策略梯度定理：

定义

当轨迹中某个 '状态-动作对' (s,a) 是令人满意的，即 $Q(s,a)>0$，那就增加 $\pi_\theta(s,a)$ 的概率。

证明

策略梯度定理 有两种形式：

$$\nabla_\theta J(\theta)=E_{\substack{s_t\sim \Pr(s_0\to s_t,t,\pi)\\ a_t\sim\pi(a_t|s_t)}}\left[\sum_{t=0}^{\infty}\gamma^t Q_\pi(s_t,a_t)\nabla\ln\pi(a_t|s_t)\right] \tag{1}$$

$$\nabla_\theta J(\theta)\propto E_{\substack{s\sim D^\pi\\ a\sim\pi(a|s)}}\left[Q_{\pi_\theta}(s,a)\nabla_\theta\ln\pi_\theta(a|s)\right] \tag{2}$$

首先，定义强化学习的优化目标：$J(\theta)\doteq V_{\pi_\theta}(s_0)$，其中，$v_{\pi_\theta}$ 是 π_θ 的真实价值函数。目标函数求梯度有：

$$\begin{aligned}
\nabla_\theta J(\theta)=\nabla_\theta V_{\pi_\theta}(s_0)=\nabla&\left[\sum_{a_0}\pi(a_0|s_0)Q_\pi(s_0,a_0)\right]\\
&=\sum_{a_0}\left[\nabla\pi(a_0|s_0)Q_\pi(s_0,a_0)+\pi(a_0|s_0)\nabla Q_\pi(s_0,a_0)\right]\\
&=\sum_{a_0}\left[\nabla\pi(a_0|s_0)Q_\pi(s_0,a_0)+\pi(a_0|s_0)\nabla\sum_{s_1,r_1}p(s_1,r_1|s_0,a_0)(r_1+\gamma V(s_1))\right]\\
&=\sum_{a_0}\nabla\pi(a_0|s_0)Q_\pi(s_0,a_0)+\sum_{a_0}\pi(a_0|s_0)\sum_{s_1}p(s_1|s_0,a_0)\cdot\gamma\nabla V(s_1)\\
&=\sum_{a_0}\nabla\pi(a_0|s_0)Q_\pi(s_0,a_0)\\
&+\sum_{a_0}\pi(a_0|s_0)\sum_{s_1}p(s_1|s_0,a_0)\cdot\gamma\sum_{a_1}\left[\nabla\pi(a_1|s_1)Q_\pi(s_1,a_1)+\pi(a_1|s_1)\sum_{s_2}p(s_2|s_1,a_1)\gamma\nabla V(s_2)\right]\\
&=\sum_{a_0}\nabla\pi(a_0|s_0)Q_\pi(s_0,a_0)\\
&+\sum_{a_0}\pi(a_0|s_0)\sum_{s_1}p(s_1|s_0,a_0)\cdot\gamma\sum_{a_1}\nabla\pi(a_1|s_1)Q_\pi(s_1,a_1)+\ldots\\
&=\sum_{s_0}\Pr(s_0\to s_0,0,\pi)\sum_{a_0}\nabla\pi(a_0|s_0)\gamma^0Q_\pi(s_0,a_0)\\
&+\sum_{s_1}\Pr(s_0\to s_1,1,\pi)\sum_{a_1}\nabla\pi(a_1|s_1)\gamma^1Q_\pi(s_1,a_1)+\ldots\\
&=\sum_{s_0}\Pr(s_0\to s_0,0,\pi)\sum_{a_0}\nabla\pi(a_0|s_0)\left[\gamma^0Q_\pi(s_0,a_0)\nabla\ln\pi(a_0|s_0)\right]\\
&+\sum_{s_1}\Pr(s_0\to s_1,1,\pi)\sum_{a_1}\nabla\pi(a_1|s_1)\left[\gamma^1Q_\pi(s_1,a_1)\nabla\ln\pi(a_1|s_1)\right]+\ldots\\
&=\sum_{t=0}^{\infty}\sum_{s_t}\Pr(s_0\to s_t,t,\pi)\sum_{a_y}\pi(a_t|s_t)\left[\gamma^tQ_\pi(s_t,a_t)\nabla\ln\pi(a_t|s_t)\right]
\end{aligned}$$

其中，$p(s_1,r_1|s_0,a_0)$ 表示环境转移概率， $\Pr(s_0\to s_t,t,\pi)$ 表示从状态 s_0 出发，在策略 π 的作用下，经过 t 步到达状态 s_t 的概率。如：

$$
\begin{aligned}
\Pr(s_0 \to s_0, 0, \pi) &= 1 \\
\Pr(s_0 \to s_1, 1, \pi) &= \sum_{a_0} \pi(a_0|s_0) p(s_1|s_0, a_0) \\
&\cdots
\end{aligned}
$$

此外，还使用了 $\nabla \pi(a|s) = \pi(a|s) \nabla \ln \pi(a|s)$ 的恒等变换。第 4 行到第 5 行对 $\nabla V(s_1)$ 进行了第 1 行到第 4 行对 $\nabla V(s_0)$ 的展开。基于此，我们得到 策略梯度定理 的基本形式：

$$
\nabla_\theta J(\theta) = \sum_{t=0}^{\infty} \sum_{s_t} \Pr(s_0 \to s_t, t, \pi) \sum_{a_t} \pi(a_t|s_t) \left[\gamma^t Q_\pi(s_t, a_t) \nabla \ln \pi(a_t|s_t)\right]
$$

形式 1 的证明：

显然有 $\sum_{a_t} \pi(a_t|s_t) = 1$。此外，还有 $\sum_{s_t} \Pr(s_0 \to s_t, t, \pi) = 1$，由于概率求和等于 1，基本形式可以写成期望的形式：

