Policy Gradient Methods with Function Approximation
Reinforcement Learning

Richard S. Sutton, David McAllester, Satinder Singh, Yishay Mansour
Proceeding NIPS'99 Proceedings of the 12th International
Conference on Neural Information Processing Systems

Bahareh Harandizadeh
email: bharandi@uci.edu

This presentation is provided using the reference paper slides, also David Silver lecture slides
“standard approach” to reinforcement learning (RL) is to
  • estimate a value function ($V$- or $Q$-function) and then
  • define a “greedy” policy on top of it

somehow “indirect”

oriented towards deterministic policies

problems:
  • “strong causality” violated (small changes have drastic effects)
  • lacking desired convergence properties
  • not really biologically plausible?
Generalized Policy Iteration

- **Sarsa**
- **Monte-Carlo**

**Policy evaluation** Estimate \( v_\pi \)
  e.g. Iterative policy evaluation

**Policy improvement** Generate \( \pi' \geq \pi \)
  e.g. Greedy policy improvement

\[
Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left( q_t^{(n)} - Q(S_t, A_t) \right)
\]
Introduction: Policy gradient approaches to RL

- approximate a stochastic policy
- represent policy directly by an independent function approximator (the “actor”) with own parameters $\theta$
- adapt policy according to

$$\Delta \theta \approx \alpha \frac{\partial \rho(\pi)}{\partial \theta}$$

where $\rho(\pi)$ is a performance measure of the policy $\pi$ and $\alpha$ a positive step-size

- subsumes known methods such as actor-critic approaches and the REINFORCE algorithms

Christian Igel: Policy Gradient Methods with Function Approximation
Notation

- discrete time $t$, states $S$, actions $A$
- $P_{ss'}^a = \Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\}$
- $R_s^a = \mathbb{E}\{r_{t+1} \mid s_t = s, a_t = a\} = \sum_{s'} P_{ss'}^a R_{ss'}^a$
- $\pi(s, a, \theta) = \pi(s, a) = \Pr\{a_t = a \mid s_t = s, \theta\}$
- $\Pr\{s \xrightarrow{k} x \mid \pi\}$: probability of going from state $s$ to state $x$ in $k$ steps under policy $\pi$ ($\Pr\{s \xrightarrow{0} s \mid \pi\} = 1$)
Average reward formulation I

expected reward per time step

\[ \rho(\pi) = \lim_{t \to \infty} \frac{1}{t} \mathbb{E}\{r_1 + \cdots + r_t \mid \pi\} = \sum_s d^\pi(s) \sum_a \pi(s, a) R^a_s \]

we assume that the stationary distribution \( d^\pi \) of states under \( \pi \) exists and is independent of \( s_0 \) (i.e., the process is ergodic)

\[ d^\pi(s) = \lim_{t \to \infty} \Pr\{s_t = s \mid s_0, \pi\} = \lim_{t \to \infty} \Pr\{s_t = s \mid \pi\} \]
An Example of Stationary Distributions

- A Markov chain:

\[
P = \begin{bmatrix}
    0.7 & 0.3 & 0.0 \\
    0.3 & 0.4 & 0.3 \\
    0.0 & 0.3 & 0.7 \\
\end{bmatrix}
\]

- The stationary distribution is \( \pi = \begin{bmatrix}
    \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\end{bmatrix} \)

\[
\begin{bmatrix}
    \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\end{bmatrix} \begin{bmatrix}
    0.7 & 0.3 & 0.0 \\
    0.3 & 0.4 & 0.3 \\
    0.0 & 0.3 & 0.7 \\
\end{bmatrix} = \begin{bmatrix}
    \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\end{bmatrix}
\]

ref: https://medium.com/@kim_hjun/markov-chain-stationary-distribution-5198941234f6
in general it holds

\[ \Pr\{s_{t+1} = s' \mid \pi\} = \sum_s \Pr\{s_t = s \mid \pi\} \Pr\{s_{t+1} = s' \mid s_t = s, \pi\} \]

the stationarity

\[ d^\pi(s') = \lim_{t \to \infty} \Pr\{s_t = s' \mid \pi\} = \lim_{t \to \infty} \Pr\{s_{t+1} = s' \mid \pi\} \]

