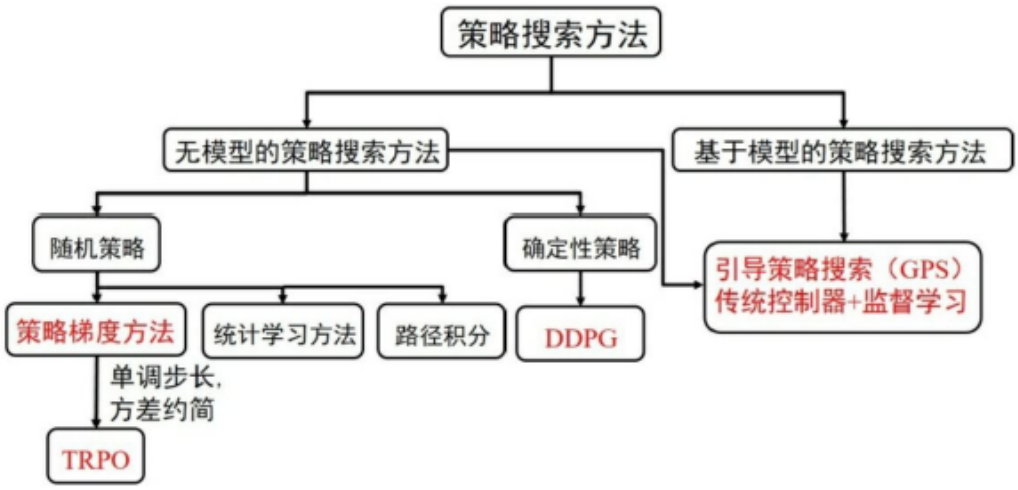
（3）直接策略搜索方法经常采用随机策略，因为随机策略可以将探索直接集成到所学习的策略之中。

与值函数方法相比，策略搜索方法也普遍存在一些缺点，比如：

（1）策略搜索的方法容易收敛到局部最小值。

（2）评估单个策略时并不充分，方差较大。

策略搜索方法分类如下图所示。




策略搜索方法按照是否利用模型可分为无模型的策略搜索方法和基于模型的策略搜索方法。其中无模型的策略搜索方法根据策略是采用随机策略还是确定性策略可分为随机策略搜索方法和确定性策略搜索方法。随机策略搜索方法最先发展起来的是策略梯度方法，但策略梯度方法存在学习速率难以确定的问题。为回避该问题，学者们又提出了基于统计学习的方法和基于路径积分的方法。但 TRPO 方法并没

有回避该问题，而是找到了替代损失函数——利用优化方法在每个局部点找到使损失函数单调非增的最优步长。

## 6.2 基于策略梯度的强化学习方法

下面分别从似然率的视角和重要性采样的视角推导策略梯度方法。

**从似然率的视角推导策略梯度。**

用 $\tau$ 表示一组状态-行为序列 $s_0, u_0, \cdots, s_H, u_H$ 。

符号 $R(\tau) = \sum_{t=0}^{H} R(s_t, u_t)$ 表示轨迹 $\tau$ 的回报， $P(\tau;\theta)$ 表示轨迹 $\tau$ 出现的概率；强化学习的目标函数可以表示为

$$U(\theta) = E(\sum_{t=0}^{H} R(s_t, u_t); \pi_\theta) = \sum_{\tau} P(\tau;\theta) R(\tau) \tag{1}$$

强化学习的目标是找到最优参数，使得

$$\max_{\theta} U(\theta) = \max_{\theta} \sum_{\tau} P(\tau;\theta) R(\tau) \tag{2}$$

这时，策略搜索方法实际上变成了一个优化问题。解决优化问题有很多方法，比如最速下降法、牛顿法、内点法等。

其中，最简单、也最常用的是最速下降法，此处称为策略梯度的方法，即 $\theta_{new} = \theta_{old} + \alpha \nabla_\theta U(\theta)$ ，问题的关键是如何计算策略梯度 $\nabla_\theta U(\theta)$ 。

