方策勾配

菱沼 徹

2020 年 3 月 9 日

1 REINFORCE [2]

軌道 $\tau = (s_0, a_0, s_1, a_1, \cdots, s_T)$ は，終端状態にたどり着くまで，方策 $a_t \sim \pi(a_t|s_t)$ に従って行動をサンプリングし，またダイナミクス $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$ に従って状態をサンプリングすることにより生成される．報酬 $r(s_t, a_t)$ を，各時間ステップにおいて受け取る．確率的定常方策 π で表し，$a_t \sim \pi(a_t|s_t)$ とする．π を特徴づけるパラメータを θ とする．

軌道 τ が得られる確率は，次のように書ける．

$$p(\tau) = p(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

$$\ln p(\tau) = \ln p(s_0) + \sum_{t=0}^{T-1} \left(\ln \pi(a_t|s_t) + \ln p(s_{t+1}|s_t, a_t) \right)$$

これを微分すると，

$$\nabla_\theta p(\tau) = p(\tau) \nabla_\theta \ln p(\tau) = p(\tau) \left(\sum_{t=0}^{T-1} \nabla_\theta \ln \pi(a_t|s_t) \right)$$

報酬和を $R(\tau)$ とすると，

$$\begin{aligned} E[R(\tau)] &= \sum_\tau p(\tau) R(\tau) \\ \nabla_\theta E[R(\tau)] &= \nabla_\theta \sum_\tau p(\tau) R(\tau) = \sum_\tau R(\tau) \nabla_\theta p(\tau) = \sum_\tau R(\tau) p(\tau) \left(\sum_{t=0}^{T-1} \nabla_\theta \ln \pi(a_t|s_t) \right) \\ &= E\left[R(\tau) \left(\sum_{t=0}^{T-1} \nabla_\theta \ln \pi(a_t|s_t) \right) \right] \end{aligned}$$

軌道の n 番目のサンプルを $\tau^{(n)} = (s_0^{(n)}, a_0^{(n)}, s_1^{(n)}, a_1^{(n)}, \cdots, s_T^{(n)})$ とすると，

$$\nabla_\theta E[R(\tau)] \approx \frac{1}{N} \sum_{n=1}^{N} \left[R(\tau^{(n)}) \left(\sum_{t=0}^{T-1} \nabla_\theta \ln \pi(a_t^{(n)}|s_t^{(n)}) \right) \right]$$

2　方策勾配 [1]

2.1　定義と算数

軌道 $\tau = (s_0, a_0, s_1, a_1, \cdots)$ は，終端状態にたどり着くまで，方策 $a_t \sim \pi(\cdot|s_t)$ に従って行動をサンプリングし，またダイナミクス $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$ に従って状態をサンプリングすることにより生成される．報酬 $r(s_t, a_t)$ を，各時間ステップにおいて受け取る．割引率を $\gamma \in (0, 1)$ とする．確率的定常方策 π で表し，$a_t \sim \pi(a_t|s_t)$ とする．

状態行動価値関数 Q_π，価値関数 V_π を次のように定義する．

$$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \cdots}\left[\sum_{\ell=0}^{\infty} \gamma^\ell r(s_{t+\ell}, a_{t+\ell})\right]$$
$$V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, \cdots}\left[\sum_{\ell=0}^{\infty} \gamma^\ell r(s_{t+\ell}, a_{t+\ell})\right]$$

次が成り立つ．

$$\begin{aligned}
Q_\pi(s_t, a_t) &= \mathbb{E}_{s_{t+1}, a_{t+1}, \cdots}\left[r(s_t, a_t) + \gamma \sum_{\ell=0}^{\infty} \gamma^\ell r(s_{t+\ell}, a_{t+\ell})\right] = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}, a_{t+1}, \cdots}\left[\sum_{\ell=0}^{\infty} \gamma^\ell r(s_{t+\ell}, a_{t+\ell})\right] \\
&= r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}\left[V_\pi(s_{t+1})\right] = r(s_t, a_t) + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) V_\pi(s_{t+1}) \\
V_\pi(s_t) &= \sum_{a_t} \pi(a_t|s_t) Q_\pi(s_t, a_t)
\end{aligned}$$

