策略梯度公式 (14.54) 证明（2021 年 5 月 18 日版）

参考资料

> https://spinningup.qiwihui.com/zh_CN/latest/spinningup/extra_pg_proof1.html

正文

从 (14.53) 开始

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta}=\mathbb{E}_{\tau\sim p_\theta(\tau)}\left\{\left[\sum_{t=0}^{T-1}\frac{\partial}{\partial\theta}\log\pi_\theta(a_t|s_t)\right]\left[G(\tau_{0:t})+\gamma^t G(\tau_{t:T})\right]\right\}$$

由期望的可加性，将 $G(\tau_{0:t})+\gamma^t G(\tau_{t:T})$ 拆为两项得

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta}=\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1}\frac{\partial}{\partial\theta}G(\tau_{0:t})\log\pi_\theta(a_t|s_t)\right]+\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1}\frac{\partial}{\partial\theta}\gamma^t G(\tau_{t:T})\log\pi_\theta(a_t|s_t)\right]$$

考虑上式第一项，代入 $G(\tau_{0:t})=\sum_{u=0}^{t-1}\gamma^u r_{u+1}$ 得

$$A=\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1}\frac{\partial}{\partial\theta}G(\tau_{0:t})\log\pi_\theta(a_t|s_t)\right]=\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1}\frac{\partial}{\partial\theta}\sum_{u=0}^{t-1}\gamma^u r_{u+1}\log\pi_\theta(a_t|s_t)\right]$$

由期望的可加性，将求和号提到期望式之外得

$$A=\sum_{t=0}^{T-1}\sum_{u=0}^{t-1}\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\frac{\partial}{\partial\theta}\gamma^u r_{u+1}\log\pi_\theta(a_t|s_t)\right]$$

对于期望式来说，折扣率 γ^u 是常量，将其提到期望式之外得

$$A=\sum_{t=0}^{T-1}\sum_{u=0}^{t-1}\gamma^u\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\frac{\partial}{\partial\theta}r_{u+1}\log\pi_\theta(a_t|s_t)\right]$$

将 r_{u+1} 提出求导式得

$$A=\sum_{t=0}^{T-1}\sum_{u=0}^{t-1}\gamma^u\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[r_{u+1}\frac{\partial}{\partial\theta}\log\pi_\theta(a_t|s_t)\right]$$

令 $f(t,u)=r_{u+1}\frac{\partial}{\partial\theta}\log\pi_\theta(a_t|s_t)$，考察下式，设轨迹 τ 是连续型随机变量，由期望的定义得

$$B=\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[f(t,u)\right]=\int_\tau p_\theta(\tau)f(t,u)$$

约定在积分下标中写上一个或多个积分变量，表示对这些积分变量在其整个定义域中积分，为求方便，使用上述表示的积分省略积分式中最后的微分。

由于 $f(t,u)$ 是关于 s_t,a_t,s_u,a_u,s_{u+1} 的函数，因此无需考虑 τ 中所有的变量（即 $s_0,a_0,s_1,a_1,\cdots,s_{T-1},a_{T-1},s_T$），而只需考虑 s_t,a_t,s_u,a_u,s_{u+1} 即可。即若设 $p_\theta(s_t,a_t,s_u,a_u,s_{u+1})$ 是 $p_\theta(\tau)$ 关于 s_t,a_t,s_u,a_u,s_{u+1} 的边缘密度函数，则得

$$B=\int_{s_t,a_t,s_u,a_u,s_{u+1}}p_\theta(s_t,a_t,s_u,a_u,s_{u+1})f(t,u)$$

> 回顾一下概率论知识，即若想求某连续型随机变量 X 的函数 $f(X)$ 的期望，而只知 X 与另一连续型随机变量 Y 的联合分布 $g(x,y)$，首先由期望的定义可知
>
> $$\mathbb{E}[f(X)]=\int_x g_X(x)f(x)$$
>
> 其中 $g_X(x)$ 是 X 的边缘密度函数，可以由
>
> $$g_X(x)=\int_y g(x,y)$$
>
> 计算得出。
> 另一方面，尽管 $f(X)$ 与 Y 无关，但可以在形式上将其视作与 Y 相关，即设 $f(X,Y)=f(X)$，从而由期望的定义
>
> $$\mathbb{E}[f(X)]=\mathbb{E}[f(X,Y)]=\int_{x,y}g(x,y)f(x,y)$$
>
> 这就说明了在计算随机变量函数的期望时，只需要使用与函数相关的那几个变量的边缘密度函数，但使用联合密度函数也是可以的。

由条件概率的定义得

$$p_\theta(s_t,a_t,s_u,a_u,s_{u+1})=p_\theta(s_t,a_t|s_u,a_u,s_{u+1})p_\theta(s_u,a_u,s_{u+1})$$

从而

$$B = \int_{s_t, a_t, s_u, a_u, s_{u+1}} p_\theta(s_t, a_t | s_u, a_u, s_{u+1}) p_\theta(s_u, a_u, s_{u+1}) f(t, u)$$