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{t=0}^{\infty} \sum_{s_t} \Pr(s_0 \to s_t, t, \pi) \sum_{a_t} \pi(a_t|s_t) \left[\gamma^t Q_\pi(s_t, a_t) \nabla \ln \pi(a_t|s_t)\right] \\
&= \sum_{t=0}^{\infty} E_{\substack{s_t \sim \Pr(s_0 \to s_t, t, \pi) \\ a_t \sim \pi(a_t|s_t)}} \left[\gamma^t Q_\pi(s_t, a_t) \nabla \ln \pi(a_t|s_T)\right] \\
&= E_{\substack{s_t \sim \Pr(s_0 \to s_t, t, \pi) \\ a_t \sim \pi(a_t|s_t)}} \left[\sum_{t=0}^{\infty} \gamma^t Q_\pi(s_t, a_t) \nabla \ln \pi(a_t|s_t)\right]
\end{aligned}
$$

形式 2 的证明：

从基本形式出发，先对 t 个时刻求和，再写成期望的形式：

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{t=0}^{\infty} \sum_{s_t} \Pr(s_0 \to s_t, t, \pi) \sum_{a_t} \pi(a_t|s_t) \left[\gamma^t Q_\pi(s_t, a_t) \nabla \ln \pi(a_t|s_t)\right] \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s_t} \Pr(s_0 \to s_t, t, \pi) \sum_{a_t} \pi(a_t|s_t) \left[Q_\pi(s_t, a_t) \nabla \ln \pi(a_t|s_t)\right] \\
&= \sum_{x \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \Pr(s_0 \to x, t, \pi) \sum_{a} \pi(a|x) \left[Q_\pi(a|x) \nabla \ln \pi(a|x)\right] \\
&= \sum_{x \in \mathcal{S}} d^\pi(x) \sum_{a} \pi(a|x) \left[Q_\pi(x, a) \nabla \ln \pi(a|x)\right]
\end{aligned}
$$

其中 $\mathcal{S}$ 是从 s_0 出发通过策略 π 能到达的所有状态的集合。$d^\pi(x) = \sum_{t=0}^{\infty} \gamma^t \Pr(s_0 \to x, t, \pi)$ 是 折扣状态分布。但它的求和不等于一。
$\sum_{x \in \mathcal{S}} d^\pi(x) = \sum_{x \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \Pr(s_0 \to x, t, \pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{s_t} \Pr(s_0 \to x, t, \pi) = \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}$ 因此，如果写成期望的形式，需要将 $d^\pi(x)$ 归一化为标准分布 $D^\pi(x)$ 即，

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{x\in\mathcal{S}} d^\pi(x) \sum_{a_t} \pi(a_t|s_t)\left[Q_\pi(x,a_t)\nabla\ln\pi(a_t|x)\right] \\
&= \frac{1}{1-\gamma}\sum_{x\in\mathcal{S}}(1-\gamma)d^\pi(x)\sum_a \pi(a|x)\left[Q_\pi(x,a)\nabla\ln\pi(a|x)\right] \\
&= \frac{1}{1-\gamma}\sum_{x\in\mathcal{S}} D^\pi(x)\sum_a \pi(a|x)\left[Q_\pi(x,a)\nabla\ln\pi(a|x)\right] \\
&= \frac{1}{1-\gamma}E_{\substack{x\sim D^\pi \\ a\sim\pi(a|x)}}\left[Q_\pi(x,a)\nabla\ln\pi(a|x)\right] \\
&\propto E_{s\sim D^\pi a\sim\pi(a|s)}\left[Q_\pi(s,a)\nabla\ln\pi(a|s)\right]
\end{aligned}
$$

可以发现，$\nabla_\theta J(\theta)$ 与 $E_{\substack{s\sim D^\pi \\ a\sim\pi(a|s)}}\left[Q_\pi(s,a)\nabla\ln\pi(s,a)\right]$ 并非严格相等而是正比关系。

讨论

形式 1 是先写成期望的形式，再对 t 个时刻求和；形式 2 是先对 t 个时刻求和，再写成期望的形式。两种形式的最大区别是**状态的概率分布不同**。在形式 1 中，状态 s_t 服从从状态 s_0 出发，在策略 π 的作用下经过 t 步能到达的所有状态的分布。每个时刻的状态都服从各自的分布。即，$s_t \sim \Pr(s_o \to s_t, t, \pi)$。在形式 2 中，当 $\gamma = 1$ 时，$D^\pi(s)$ 为**轨迹中状态 s 在每个时刻出现的平均概率**。

用形式 1 计算策略梯度：

$$
\begin{aligned}
\nabla_\theta J(\theta) &= E_{s_t,a_t\sim\tau_\pi}\left[\sum_{t=0}^{2}\gamma^t Q_\pi(s_t,a_t)\nabla\ln\pi(a_t|s_t)\right] \\
&= \sum_{t=0}^{2}\sum_{s_t}\Pr(s_0\to s_t,t,\pi)\sum_{a_t}\pi(a_t|s_t)Q_\pi(s_t,a_t)\nabla\ln\pi(a_t|s_t)
\end{aligned}
$$

用形式 2 计算策略梯度：($\gamma = 1$时，归一化系数为 T 而非 $\frac{1}{1-\gamma}$)

$$
\begin{aligned}
\nabla_\theta J(\theta) &= T\times E_{\substack{s\sim D^\pi \\ a\sim\pi(a|s)}}\left[Q_\pi(s,a)\nabla\ln\pi(a|s)\right] \\
&= T\cdot\sum_{s\in\mathcal{S}} D^\pi(s)\sum_a \pi(a|s)Q_\pi(s,a)\nabla\ln\pi(a|s)
\end{aligned}
$$