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

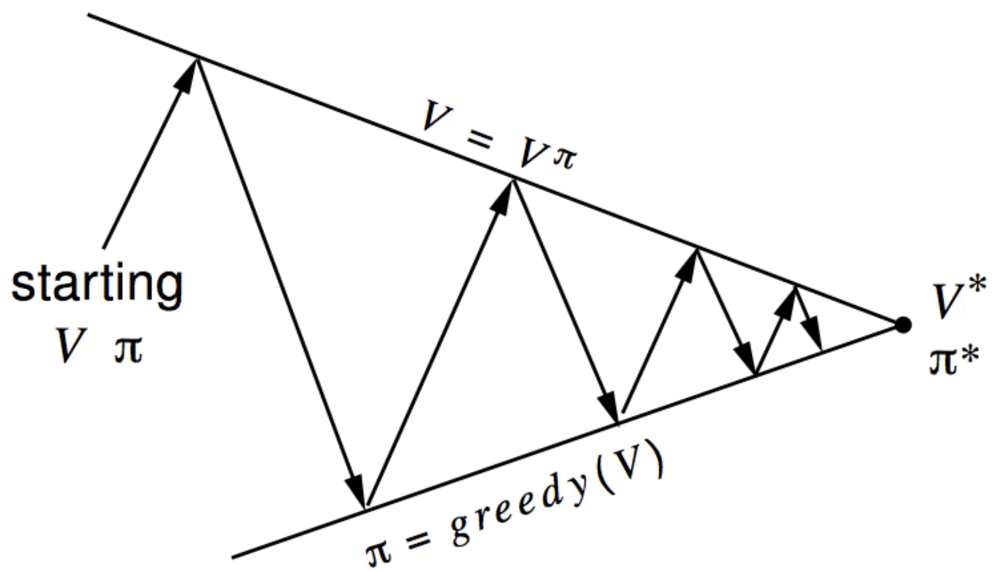
This presentation is provided using the reference paper slides, also David Silver lecture slides

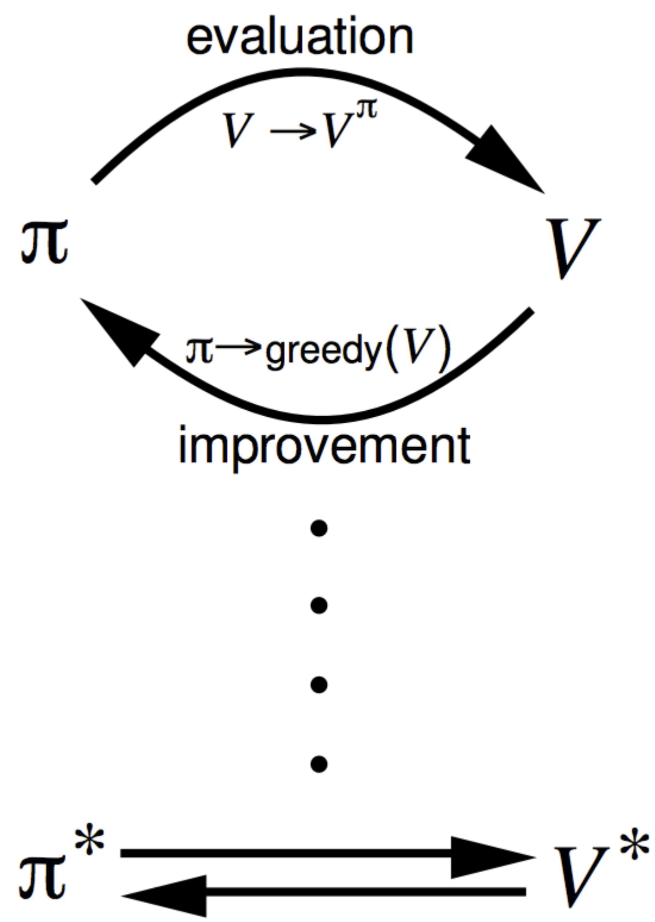
Introduction: Value function approaches to RL