我们对目标函数求导：

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_{\tau} P(\tau;\theta) R(\tau) \\
&= \sum_{\tau} \nabla_\theta P(\tau;\theta) R(\tau) \\
&= \sum_{\tau} \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta) R(\tau) \\
&= \sum_{\tau} P(\tau;\theta) \frac{\nabla_\theta P(\tau;\theta) R(\tau)}{P(\tau;\theta)} \\
&= \sum_{\tau} P(\tau;\theta) \nabla_\theta \log P(\tau;\theta) R(\tau)
\end{aligned} \tag{3}$$

最终策略梯度变成求 $\nabla_\theta \log P(\tau;\theta) R(\tau)$ 的期望，这可以利用经验平均估算。因此，当利用当前策略 $\pi_\theta$ 采样 m 条轨迹后，可以利用 m 条轨迹的经验平均逼近策略梯度：

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau;\theta) R(\tau) \tag{4}$$

**从重要性采样的角度推导策略梯度。**

目标函数为 $U(\theta) = E(\sum_{t=0}^{H} R(s_t, u_t); \pi_\theta) = \sum_{\tau} P(\tau;\theta) R(\tau)$ 。

利用参数 $\theta_{old}$ 产生的数据评估参数 $\theta$ 的回报期望，由重要性采样得

$$\begin{aligned}
U(\theta) &= \sum_{\tau} P(\tau|\theta_{old}) \frac{P(\tau;\theta)}{P(\tau;\theta_{old})} R(\tau) \\
&= E_{\tau \sim \theta_{old}}[\frac{P(\tau;\theta)}{P(\tau;\theta_{old})} R(\tau)]
\end{aligned} \tag{5}$$

导数为

$$\nabla_\theta U(\theta) = E_{\tau\sim\theta_{old}}[\frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta_{old})}R(\tau)] \tag{6}$$

令 $\theta = \theta_{old}$ ，得到当前策略的导数：

$$\begin{aligned}&\nabla_\theta U(\theta)|_{\theta=\theta_{old}}\\ &= E_{\tau\sim\theta_{old}}[\frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta_{old})}R(\tau)]\\ &= E_{\tau\sim\theta_{old}}[\nabla_\theta \log P(\tau|\theta)|_{\theta_{old}}R(\tau)]\end{aligned} \tag{7}$$

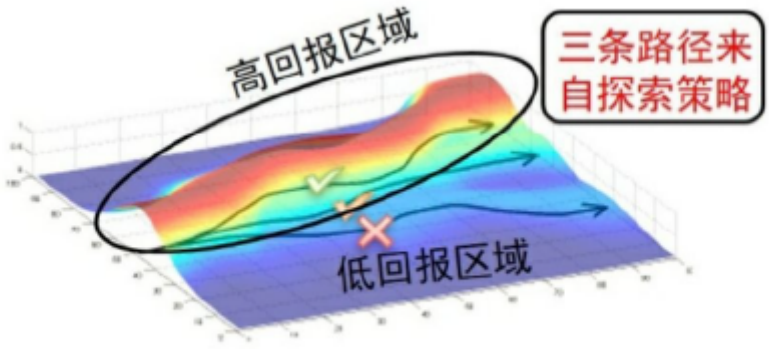
从重要性采样的视角推导策略梯度，不仅得出与似然率的视角相同的结果，更重要的是得到了原来目标函数新的损失函数： $U(\theta) = E_{\tau\sim\theta_{old}}[\frac{P(\tau|\theta)}{P(\tau|\theta_{old})}R(\tau)]$ 。

