This is the first article of this new series. All the posts will be written in English to reduce the unclear translation of specialised vocabulary. Kindly please feel free to contact me if any grammatical mistakes or whatever.

Policy gradient is an approch to solve such reinforcement learning problems. This article will focus on the most fundamental paper for this kind of methods. Most of this article is adapted from the original paper that named [*Policy Gradient Methods for Reinforcement Learning with Function Approximation*][1] (you can find the paper via this link).

Notation

Symbol	Meaning
$s \in S, a \in A, r \in R$	States, Actions, Rewards
s_t, a_t, r_t	State, action, and reward at time step t of one trajectory
$\gamma \in (0, 1]$	Discount factor
$P^a_{ss'} = Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\}$	State transition probability from state s to s' with action a
$V^\pi(s)$	State-value function when following a policy π
$Q^\pi(s, a)$	Action-value function when following a policy π

Why Policy Gradient

The goal fo reinforcement learning is to find an optimal policy for the agent to obtain optimal rewards. Value-function approach, an alternative to solve reinforcement learning problems, has several limitations:

1. aim to find deterministic policies (greedy policy over values)
   - optimal policy is often stochastic, selecting different actions with specific probabilities
2. an arbitrarily small change in the estimated value of an action can cause it to be, or not be, selected.
   - Such discontinuous changes have been identified as a key obstacle to establishing convergence assurances for algorithms following the value-function approach.
3. too expensive computationally in the high dimensional space (continuous), suffering from the curse of dimensionality.

Instead of value-based methods, policy gradient methods can approximate a stochastic directly using an independent function approximator with its own parameters. The policy can be written as $\pi(a \mid s, \theta) = Pr\{a_t = a \mid s_t = s, \theta\}$, where θ is a parameter vector. With abbreviating $\pi(a \mid s, \theta)$ to $\pi(a \mid s)$, we assume that $\nabla_\theta \pi(a \mid s)$ always existis.

Policy Gradient

We first define the objective function $J(\theta)$ to update θ with a postitive-definite step size α, i.e. $\Delta\theta \approx \alpha \frac{\partial J(\theta)}{\partial \theta}$. $J(\theta)$ can be defined in several ways and here lists two most common definitions (we only consider the discrete cases and the results are easy to apply in the continuos cases by simply replacing summations with integrals).

1. **Average Reward Formulation**

   $J(\theta) = \lim_{n \to \infty} E\{r_1 + r_2 + \cdots + r_n \mid \pi\} = \sum_s d^\pi(s) \sum_a \pi(a \mid s) R_s^a$

   where $d^\pi(s) = lim_{t \to \infty} Pr\{s_t = s \mid s_0, \pi\}$ is the stationary distribution of states under π. We assume that the distribution exists and is independent of s_0 for all policies.

   The value of a state-action pair given a policy (also called Q function) is defined as $Q^\pi(s, a) = \sum_{t=1}^\infty E\{r_t - J(\theta) \mid s_0 = s, a_0 = a, \pi\}, \forall s \in S, a \in A$.

2. **Start-State Formulation**: there is a designated start state s_0 and we only care about the long-term reward obtained from the start state.

   $J(\theta) = E\{\sum_{t=1}^\infty \gamma^{t-1} r_t \mid s_0, \pi\} = V^\pi(s_0)$ and $d^\pi(s) = \sum_{t=0}^\infty \gamma^t Pr\{s_t = s \mid s_0, \pi\}$.

   Q function under the definition can be written as $Q^\pi(s, a) = E\{\sum_{k=1}^\infty \gamma^{k-1} r_{t+k} \mid s_t = s, a_t = a, \pi\}$.

It is tricky to compute the gradient $\nabla_\theta J(\theta)$ since it depends not only on the action selection (directly determined by π_θ), also on the stationay distribution of states following the target selection behaviour (indirectly determined by π_θ. Given that the environment is generally unknown, it is difficult to estimate the effect on the state distribution by a policy update. To address such issues, the paper proposed **Policy Gradient Theorem**, providing a reformation of policy gradient without the derivative of the state distribution $d^\pi(\cdot)$.

Policy Gradient Theorem

> For any MDP, in either the average-reward or start-state formulations,
>
> $$\frac{\partial J(\theta)}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s, a)$$

Proof: Involves several basic derivations of reinforment learning that may not be friendly to freshers.