$d_\pi(s)$ を，方策 π の下での状態 s の訪問頻度とする．

$$d_\pi(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t p(s_t = s|\pi)$$

ここで，$p(s_t = s|\pi)$ は，方策 π の下で時刻 t において状態 s に存在する確率である．初期状態 s_0 の分布は方策 π に依存しないため実際には $p(s_0 = s|\pi) = p(s_0 = s)$ であるが，表記の単純化のために $p(s_t = s|\pi)$ と書く．

次が成り立つ．

$$\sum_s d_\pi(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \sum_s p(s_t = s|\pi) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t = 1$$

最適化の評価指標として，期待割引報酬 $\rho(\pi)$ を次のように定義する．

$$\rho(\pi) = \mathbb{E}_{s_0, a_0, \cdots}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] = \sum_s P(s_0 = s) V_\pi(s_0)$$

2.2 勾配の導出

方策を特徴づけるパラメータを θ とすると，

$$
\begin{aligned}
\frac{\partial V_\pi(s_t)}{\partial\theta} &= \frac{\partial}{\partial\theta}\sum_{a_t}\pi(a_t|s_t)Q_\pi(s_t,a_t) \\
&= \sum_{a_t}\left\{\left[\frac{\partial}{\partial\theta}\pi(a_t|s_t)\right]Q_\pi(s_t,a_t)+\pi(a_t|s_t)\left[\frac{\partial}{\partial\theta}Q_\pi(s_t,a_t)\right]\right\} \\
&= \sum_{a_t}\left\{\left[\frac{\partial}{\partial\theta}\pi(a_t|s_t)\right]Q_\pi(s_t,a_t)+\pi(a_t|s_t)\left[\frac{\partial}{\partial\theta}\left(r(s_t,a_t)+\gamma\sum_{s_{t+1}}P(s_{t+1}|s_t,a_t)V_\pi(s_{t+1})\right)\right]\right\} \\
&= \sum_{a_t}\left\{\left[\frac{\partial}{\partial\theta}\pi(a_t|s_t)\right]Q_\pi(s_t,a_t)+\pi(a_t|s_t)\gamma\sum_{s_{t+1}}P(s_{t+1}|s_t,a_t)\left[\frac{\partial}{\partial\theta}V_\pi(s_{t+1})\right]\right\}
\end{aligned}
$$

初期状態 s_0 の分布は方策 π に依存しないため実際には $p(s_0=s|\pi)=p(s_0=s)$ であり，従って $\frac{\partial p(s_0=s|\pi)}{\partial\theta}=\frac{\partial p(s_0=s)}{\partial\theta}=0$ であることに注意すると，