- "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**
- **Monto-Carlo**





Policy evaluation Estimate v_π
e.g. Iterative policy evaluation

Policy improvement Generate $\pi' \geq \pi$
e.g. Greedy policy improvement

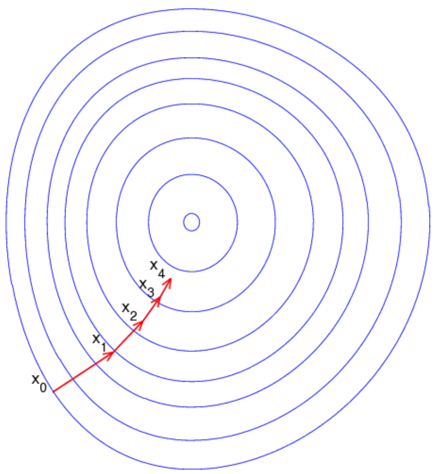
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(q_t^{(n)} - Q(S_t, A_t) \right)$$

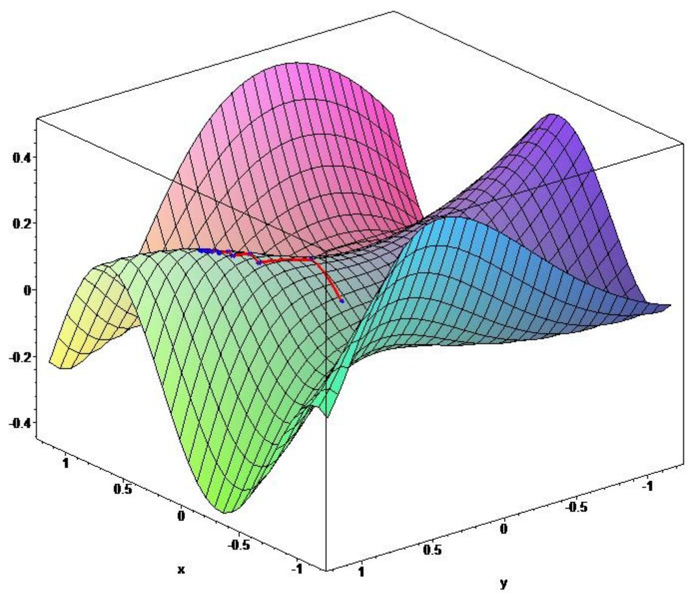
Introduction: Policy gradient approaches to RL

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

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

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





- Where $\nabla_\theta J(\theta)$ is the policy gradient

$$\nabla_\theta J(\theta) = \begin{pmatrix} \frac{\partial J(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{pmatrix}$$

Notation

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

Average reward formulation I

expected reward per time step

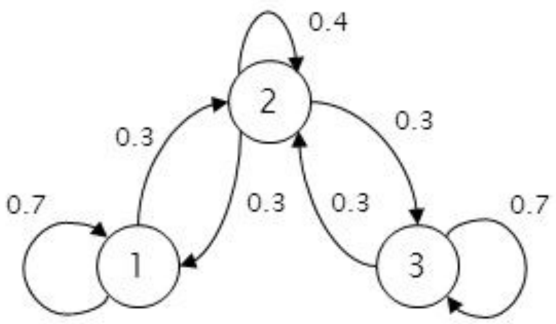
$$\rho(\pi) = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\{r_1 + \cdots + r_t \,|\, \pi\} = \sum_s d^\pi(s) \sum_a \pi(s,a) R_s^a$$

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

$$d^\pi(s) = \lim_{t\to\infty} \Pr\{s_t = s \,|\, s_0, \pi\} = \lim_{t\to\infty} \Pr\{s_t = s \,|\pi\}$$

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 $\underset{\sim}{\pi} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$