1. Average Reward Formulation

$$\begin{aligned}
\frac{\partial V^\pi(s)}{\partial \theta} &= \frac{\partial}{\partial \theta}\sum_a \pi(a \mid s) Q^\pi(s,a) \\
&= \sum_a \left[\frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) + \pi(a \mid s)\frac{\partial}{\partial \theta} Q^\pi(s,a)\right] \\
&= \sum_a \left[\frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) + \pi(a \mid s)\frac{\partial}{\partial \theta}\left[R_s^a - J(\theta) + \sum_{s'} P_{ss'}^a V^\pi(s')\right]\right] \\
&= \sum_a \left[\frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) + \pi(a \mid s)\left[-\frac{\partial J(\theta)}{\partial \theta} + \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta}\right]\right]
\end{aligned}$$

Since $-\sum_a \pi(a \mid s)\frac{\partial J(\theta)}{\partial \theta} = -\frac{\partial J(\theta)}{\partial \theta}$, we can have

$$\frac{\partial J(\theta)}{\partial \theta} = \sum_a \left[\frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) + \pi(a \mid s)\sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta}\right] - \frac{\partial V^\pi(s)}{\partial \theta}$$

Summing both sides over the stationary distribution d^π,

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= \sum_s d^\pi(s)\frac{\partial J(\theta)}{\partial \theta} \\
&= \sum_s d^\pi(s)\sum_a \frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) + \sum_s d^\pi(s)\sum_a \pi(a \mid s)\sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} - \sum_s d^\pi(s)\frac{\partial V^\pi(s)}{\partial \theta}
\end{aligned}$$

Since d^π is stationary,

$$\begin{aligned}
\sum_s d^\pi(s)\sum_a \pi(a \mid s)\sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} &= \sum_{s,a,s'} d^\pi(s)\pi(a \mid s) P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \\
&= \sum_{s'} d^\pi(s')\frac{\partial V^\pi(s')}{\partial \theta}
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= \sum_s d^\pi(s)\sum_a \frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) + \sum_{s'} d^\pi(s')\frac{\partial V^\pi(s')}{\partial \theta} - \sum_s d^\pi(s)\frac{\partial V^\pi(s)}{\partial \theta} \\
&= \sum_s d^\pi(s)\sum_a \frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a)
\end{aligned}$$

Q.E.D

2. Start-State Formulation

$$
\begin{aligned}
\frac{\partial V^{\pi}(s)}{\partial \theta} &= \frac{\partial}{\partial \theta} \sum_{a} \pi(a \mid s) Q^{\pi}(s, a) \\
&= \sum_{a}\left[\frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a)+\pi(a \mid s) \frac{\partial}{\partial \theta} Q^{\pi}(s, a)\right] \\
&= \sum_{a}\left[\frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a)+\pi(a \mid s) \frac{\partial}{\partial \theta}\left[R_{s}^{a}+\sum_{s^{\prime}} \gamma P_{s s^{\prime}}^{a} V^{\pi}\left(s^{\prime}\right)\right]\right] \\
&= \sum_{a}\left[\frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a)+\pi(a \mid s) \sum_{s^{\prime}} \gamma P_{s s^{\prime}}^{a} \frac{\partial V^{\pi}\left(s^{\prime}\right)}{\partial \theta}\right] \\
&= \sum_{a} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a)+\sum_{a} \pi(a \mid s) \sum_{s^{\prime}} \gamma P_{s s^{\prime}}^{a} \frac{\partial V^{\pi}\left(s^{\prime}\right)}{\partial \theta} \\
&= \sum_{a} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a)+\sum_{s^{\prime}} \sum_{a} \pi(a \mid s) \gamma P_{s s^{\prime}}^{a} \frac{\partial V^{\pi}\left(s^{\prime}\right)}{\partial \theta} \\
&= \sum_{a} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a)+\sum_{s^{\prime}} \gamma Pr\left\{s \rightarrow s^{\prime}, 1, \pi\right\} \frac{\partial V^{\pi}\left(s^{\prime}\right)}{\partial \theta} \\
&= \sum_{a} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a) \\
&+\sum_{s^{\prime}} \gamma Pr\left\{s \rightarrow s^{\prime}, 1, \pi\right\}\left[\sum_{a^{\prime}} \frac{\partial \pi\left(a^{\prime} \mid s^{\prime}\right)}{\partial \theta} Q^{\pi}\left(s^{\prime}, a^{\prime}\right)+\sum_{s^{\prime \prime}} \gamma Pr\left\{s^{\prime} \rightarrow s^{\prime \prime}, 1, \pi\right\} \frac{\partial V^{\pi}\left(s^{\prime \prime}\right)}{\partial \theta}\right] \\
&= \sum_{a} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a) \\
&+\sum_{s^{\prime}} \gamma Pr\left\{s \rightarrow s^{\prime}, 1, \pi\right\} \sum_{a^{\prime}} \frac{\partial \pi\left(a^{\prime} \mid s^{\prime}\right)}{\partial \theta} Q^{\pi}\left(s^{\prime}, a^{\prime}\right)+\sum_{s^{\prime \prime}} \gamma^{2} Pr\left\{s \rightarrow s^{\prime \prime}, 2, \pi\right\} \frac{\partial V^{\pi}\left(s^{\prime \prime}\right)}{\partial \theta} \\
&= \cdots \\
&= \sum_{x} \sum_{k=0}^{\infty} \gamma^{k} Pr\{s \rightarrow x, k, \pi\} \sum_{a} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a)
\end{aligned}
$$

