Policy Gradient Reinforcement Learning

Jeroen van Riel

April 2025

1 Stationary distribution for finite episodes

For some fixed policy π and initial state distribution h, we consider the underlying *induced Markov chain* over states. Because we are working with finite episodes, the induced state process is a Markov chain with absorbing states. We want to analyze how often states are visited on average, over multiple episodes. To see what *on average* means here, imagine that we link together separate episodes to create a regular Markov chain without absorbing states, in the following way: from each final state, we introduce state transitions to the initial states according to distribution h, see also Figure 1. Furthermore, we will write $S_t^{(i)}$ to denote the state at step t of episode i.

Consider an absorbing Markov chain with transition matrix

$$P_{xy} = \sum_a \pi(a|x)p(y|x,a).$$

There are t transient states and r absorbing states, so P can be written as

$$P = \begin{pmatrix} Q & R \\ \mathbf{0} & I_r \end{pmatrix},$$

where Q is a t-by-t matrix, R is a nonzero t-by-r matrix, I_r is the r-by-r identify matrix and $\mathbf{0}$ is the zero matrix. Observe that $(Q^k)_{xs}$ is the probability of reaching state s in k steps without being absorbed, starting from state x. Hence, the expected number of visits to state s without being absorbed, starting from state x, is given by

$$\eta(s|x) := \sum_{k=0}^{\infty} (Q^k)_{xs}.$$

Writing this in matrix form $N_{xs} = \eta(s|x)$, we can use the following property of this so-called Neumann series, to obtain

$$N = \sum_{k=0}^{\infty} Q^k = (I_t - Q)^{-1}.$$

Now we can derive two equivalent equations

$$N = (I_t - Q)^{-1} \iff \begin{cases} N(I_t - Q) = I_t \iff N = I_t + NQ, & \text{or} \\ (I_t - Q)N = I_t \iff N = I_t + QN. \end{cases}$$

Expanding the first equation in terms of matrix entries $N_{xs} = \eta(s|x)$ gives

$$\begin{aligned} \eta(s|x) &= \mathbb{1}\{x = s\} + \sum_y \eta(y|x) Q_{ys} \\ &= \mathbb{1}\{x = s\} + \sum_y \eta(y|x) \sum_a \pi(a|y)p(y|x,a) \end{aligned}$$

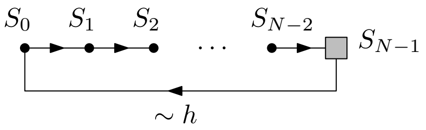

Figure 1: Illustration of the induced Markov chain when dealing with finite episodes. The next state after the final state, indicated as the grey rectangle, is sampled according to initial state distribution h.

and similarly, the second equation gives

$$\begin{aligned}\eta(s|x) &= \mathbb{1}\{x = s\} + \sum_y Q_{xy}\eta(s|y) \\ &= \mathbb{1}\{x = s\} + \sum_a \pi(a|x) \sum_y p(y|x,a)\eta(s|y)\end{aligned}$$

Now since the initial state is chosen according to distribution h, the expected number of visits $\eta(s)$ to state s in some episode is given by

$$\eta(s) = \sum_x h(x)\eta(s|x),$$

or written in matrix form $\eta = hN$, where η and h are row vectors. Therefore, we can also work with the equations

$$\begin{cases} hN = h + hNQ, & \text{or} \\ hN = h + hQN, \end{cases}$$

which are generally called *balance equations.* By writing the first variant as $\eta = h + \eta Q$ and expanding the matrix multiplication, we obtain

$$\eta(s) = h(s) + \sum_y \eta(y) \sum_a \pi(a|y)p(s|y,a).$$

Through appropriate normalization of the expected number of visits, we obtain the average fraction of time spent in state s, given by