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

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

Average reward formulation II

in general it holds

$$\Pr\{s_{t+1} = s' \,|\, \pi\} = \sum_s \Pr\{s_t = s \,|\, \pi\} \Pr\{s_{t+1} = s' \,|\, s_t = s, \pi\}$$

the stationarity

$$d^\pi(s') = \lim_{t \to \infty} \Pr\{s_t = s' \,|\, \pi\} = \lim_{t \to \infty} \Pr\{s_{t+1} = s' \,|\, \pi\}$$

implies for $t \to \infty$

$$d^\pi(s') = \sum_s \underbrace{d^\pi(s)}_{\Pr\{s_t = s \,|\pi\}} \underbrace{\sum_a \pi(s,a) P^a_{ss'}}_{\Pr\{s_{t+1} = s' \,|\, s_t = s, \pi\}}$$

Average reward formulation III

we define

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

with

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

we have

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

Start-state formulation

expected reward per time step

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

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

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

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

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

Policy gradient theorem

Theorem

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

$$\frac{\partial \rho(\pi)}{\partial \boldsymbol{\theta}} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} Q^\pi(s,a) \quad .$$

Proof policy gradient, average reward I

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

Proof policy gradient, average reward II

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

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

$$\sum_s d^\pi(s) \frac{\partial \rho(\pi)}{\partial \boldsymbol{\theta}} = \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} Q^\pi(s,a)$$
$$+ \sum_s d^\pi(s) \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} \frac{\partial V^\pi(s')}{\partial \boldsymbol{\theta}} - \sum_s d^\pi(s) \frac{\partial V^\pi(s)}{\partial \boldsymbol{\theta}}$$

Proof policy gradient, average reward III

$$\sum_s d^\pi(s)\frac{\partial\rho(\pi)}{\partial\boldsymbol{\theta}} = \sum_s d^\pi(s)\sum_a \frac{\partial\pi(s,a)}{\partial\boldsymbol{\theta}}Q^\pi(s,a)$$

$$+\underbrace{\sum_s d^\pi(s)\sum_a \pi(s,a)\sum_{s'} P^a_{ss'}}_{d^\pi(s),\ \text{see slide 6}}\frac{\partial V^\pi(s')}{\partial\boldsymbol{\theta}} - \sum_s d^\pi(s)\frac{\partial V^\pi(s)}{\partial\boldsymbol{\theta}}$$

$$= \sum_s d^\pi(s)\sum_a \frac{\partial\pi(s,a)}{\partial\boldsymbol{\theta}}Q^\pi(s,a) + \sum_s d^\pi(s)\frac{\partial V^\pi(s)}{\partial\boldsymbol{\theta}} - \sum_s d^\pi(s)\frac{\partial V^\pi(s)}{\partial\boldsymbol{\theta}}$$

$$\boxed{\frac{\partial\rho(\pi)}{\partial\boldsymbol{\theta}} = \sum_s d^\pi(s)\sum_a \frac{\partial\pi(s,a)}{\partial\boldsymbol{\theta}}Q^\pi(s,a)}$$

Proof policy gradient, start-state I

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

Proof policy gradient, start-state II

$$\frac{\partial V^\pi(s)}{\partial \boldsymbol{\theta}} = \sum_a \left[\frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} Q^\pi(s,a) + \pi(s,a) \sum_{s'} \gamma P^a_{ss'} \frac{\partial V^\pi(s')}{\partial \boldsymbol{\theta}} \right]$$

$$= \gamma^0 \Pr\{s \xrightarrow{0} s \mid \pi\} \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} Q^\pi(s,a) + \sum_{s'} \gamma^1 \Pr\{s \xrightarrow{1} s' \mid \pi\} \frac{\partial V^\pi(s')}{\partial \boldsymbol{\theta}}$$

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

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

Proof policy gradient, start-state III

$$\frac{\partial \rho(\pi)}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\left\{ \sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_0, \pi \right\} = \frac{\partial V^{\pi}(s_0)}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial V^{\pi}(s_0)}{\partial \boldsymbol{\theta}} = \sum_s \underbrace{\sum_{k=0}^{\infty} \gamma^k \Pr\{s_0 \xrightarrow{k} s \mid \pi\}}_{d^{\pi}(s),\ \text{see slide 8}} \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} Q^{\pi}(s,a)$$

$$\boxed{= \sum_s d^{\pi}(s) \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} Q^{\pi}(s,a)}$$