下面分别阐述公式中的 $\nabla_\theta \log P(\tau;\theta)$ 和 $R(\tau)$ 。

第一项 $\nabla_\theta \log P(\tau;\theta)$ 是轨迹 $\tau$ 的概率随参数 $\theta$ 变化最陡的方向。参数在该方向更新时，若沿着正方向，则该轨迹 $\tau$ 的概率会变大；若沿着负方向更新，则该轨迹 $\tau$ 的概率会变小。

第二项 $R(\tau)$ 控制了参数更新的方向和步长。 $R(\tau)$ 为正且越大则参数更新后该轨迹的概率越大； $R(\tau)$ 为负，则降低该轨迹的概率，抑制该轨迹的发生。

因此，从直观上理解策略梯度时，我们发现策略梯度会增加高回报路径的概率，减小低回报路径的概率。如下图所示，高回报的轨迹概率被增大，低回报区域的轨迹概率被减小。




前面推导出策略梯度的求解公式为

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\nabla_\theta \log P(\tau;\theta)R(\tau) \tag{8}$$

现在，解决似然率的梯度问题，即如何求解 $\nabla_\theta \log P(\tau;\theta)$ 。

已知 $\tau = s_0, u_0, \cdots, s_H, u_H$ ，则轨迹的似然率可写成

$$P(\tau^{(i)};\theta) = \prod_{t=0}^{H} P(s_{t+1}^{(i)}|s_t^{(i)}, u_t^{(i)}) \cdot \pi_\theta(u_t^{(i)}|s_t^{(i)}) \tag{9}$$

其中， $P(s_{t+1}^{(i)}|s_t^{(i)}, u_t^{(i)})$ 表示动力学，无参数，因此可在求导过程中消掉。具体推导如下。

$$\nabla_\theta \log P(\tau^{(i)};\theta) = \nabla_\theta \log\ [\prod_{t=0}^{H} P(s_{t+1}^{(i)}|s_t^{(i)}, u_t^{(i)}) \cdot \pi_\theta(u_t^{(i)}|s_t^{(i)})]$$

$$= \nabla_\theta[\sum^{H} \log P(s_{t+1}^{(i)}|s_t^{(i)}, u_t^{(i)}) + \sum^{H} \log \pi_\theta(u_t^{(i)}|s_t^{(i)})]$$

$$
\begin{aligned}
&= \nabla_\theta [\sum_{t=0}^{H} \log \pi_\theta(u_t^{(i)}|s_t^{(i)})] \\
&= \sum_{t=0}^{H} \nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)})
\end{aligned} \tag{10}
$$

从上述结果来看，似然率梯度转化为动作策略的梯度，与动力学无关，那么如何求解策略的梯度呢?

我们看一下常见的策略表示方法。

通常，随机策略可以写成确定性策略加随机部分，即

$$
\pi_\theta = \mu_\theta + \epsilon \tag{11}
$$

高斯策略 $\epsilon \sim N(0, \sigma^2)$ ，是均值为零，标准差为 $\sigma$ 的高斯分布。

和值函数逼近一样，确定性部分通常表示成以下方式。

线性策略： $\mu(s) = \phi(s)^T \theta$ 。

径向基策略： $\pi_\theta(s) = \omega^T \phi(s)$ ，其中， $\phi(s) = \exp\left(-\frac{1}{2}(s-\mu_i)\right)^T D_i(s-\mu_i))$ ；参数为 $\theta = \{\omega, \mu_i, d_i\}$ 。

我们以确定性部分策略是线性策略为例说明 $\log \pi_\theta(u_t^{(i)}|s_t^{(i)})$ 是如何计算的。

首先， $\pi(u|s) \sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(u-\phi(s)^T\theta)^2}{2\sigma^2}\right)$ ，利用该分布采样，得到 $u_t^{(i)}$ ，然后将 $(s_t^{(i)}, u_t^{(i)})$ 代入，得

$$
\nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)}) = \frac{(u_t^{(i)} - \phi(s_t^{(i)})^T\theta)\phi(s_t^{(i)})}{\sigma^2} \tag{12}
$$