$$
\begin{aligned}
&\frac{\partial\rho(\pi)}{\partial\theta} = \frac{\partial}{\partial\theta}\left[\sum_s p(s_0=s|\pi)V_\pi(s)\right] = \sum_s p(s_0=s|\pi)\frac{\partial V_\pi(s)}{\partial\theta} \\
&= \sum_s\sum_a\left\{p(s_0=s|\pi)Q_\pi(s,a)\left[\frac{\partial}{\partial\theta}\pi(a|s)\right]+p(s_0=s|\pi)\pi(a|s)\gamma\sum_{s'}P(s'|s,a)\left[\frac{\partial}{\partial\theta}V_\pi(s')\right]\right\} \\
&= \sum_s\sum_a\left\{p(s_0=s|\pi)Q_\pi(s,a)\left[\frac{\partial}{\partial\theta}\pi(a|s)\right]+\gamma\sum_{s'}p(s_1=s'|\pi)\left[\frac{\partial}{\partial\theta}V_\pi(s')\right]\right\} \\
&= \sum_s\sum_a\left\{p(s_0=s|\pi)Q_\pi(s,a)\left[\frac{\partial}{\partial\theta}\pi(a|s)\right]\right\}+\gamma\sum_{s'}p(s_1=s'|\pi)\left[\frac{\partial}{\partial\theta}V_\pi(s')\right] \\
&= \sum_s\sum_a\left\{p(s_0=s|\pi)Q_\pi(s,a)\left[\frac{\partial}{\partial\theta}\pi(a|s)\right]+\gamma\sum_{s'}p(s_1=s'|\pi)\left[\frac{\partial}{\partial\theta}\pi(a|s)\right]\right\}+\gamma^2\sum_{s''}p(s_2=s''|\pi)\left[\frac{\partial}{\partial\theta}V_\pi(s'')\right] \\
&= \sum_s\sum_a\left\{\sum_t\gamma^t p(s_t=s|\pi)Q_\pi(s,a)\left[\frac{\partial}{\partial\theta}\pi(a|s)\right]\right\} = \sum_s\sum_a\left\{\left[\sum_t\gamma^t p(s_t=s|\pi)\right]Q_\pi(s,a)\left[\frac{\partial}{\partial\theta}\pi(a|s)\right]\right\} \\
&= \frac{1}{1-\gamma}\sum_s\sum_a\left\{d_\pi(s)Q_\pi(s,a)\left[\frac{\partial}{\partial\theta}\pi(a|s)\right]\right\} = \frac{1}{1-\gamma}\sum_s\sum_a\left\{d_\pi(s)Q_\pi(s,a)\left[\pi(a|s)\frac{\partial}{\partial\theta}\ln\pi(a|s)\right]\right\} \\
&= \frac{1}{1-\gamma}\sum_s\sum_a\left\{d_\pi(s)\pi(a|s)Q_\pi(s,a)\left[\frac{\partial}{\partial\theta}\ln\pi(a|s)\right]\right\}
\end{aligned}
$$

ここで，$d_\pi(s)\pi(a|s)$ は π の下で (s,a) を訪問する頻度である事に注意すれば，$(s^{(n)},a^{(n)})$ を n 番目のサンプルとして，次のようにサンプル近似できる．

$$
\frac{\partial\rho(\pi)}{\partial\theta} \approx \frac{1}{1-\gamma}\sum_n\left\{Q_\pi(s^{(n)},a^{(n)})\left[\frac{\partial}{\partial\theta}\ln\pi(a^{(n)}|s^{(n)})\right]\right\}
$$

2.3 baseline 関数の利用

次が成り立つ．

$$
\sum_a \pi(a|s) = 1
$$

$$
\frac{\partial}{\partial\theta}\sum_a \pi(a|s) = \sum_a \pi(a|s)\frac{\partial}{\partial\theta}\pi(a|s) = \sum_a \pi(a|s)\pi(a|s)\frac{\partial}{\partial\theta}\ln\pi(a|s) = \frac{\partial}{\partial\theta}1 = 0
$$

従って，任意の関数 $b(s)$ に対して次が成り立つ．

$$
\frac{1}{1-\gamma}\sum_s\sum_a\left\{d_\pi(s)\pi(a|s)b(s)\left[\frac{\partial}{\partial\theta}\ln\pi(a|s)\right]\right\} = \frac{1}{1-\gamma}\sum_s d_\pi(s)b(s)\sum_a\left\{\pi(a|s)\left[\frac{\partial}{\partial\theta}\ln\pi(a|s)\right]\right\} = 0
$$

従って，方策を勾配を次のように書くことができる．

$$
\frac{\partial\rho(\pi)}{\partial\theta} = \frac{1}{1-\gamma}\sum_s\sum_a\left\{d_\pi(s)\pi(a|s)\left(Q_\pi(s,a)-b(s)\right)\left[\frac{\partial}{\partial\theta}\ln\pi(a|s)\right]\right\}
$$

ここで，$b(s)$ は baseline 関数と呼ばれる．特に，$b(s) = V_\pi(s)$ としてた場合の $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$ はアドバンテージ関数と呼ばれる．

参考文献

[1] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[2] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.