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*

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

Then the policy gradient is exact,

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\ Q_w(s,a)\right]$$

Function approximation I

- Q^π is approximated by a learned function approximator (e.g., neural net)

$$f_{\boldsymbol{w}} : S \times A \to \mathbb{R}$$

with parameters $\boldsymbol{w}$ (the "critic")

- natural choice of learning rule

$$\begin{aligned}\Delta \boldsymbol{w}_t &\propto \frac{\partial}{\partial \boldsymbol{w}} \left[\hat{Q}^\pi(s_t, a_t) - f_{\boldsymbol{w}}(s_t, a_t) \right]^2 \\ &\propto \left[\hat{Q}^\pi(s_t, a_t) - f_{\boldsymbol{w}}(s_t, a_t) \right] \frac{\partial f_{\boldsymbol{w}}(s_t, a_t)}{\partial \boldsymbol{w}}\end{aligned}$$

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

Function approximation I

convergence to local optimum

$$\mathbf{0} = \Delta \boldsymbol{w}_t \propto \left[\hat{Q}^\pi(s_t, a_t) - f_{\boldsymbol{w}}(s_t, a_t)\right] \frac{\partial f_{\boldsymbol{w}}(s_t, a_t)}{\partial \boldsymbol{w}}$$

in expectation given π implies "convergence condition"

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

Policy gradient with function approximation theorem

Theorem

If $f_{\boldsymbol{w}}$ satisfies the "convergence condition" and is compatible with the policy parameterization in the sense of

$$\frac{\partial f_{\boldsymbol{w}}(s,a)}{\partial \boldsymbol{w}} = \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} \frac{1}{\pi(s,a)}$$

Proof of policy gradient with function approximation I

"compatibility"

$$\frac{\partial f_{\boldsymbol{w}}(s,a)}{\partial \boldsymbol{w}} = \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} \frac{1}{\pi(s,a)}$$

and "convergence condition"

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

imply

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

Proof of policy gradient with function approximation II

$$\frac{\partial \rho(\pi)}{\partial \boldsymbol{\theta}} = \underbrace{\sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} Q^\pi(s,a)}_{\text{policy gradient theorem}}$$

$$- \underbrace{\sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} \left[Q^\pi(s,a) - f_{\boldsymbol{w}}(s,a)\right]}_{\mathbf{0}\text{, see previous slide}}$$

$$= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} \left[Q^\pi(s,a) - Q^\pi(s,a) + f_{\boldsymbol{w}}(s,a)\right]$$

$$= \sum_s d^\pi(s) \sum_a \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} f_{\boldsymbol{w}}(s,a)$$

Application

consider vector of features ($\rightarrow$ RBF networks, CMACs etc.) $\boldsymbol{\phi}(s,a), \forall a \in A, s \in S$; policy is a Gibbs distribution in a linear combination of the features

$$\pi(s,a) = \frac{e^{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\phi}(s,a)}}{\sum_b e^{\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\phi}(s,b)}}$$

compatibility condition requires

$$\frac{\partial f_{\boldsymbol{w}}(s,a)}{\partial \boldsymbol{w}} = \frac{\partial \pi(s,a)}{\partial \boldsymbol{\theta}} \frac{1}{\pi(s,a)} = \boldsymbol{\phi}(s,a) - \sum_b \pi(s,b)\boldsymbol{\phi}(s,b)$$

leading to the natural parameterization of $f_{\boldsymbol{w}}$

$$f_{\boldsymbol{w}}(s,a) = \boldsymbol{w}^\mathsf{T}\left[\boldsymbol{\phi}(s,a) - \sum_b \pi(s,b)\boldsymbol{\phi}(s,b)\right]$$