implies for \( t \to \infty \)

\[ d^\pi(s') = \sum_s d^\pi(s) \sum_a \pi(s, a) P^a_{ss'} \sum_{s_{t+1}=s'} \Pr\{s_{t+1}=s' \mid s_t=s, \pi\} \]
we define

$$Q^\pi(s, a) = \sum_{t=1}^{\infty} \mathbb{E}\{r_t - \rho(\pi) \mid s_0 = s, a_0 = a, \pi\}$$

with

$$V^\pi(s) = \sum_a \pi(s, a) Q^\pi(s, a)$$

we have

$$Q^\pi(s, a) = \sum_{t=1}^{\infty} \mathbb{E}\{r_t - \rho(\pi) \mid s_0 = s, a_0 = a, \pi\} = R_s^a - \rho(\pi) + \sum s' P_{ss'}^a V^\pi(s')$$
Start-state formulation

expected reward per time step

$$\rho(\pi) = \mathbb{E}\left\{ \sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_0, \pi \right\}$$

with $$\gamma \in [0, 1]$$, $$\gamma = 1$$ only for episodic tasks, we have

$$Q^\pi(s, a) = \mathbb{E}\left\{ \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k} \mid s_t = s, a_t = a, \pi \right\}$$

we define (ignoring the normalization constant $$(1 - \gamma)$$)

$$d^\pi(s) = \sum_{t=0}^{\infty} \gamma^{t} \Pr\{s_t = s \mid s_0, \pi\}$$

it holds

$$d^\pi(s) = \sum_{k=0}^{\infty} \gamma^{k} \Pr\{s_0 \rightarrow s \mid \pi\}$$
Theorem

For any MDP, in either average-reward or start-state formulations,

\[
\frac{\partial \rho(\pi)}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a).
\]
\[
\frac{\partial V^{\pi}(s)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_a \pi(s, a) Q^{\pi}(s, a)
\]

\[
= \sum_a \left[ \frac{\partial \pi(s, a)}{\partial \theta} Q^{\pi}(s, a) + \pi(s, a) \frac{\partial}{\partial \theta} Q^{\pi}(s, a) \right]
\]

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

\[
= \sum_a \left[ \frac{\partial \pi(s, a)}{\partial \theta} Q^{\pi}(s, a) + \pi(s, a) \left[ - \frac{\partial \rho(\pi)}{\partial \theta} + \sum_{s'} P_{ss'}^{a} \frac{\partial V^\pi(s')}{\partial \theta} \right] \right]
\]
\[
\frac{\partial V^\pi(s)}{\partial \theta} = \sum_a \left[ \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) + \pi(s, a) \left( -\frac{\partial \rho(\pi)}{\partial \theta} + \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \right) \right] \Rightarrow \\
\frac{\partial \rho(\pi)}{\partial \theta} = \sum_a \left[ \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) + \pi(s, a) \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \right] - \frac{\partial V^\pi(s)}{\partial \theta} \Rightarrow \\
\sum_s d^\pi(s) \frac{\partial \rho(\pi)}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) \\
+ \sum_s d^\pi(s) \sum_a \pi(s, a) \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} - \sum_s d^\pi(s) \frac{\partial V^\pi(s)}{\partial \theta}
\]
\[
\sum_s d^\pi(s) \frac{\partial \rho(\pi)}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a)
\]
\[
+ \sum_s d^\pi(s) \sum_a \pi(s, a) \sum_{s'} P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} - \sum_s d^\pi(s) \frac{\partial V^\pi(s)}{\partial \theta}
\]
\[
= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) + \sum_s d^\pi(s) \frac{\partial V^\pi(s)}{\partial \theta} - \sum_s d^\pi(s) \frac{\partial V^\pi(s)}{\partial \theta}
\]