上述积分形式是五重积分，将其分解为下式

$$B = \int_{s_u, a_u, s_{u+1}} \int_{s_t, a_t} p_\theta(s_t, a_t | s_u, a_u, s_{u+1}) p_\theta(s_u, a_u, s_{u+1}) f(t, u)$$

$p_\theta(s_u, a_u, s_{u+1})$ 与 s_t, a_t 无关，将它丢到积分号外得

$$B = \int_{s_u, a_u, s_{u+1}} p_\theta(s_u, a_u, s_{u+1}) \int_{s_t, a_t} p_\theta(s_t, a_t | s_u, a_u, s_{u+1}) f(t, u)$$

其实 $f(t,u) = r_{u+1} \frac{\partial}{\partial \theta} \log \pi_\theta(a_t | s_t)$ 中 r_{u+1} 也与 s_t, a_t 无关，因此

$$B = \int_{s_u, a_u, s_{u+1}} p_\theta(s_u, a_u, s_{u+1}) r_{u+1} \int_{s_t, a_t} p_\theta(s_t, a_t | s_u, a_u, s_{u+1}) \frac{\partial}{\partial \theta} \log \pi_\theta(a_t | s_t)$$

考察上式右侧积分

$$C = \int_{s_t, a_t} p_\theta(s_t, a_t | s_u, a_u, s_{u+1}) \frac{\partial}{\partial \theta} \log \pi_\theta(a_t | s_t)$$

由条件概率的定义，将 s_t 分离得

$$p_\theta(s_t, a_t | s_u, a_u, s_{u+1}) = p_\theta(a_t | s_t, s_u, a_u, s_{u+1}) p_\theta(s_t | s_u, a_u, s_{u+1})$$

由于轨迹 τ 是一个马尔可夫决策过程的轨迹，拥有马尔可夫性质，即在 t 时刻做出的动作只与 s_t 有关，而与 t 时刻之前的状态和动作无关，因此得

$$p_\theta(a_t | s_t, s_u, a_u, s_{u+1}) = p_\theta(a_t | s_t) = \pi_\theta(a_t | s_t)$$

这说明上式实际上可以化简为策略 π 在状态为 s_t 时给出动作 a_t 的概率，因此

$$C = \int_{s_t, a_t} p_\theta(s_t | s_u, a_u, s_{u+1}) \pi_\theta(a_t | s_t) \frac{\partial}{\partial \theta} \log \pi_\theta(a_t | s_t)$$

将两重积分分解为下式

$$C = \int_{s_t} \int_{a_t} p_\theta(s_t | s_u, a_u, s_{u+1}) \pi_\theta(a_t | s_t) \frac{\partial}{\partial \theta} \log \pi_\theta(a_t | s_t)$$

将无关项丢出积分式得

$$C = \int_{s_t} p_\theta(s_t | s_u, a_u, s_{u+1}) \int_{a_t} \pi_\theta(a_t | s_t) \frac{\partial}{\partial \theta} \log \pi_\theta(a_t | s_t)$$

以下推导被不正式地称作“期望梯度对数概率引理”。为了理解什么叫“期望梯度对数概率”，注意到上式右侧积分其实就是概率密度函数 $\pi_\theta(a_t|s_t)$ 的对数的梯度的期望，从右向左读就是“期望梯度对数概率”。

将积分中的偏导展开并化简得

$$C=\int_{s_t} p_\theta(s_t|s_u,a_u,s_{u+1})\int_{a_t}\frac{\partial}{\partial\theta}\pi_\theta(a_t|s_t)$$

积分与偏导次序交换得

$$C=\int_{s_t} p_\theta(s_t|s_u,a_u,s_{u+1})\frac{\partial}{\partial\theta}\int_{a_t}\pi_\theta(a_t|s_t)$$

对概率密度函数在整个定义域上积分的结果是 1，即

$$C=\int_{s_t} p_\theta(s_t|s_u,a_u,s_{u+1})\frac{\partial}{\partial\theta}1$$

对常数偏导为 0，即

$$C=\int_{s_t} p_\theta(s_t|s_u,a_u,s_{u+1})\cdot 0=0$$

将这一重要结论代入到原式中，即得

$$\frac{\partial\mathcal{J}(\theta)}{\partial\theta}=\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\gamma^t G(\tau_{t:T})\frac{\partial}{\partial\theta}\log\pi_\theta(a_t|s_t)\right]$$

当 $u\geqslant t$ 时，$p_\theta(a_t|s_t,s_u,a_u,s_{u+1})\neq p_\theta(a_t|s_t)$。从直观上理解，在 t 时刻之后的状态和动作均受到 a_t 的影响，再直观一点，假设执行某个动作 a_t 后不可能出现某个状态 s_{t+1}，那么若 s_{t+1} 出现了，就一定不可能执行动作 a_t。这样一来，对于 $\tau_{t:T}$ 部分就不能像上述一样化简了。