其中，方差参数 $\sigma^2$ 用来控制策略的探索性。

由此，推导出策略梯度的计算公式:

$$
\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} (\sum_{t=0}^{H} \nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)}) R(\tau^{(i)})) \tag{13}
$$

上述式的策略梯度是无偏的，但方差很大，我们在回报中引入常数基线 b 减小方差。

当回报引入常数 b 时，策略梯度不变，即

$$
\begin{aligned}
\nabla_\theta U(\theta) \approx \hat{g} &= \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)}; \theta) R(\tau^{(i)}) \\
&= \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)}; \theta) R(\tau^{(i)} - b)
\end{aligned} \tag{14}
$$

我们求使得策略梯度的方差最小时的基线 b 。

令 $X = \nabla_\theta \log P(\tau^{(i)}; \theta)(R(\tau^{(i)} - b)$ ，则方差为

$$
Var(X) = E(X - \bar{X})^2 = EX^2 - E\bar{X}^2 \tag{15}
$$

方差最小处，方差对 b 的导数为零，即

$$
\frac{\partial Var(X)}{\partial b} = E(X \frac{\partial X}{\partial b}) = 0 \tag{16}
$$

其中，$\bar{X} = EX$ 与 b 无关。将 $X$ 代入，可得

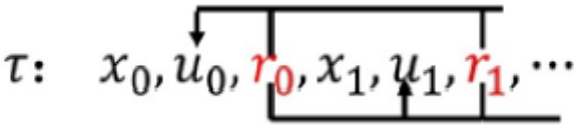
$$b = \frac{\sum_{i=1}^{m}[(\sum_{t=0}^{H} \nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)}))^2 R(\tau)]}{\sum_{i=1}^{m}[(\sum_{t=0}^{H} \nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)}))^2]} \tag{17}$$

除了增加基线的方法外，修改回报函数也可以进一步减小方差。

引入基线后，策略梯度公式变成

$$\nabla_\theta U(\theta) \approx \frac{1}{m}\sum_{i=1}^{m}(\sum_{t=0}^{H} \nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)})(R(\tau^{(i)}) - b)) \tag{18}$$

其中，b 取（17）式。在式（18）中，每个动作 $u_t^{(i)}$ 所对应的 $\nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)})$ 都乘以相同的该轨迹的总回报 $(R(\tau^{(i)}) - b)$，如下图所示。



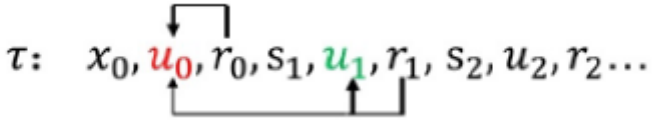

然而，当前的动作与过去的回报实际上是没有关系的，即

$$E_p[\partial \log \pi_\theta(u_t|x_t, t) r_j] = 0 \quad for\ j < t \tag{19}$$

因此，我们可以修改式（18）中的回报函数，有两种修改方法。

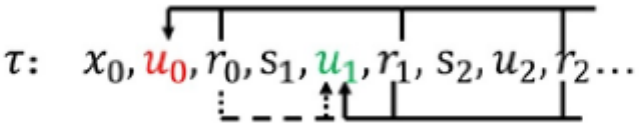
第一种方法称为 G(PO)MDP，如下图所示。

$$\nabla_\theta U(\theta) \approx \frac{1}{m}\sum_{i=1}^{m}\sum_{j=0}^{H-1}(\sum_{t=0}^{j} \nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)})(r_j - b_j)) \tag{20}$$




第二种方法称为策略梯度理论，如下图所示。

$$\nabla_\theta U(\theta) \approx \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(u_t^{(i)}|s_t^{(i)})(\sum_{k=t}^{H-1}(R(s_k^{(i)}) - b)) \tag{21}$$




为了使方差最小，可以利用前面的方法求解相应的基线 b 。