Then,

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} E\left\{\sum_{t=1} \gamma^{t-1} r_{t} \mid s_{0}, \pi\right\}=\frac{\partial}{\partial \theta} V^{\pi}\left(s_{0}\right) \\
&= \sum_{s} \sum_{k=0}^{\infty} \gamma^{k} Pr\left\{s_{0} \rightarrow s, k, \pi\right\} \sum_{a} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a) \\
&= \sum_{s} d^{\pi}(s) \sum_{a} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^{\pi}(s, a)
\end{aligned}
$$

Q.E.D

We can re-write the gradient in expectation form

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta} &= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) \\ &= \sum_s d^\pi(s) \sum_a \pi(a \mid s) \frac{1}{\pi(a \mid s)} \frac{\partial \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) \\ &= \sum_s d^\pi(s) \sum_a \pi(a \mid s) \frac{\partial \ln \pi(a \mid s)}{\partial \theta} Q^\pi(s,a) \\ &= E_\pi[\nabla_\theta \ln \pi(a \mid s) Q^\pi(s,a)]\end{aligned}$$

As we can see, there is no terms of the form $\nabla_\theta d^\pi(s)$, namely, the effect of policy changes on the distribution of states does not appear. However, $Q^\pi(s,a)$ is normally unkown and must be estimated. One approach is to use the actual returns R_t as an approximation for each $Q^\pi(s_t, a_t)$. This leads to Williams's episodic [*REINFORCE*][2] algorithm, obtaining an unbiased estimate of the gradient.

Consider the case in which Q^π is approximated by a learned function approximator. If the approximation is sufficiently good, we might hope to use it in place of Q^π in the gradient and still point roughly in the direction of the gradient.

Luckily, **Compatible Function Approximation Theorem** tells us that the approximator can avoid introducing any bias, still following the exact policy gradient if satisfies two conditions.

Compatible Function Approximation Theorem

> If the following two conditions are satisfied:
>
> 1. Value function approximator is compatible to the policy
>
> $$\nabla_w Q_w(s,a) = \nabla_\theta \ln \pi(a \mid s)$$
>
> 2. Value function parameters w minimise the mean-squared error
>
> $$\varepsilon = E_\pi[(Q^\pi(s,a) - Q_w(s,a))^2]$$
>
> Then the policy gradient is exact,
>
> $$\nabla_\theta(J(\theta)) = E_\pi[\nabla \ln \pi(a \mid s) Q_w(s,a)]$$

Proof:

If the process of minimising the mean-squared error has converged to a local optimum, then

$$\sum_s d^\pi(s) \sum_a \pi(a \mid s)[Q^\pi(s,a) - Q_w(s,a)]\nabla_w Q_w(s,a) = 0$$

According to the condition 1,

$$\sum_s d^\pi(s) \sum_a \nabla_\theta \pi(a \mid s)[Q^\pi(s,a) - Q_w(s,a)] = 0$$

The above equation tells us that the error in $Q_w(s,a)$ is orthogonal to the gradient of the policy parameterisation. Since the expression is zero, we can subtract it from the policy gradient.

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi(a \mid s) Q^\pi(s,a) \\
&= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi(a \mid s) Q^\pi(s,a) - \sum_s d^\pi(s) \sum_a \nabla_\theta \pi(a \mid s)[Q^\pi(s,a) - Q_w(s,a)] \\
&= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi(a \mid s)[Q^\pi(s,a) - Q^\pi(s,a) + Q_w(s,a)] \\
&= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi(a \mid s) Q_w(s,a)
\end{aligned}
$$

Q.E.D

Remainder

This vanilla policy gradient update as shown above has no bias but high variance. The following algorithms were proposed to reduce the variance while keeping the bias unchanged. A general form of policy gradient forms was summarised by a paper named [*High-dimensional Continuous Control using Generalized Advantage Estimation*][3], which will be introduced later.

Also, the paper proves that a form of policy iteration with function approximation is convergent to a locally optimal policy that will be covered with other materials.

Reference

[1]: Sutton, Richard S., et al. "Policy gradient methods for reinforcement learning with function approximation." *Advances in neural information processing systems*. 2000.

[2]: Williams, Ronald J. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." *Machine learning* 8.3-4 (1992): 229-256.

[3]: Schulman, John, et al. "High-dimensional continuous control using generalized advantage estimation." *arXiv preprint arXiv:1506.02438* (2015).