Remarks

- $f_{\boldsymbol{w}}(s,a) = \boldsymbol{w}^{\mathsf{T}}\left[\boldsymbol{\phi}(s,a) - \sum_b \pi(s,b)\boldsymbol{\phi}(s,b)\right]$ implies $\forall s \in S$

$$\sum_a \pi(s,a) f_{\boldsymbol{w}}(s,a) = 0$$

$\rightarrow f_{\boldsymbol{w}}$ approximates advantage function $A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$ rather than $Q^{\pi}(s,a)$

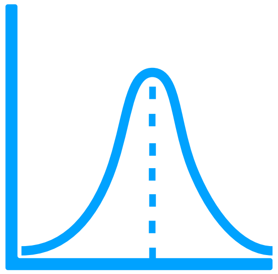
function QAC
  Initialise s, θ
  Sample $a \sim \pi_\theta$
  for each step **do**
    Sample reward $r = \mathcal{R}_s^a$; sample transition $s' \sim \mathcal{P}_{s,\cdot}^a$
    Sample action $a' \sim \pi_\theta(s', a')$
    $\delta = r + \gamma Q_w(s', a') - Q_w(s, a)$
    $\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(s,a) Q_w(s,a)$
    $w \leftarrow w + \beta \delta \phi(s,a)$
    $a \leftarrow a', s \leftarrow s'$
  end for
end function

Demo Time

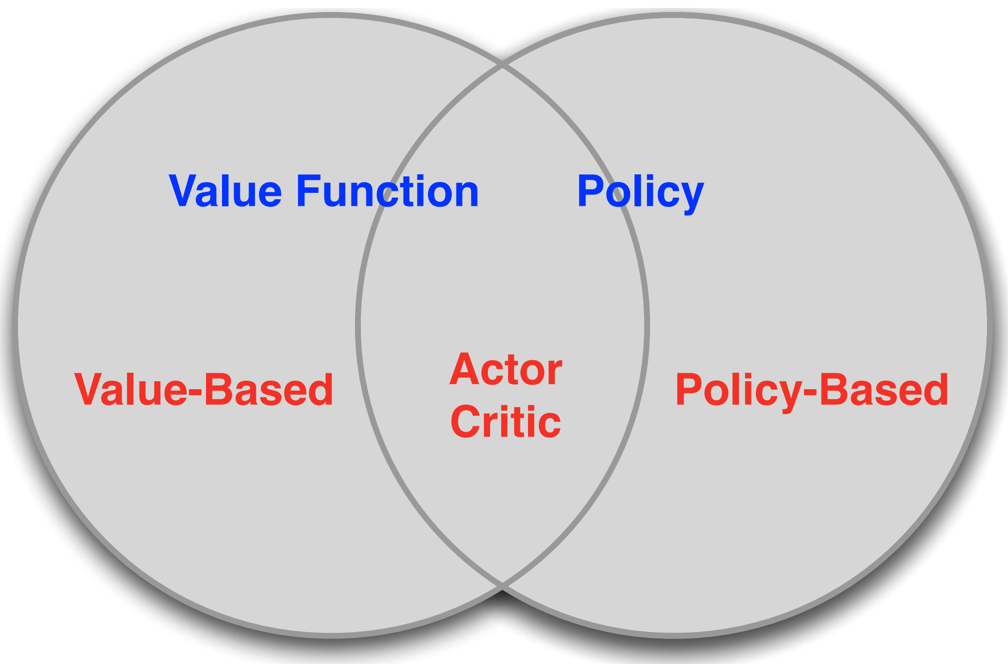
https://colab.research.google.com/drive/1YGhWx20p-slbKvRpBXCpUzWVWAkpzjj3

https://gym.openai.com/envs/CartPole-v0/

question?



Please answer the following question based on this paper and David Silver lecture on Policy Gradient(lecture 7)



What is the Value_based, Policy based and Actor critic and what are the advantages and disadvantages of each of them?