\[
\frac{\partial \rho(\pi)}{\partial \theta} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a)
\]
\[
\frac{\partial V^\pi(s)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_a \pi(s, a) Q^\pi(s, a) \\
= \sum_a \left[ \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) + \pi(s, a) \frac{\partial}{\partial \theta} Q^\pi(s, a) \right] \\
= \sum_a \left[ \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) + \pi(s, a) \frac{\partial}{\partial \theta} \left[ R_s^a + \sum_{s'} \gamma P_{ss'}^a V^\pi(s') \right] \right] \\
= \sum_a \left[ \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) + \pi(s, a) \sum_{s'} \gamma P_{ss'}^a \frac{\partial V^\pi(s')}{\partial \theta} \right]
\]
\[
\frac{\partial V_\pi(s)}{\partial \theta} = \sum_a \left[ \frac{\partial \pi(s, a)}{\partial \theta} Q_\pi(s, a) + \pi(s, a) \sum_{s'} \gamma P_{ss'}^a \frac{\partial V_\pi(s')}{\partial \theta} \right]
\]

\[
= \gamma^0 \Pr\{s \xrightarrow{0} s' \mid \pi\} \sum_a \frac{\partial \pi(s, a)}{\partial \theta} Q_\pi(s, a) + \sum_{s'} \gamma^1 \Pr\{s \xrightarrow{1} s' \mid \pi\} \frac{\partial V_\pi(s')}{\partial \theta}
\]

\[
= \sum_{s'} \left[ \gamma^0 \Pr\{s \xrightarrow{0} s' \mid \pi\} \sum_a \frac{\partial \pi(s', a)}{\partial \theta} Q_\pi(s', a) + \gamma^1 \Pr\{s \xrightarrow{1} s' \mid \pi\} \frac{\partial V_\pi(s')}{\partial \theta} \right]
\]

\[
= \sum_{s'} \sum_{k=0}^{\infty} \gamma^k \Pr\{s \xrightarrow{k} s' \mid \pi\} \sum_a \frac{\partial \pi(s', a)}{\partial \theta} Q_\pi(s', a)
\]

Christian Igel: Policy Gradient Methods with Function Approximation

15 / 25
Proof policy gradient, start-state III

\[
\frac{\partial \rho(\pi)}{\partial \theta} = \frac{\partial}{\partial \theta} \mathbb{E} \left\{ \sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_0, \pi \right\} = \frac{\partial V^\pi(s_0)}{\partial \theta}
\]

\[
\frac{\partial V^\pi(s_0)}{\partial \theta} = \sum_{s} \sum_{k=0}^{\infty} \gamma^k \Pr\{s_0 \xrightarrow{k} s \mid \pi\} \sum_{a} \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a) \tag{d^\pi(s), see slide 8}
\]

\[
= \sum_{s} d^\pi(s) \sum_{a} \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a)
\]
Theorem (Compatible Function Approximation Theorem)

If the following two conditions are satisfied:

1. Value function approximator is compatible to the policy

\[ \nabla_w Q_w(s, a) = \nabla_\theta \log \pi_\theta(s, a) \]

2. Value function parameters \( w \) minimise the mean-squared error

\[ \varepsilon = \mathbb{E}_{\pi_\theta} \left[ (Q^{\pi_\theta}(s, a) - Q_w(s, a))^2 \right] \]

Then the policy gradient is exact,

\[ \nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta} \left[ \nabla_\theta \log \pi_\theta(s, a) \ Q_w(s, a) \right] \]
• $Q^\pi$ is approximated by a learned function approximator (e.g., neural net)
  
  $$f_w : S \times A \to \mathbb{R}$$

  with parameters $w$ (the “critic”)

• natural choice of learning rule

  $$\Delta w_t \propto \frac{\partial}{\partial w} \left[ \hat{Q}^\pi(s_t, a_t) - f_w(s_t, a_t) \right]^2$$

  $$\propto \left[ \hat{Q}^\pi(s_t, a_t) - f_w(s_t, a_t) \right] \frac{\partial f_w(s_t, a_t)}{\partial w}$$

  where $\hat{Q}^\pi(s_t, a_t)$ is some unbiased estimate of $Q^\pi(s_t, a_t)$ (e.g., $R_t$)
convergence to local optimum

\[ 0 = \Delta w_t \propto \left[ \hat{Q}^\pi(s_t, a_t) - f_w(s_t, a_t) \right] \frac{\partial f_w(s_t, a_t)}{\partial w} \]

in expectation given \( \pi \) implies “convergence condition”

\[ \sum_s d^\pi(s) \sum_a \pi(s, a) \left[ Q^\pi(s, a) - f_w(s, a) \right] \frac{\partial f_w(s, a)}{\partial w} = 0 \]
Theorem