$$\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}.$$

1.1 Monte Carlo sampling

Suppose we have some function $f : \mathcal{S} \to \mathbb{R}$ over states and we are interested in estimating $\mathbb{E}_{S_t^{(i)} \sim \mu}[f(S_t^{(i)})]$. We can just take random samples of $S_t^{(i)}$, by sampling initial state $S_0^{(i)} \sim h$ and then *rolling out* π to obtain $\tau^{(i)} = (S_0^{(i)}, A_0^{(i)}, R_1^{(i)}, S_1^{(i)}, A_1^{(i)}, R_2^{(i)}, S_2^{(i)}, \ldots, S_{N^{(i)}-1}^{(i)}) \sim \pi(\tau^{(i)}|S_0^{(i)})$, where $N^{(i)}$ denotes the total number of states visited in this episode. Given M such episode samples, we compute the estimate as

$$\mathbb{E}_{S_t^{(i)} \sim \mu}[f(S_t^{(i)})] \approx \left(\sum_{i=1}^{M} \sum_{t=0}^{N^{(i)}-1} f(S_t^{(i)}) \right) / \left(\sum_{i=1}^{M} N^{(i)} \right).$$

Observe that the analysis of the induced Markov chain can be extended to explicitly include actions and rewards as part of the state and derive the stationary distribution of the resulting Markov chain. However, we do not need this distribution explicitly in practice, because we can again use episode samples $\tau^{(i)}$. To keep notation concise, we will from now on denote this type of expectation as $\mathbb{E}_{\tau\sim h,\pi}[f(\tau)]$ and omit episode superscripts. Using this new notation, note that the average episode length is given by

$$\mathbb{E}_{h,\pi}[N] = \sum_{s'} \eta(s').$$

2 Policy gradient estimation

Let $v_{\pi_\theta} = \mathbb{E}_{h,\pi_\theta}[G_0]$ denote the expected episodic reward under policy π, where G_t is called the reward-to-go at step t, which is defined as

$$G_t := \sum_{k=t+1}^{\infty} R_k.$$

The main idea of policy gradient methods is to update the policy parameters θ in the direction that increases the expected episodic reward the most. This means that the policy parameters are updated as

$$\theta_{k+1} = \theta_k + \alpha \nabla v_{\pi_\theta},$$

where α is the learning rate and the gradient is with respect to θ. Instead of trying to derive or compute the gradient exactly, we often use some statistical estimate based on sampled episode. The basic policy gradient algorithm is to repeat the following three steps:

```
repeat:
1.  sample $M$ episodes $\tau^{(1)},\dots,\tau^{(M)}$ following $\pi_\theta$
2.  compute gradient estimate $\widehat{\nabla v_{\pi_\theta}}(\tau^{(1)},\dots,\tau^{(M)})$
3.  update $\theta \leftarrow \theta + \alpha\widehat{\nabla v_{\pi_\theta}}$
```

2.1 Policy gradient theorem and REINFORCE estimator

We will now present the fundamental policy gradient theorem, which essentially provides a function f such that

$$\nabla v_{\pi_\theta} = \mathbb{E}_{\tau\sim h,\pi_\theta}[f(\tau)],$$

which allows us to estimate the policy gradient using episode samples. To align with the notation of [1], we write $\Pr(x \to s, k, \pi) := (Q^k)_{xs}$, for the probability of reaching state s in k steps under policy π, starting from state some x, so that the expected number of visits can also be written as

$$\eta(s) = \sum_x h(x) \sum_{k=0}^{\infty} \Pr(x \to s, k, \pi)$$

As proven in the chapter on policy gradient methods in [1], the gradient of the value function for a fixed initial state s_0 with respect to the parameters is given by

$$\nabla v_\pi(s_0) = \sum_s \sum_{k=0}^{\infty} \Pr(s_0 \to s, k, \pi) \sum_a q_\pi(s,a) \nabla \pi(a|s). \tag{1}$$