If \( f_w \) satisfies the “convergence condition” and is compatible with the policy parameterization in the sense of

\[
\frac{\partial f_w(s, a)}{\partial w} = \frac{\partial \pi(s, a)}{\partial \theta} \frac{1}{\pi(s, a)}
\]
Proof of policy gradient with function approximation I

“compatibility”

\[
\frac{\partial f_w(s, a)}{\partial w} = \frac{\partial \pi(s, a)}{\partial \theta} \frac{1}{\pi(s, a)}
\]

and “convergence condition”

\[
\sum_s d^\pi(s) \sum_a \pi(s, a) \left[ \hat{Q}^\pi(s, a) - f_w(s, a) \right] \frac{\partial f_w(s, a)}{\partial w} = 0
\]

imply

\[
\sum_s d^\pi(s) \sum_a \frac{\partial \pi(s, a)}{\partial \theta} \left[ \hat{Q}^\pi(s, a) - f_w(s, a) \right] = 0
\]
\[
\frac{\partial \rho(\pi)}{\partial \theta} = \sum_s \sum_a d^\pi(s) \frac{\partial \pi(s, a)}{\partial \theta} Q^\pi(s, a)
\]

\[= \sum_s \sum_a d^\pi(s) \frac{\partial \pi(s, a)}{\partial \theta} [Q^\pi(s, a) - f_w(s, a)]\]

0, see previous slide

\[= \sum_s \sum_a d^\pi(s) \frac{\partial \pi(s, a)}{\partial \theta} [Q^\pi(s, a) - Q^\pi(s, a) + f_w(s, a)]\]

\[= \sum_s \sum_a d^\pi(s) \frac{\partial \pi(s, a)}{\partial \theta} f_w(s, a)\]
consider vector of features (→ RBF networks, CMACs etc.)
φ(s, a), ∀a ∈ A, s ∈ S; policy is a Gibbs distribution in a linear combination of the features

\[ \pi(s, a) = \frac{e^{\theta^T \phi(s,a)}}{\sum_b e^{\theta^T \phi(s,b)}} \]

compatibility condition requires

\[ \frac{\partial f_w(s, a)}{\partial w} = \frac{\partial \pi(s, a)}{\partial \theta} \frac{1}{\pi(s, a)} = \phi(s, a) - \sum_b \pi(s, b) \phi(s, b) \]

leading to the natural parameterization of \( f_w \)

\[ f_w(s, a) = w^T \left[ \phi(s, a) - \sum_b \pi(s, b) \phi(s, b) \right] \]
• $f(w)(s, a) = w^T [\phi(s, a) - \sum b \pi(s, b)\phi(s, b)]$ implies $\forall s \in S$

$$
\sum a \pi(s, a)f_w(s, a) = 0
$$

$\rightarrow f_w$ approximates advantage function $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$ rather than $Q^\pi(s, a)$

**function QAC**

Initialise $s, \theta$
Sample $a \sim \pi_\theta$
for each step do
    Sample reward $r = R^a_s$; sample transition $s' \sim P^a_s$.
    Sample action $a' \sim \pi_\theta(s', a')$
    $\delta = r + \gamma Q_w(s', a') - Q_w(s, a)$
    $\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(s, a)Q_w(s, a)$
    $w \leftarrow w + \beta \delta \phi(s, a)$
    $a \leftarrow a', s \leftarrow s'$
end for
end function
Demo Time

https://colab.research.google.com/drive/1YGhWx20p-sIbKvRpBXCpUzWVWAkpzjj3

https://gym.openai.com/envs/CartPole-v0/
question?
What is the Value-based, Policy based and Actor critic and what are the advantages and disadvantages of each of them?