When choosing the initial state s_0 according to some distribution $h(s_0)$, we verify that the final result is still the same as in [1]:

$$
\begin{aligned}
\nabla v_\pi :=& \nabla \mathbb{E}_{s_0 \sim h}[v_\pi(s_0)] && \text{(2a)}\\
=& \sum_{s_0} h(s_0) \sum_s \sum_{k=0}^{\infty} \Pr(s_0 \to s, k, \pi) \sum_a q_\pi(s,a) \nabla \pi(a|s) && \text{(2b)}\\
=& \sum_s \eta(s) \sum_a q_\pi(s,a) \nabla \pi(a|s) && \text{(2c)}\\
=& \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s) && \text{(2d)}\\
\propto& \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s), && \text{(2e)}
\end{aligned}
$$

where the constant of proportionality is just the average episode length. Because we do not know μ or q_π explicitly, we would like to estimate ∇v_π based on samples. If we sample episodes according to h and π as explained above, we encounter states according to μ, so we have

$$
\begin{aligned}
\nabla v_\pi &\propto \mathbb{E}_{h,\pi}\left[\sum_a q_\pi(S_t,a)\nabla\pi(a|S_t)\right] && \text{(3a)}\\
&= \mathbb{E}_{h,\pi}\left[\sum_a \pi(a|S_t) q_\pi(S_t,a) \frac{\nabla\pi(a|S_t)}{\pi(a|S_t)}\right] && \text{(3b)}\\
&= \mathbb{E}_{h,\pi}\left[q_\pi(S_t,A_t) \frac{\nabla\pi(A_t|S_t)}{\pi(A_t|S_t)}\right] && \text{(3c)}\\
&= \mathbb{E}_{h,\pi}\left[G_t \nabla \log \pi(A_t|S_t)\right]. && \text{(3d)}
\end{aligned}
$$

2.2 Baseline

Let $b(s)$ be some function of the state s only, then we have for any $s \in \mathcal{S}$

$$
\sum_a b(s)\nabla\pi(a|s) = b(s)\nabla\sum_a \pi(a|s) = b(s)\nabla 1 = 0. \tag{4}
$$

This yields the so-called REINFORCE estimate with *baseline*

$$
\begin{aligned}
\nabla v_\pi &\propto \sum_s \mu(s) \sum_a (q_\pi(s,a) + b(s))\nabla\pi(a|s) && \text{(5a)}\\
&= \mathbb{E}_{h,\pi}\left[\big(q_\pi(S_t,A_t) + b(S_t)\big)\nabla\log\pi(A_t|S_t)\right] && \text{(5b)}\\
&= \mathbb{E}_{h,\pi}\left[\big(G_t + b(S_t)\big)\nabla\log\pi(A_t|S_t)\right]. && \text{(5c)}
\end{aligned}
$$

Although estimates (3d) and (5c) are both equivalent in terms of their expected value, they may differ in higher moments, which is why an appropriate choice of b can make a lot of difference in how well the policy gradient algorithm converges to an optimal policy. As a specific baseline, consider the expected cumulative sum of rewards up to step the current step t, defined as

$$
b(s) = \mathbb{E}_{h,\pi}\left[\sum_{k=1}^{t} R_k \middle| S_t = s\right], \tag{6}
$$

then observe that

$$q_\pi(s,a) + b(s) = \mathbb{E}_{h,\pi}\left[\sum_{k=t+1}^{\infty} R_k \middle| S_t = s, A_t = a\right] + \mathbb{E}_{h,\pi}\left[\sum_{k=1}^{t} R_k \middle| S_t = s\right] \tag{7a}$$

$$= \mathbb{E}_{h,\pi}\left[\sum_{k=1}^{\infty} R_k \middle| S_t = s, A_t = a\right] \tag{7b}$$

$$= \mathbb{E}_{h,\pi}[G_0|S_t = s, A_t = a], \tag{7c}$$

which is just the expected total episodic reward. Now define function f to be

$$f(s,a) := (q_\pi(s,a) + b(s))\nabla \log \pi(a|s) = \mathbb{E}_{h,\pi}\left[G_0|S_t = s, A_t = a\right] \nabla \log \pi(a|s) \tag{8a}$$

$$= \mathbb{E}_{h,\pi}\left[G_0 \nabla \log \pi(a|s)|S_t = s, A_t = a\right], \tag{8b}$$

then applying the law of total expectation yields

$$\nabla v_\pi \propto \mathbb{E}_{h,\pi}[f(S_t, A_t)] = \mathbb{E}_{h,\pi}\left[G_0 \nabla \log \pi(A_t|S_t)\right]. \tag{9}$$

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* Adaptive Computation and Machine Learning Series, Cambridge, Massachusetts: The MIT Press, second edition ed., 2018.