Approximate Policy Iteration and Policy-Based Learning Methods

Chen-Yu Wei

Approximate Policy Iteration (API)

For $k = 1,\ 2, \ldots$

Evaluate $\hat{Q}_k \approx Q^{\pi_k}$

$$\pi_{k+1}(s) \leftarrow \underset{a}{\operatorname{argmax}}\, \hat{Q}_k(s,a)$$

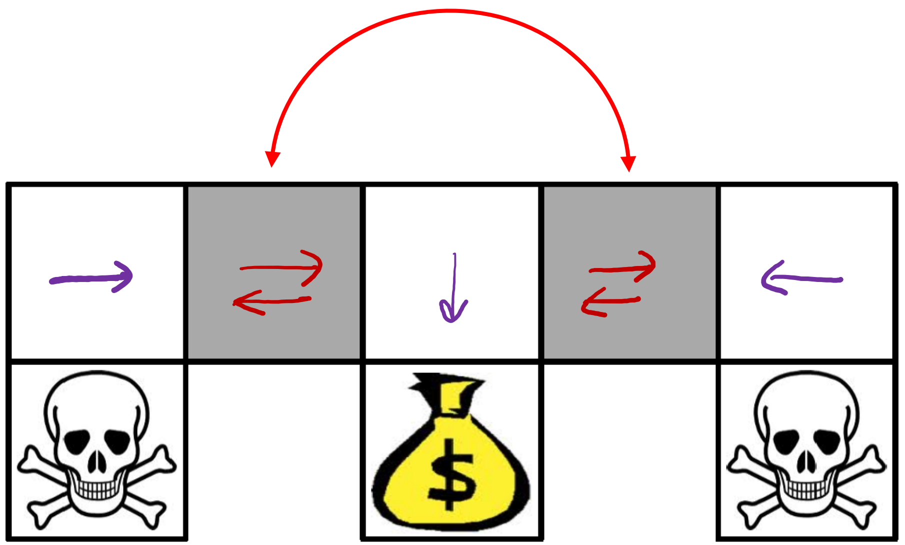
Value-based : Q^*, V^π, V^* $\approx \boxed{V_\theta}$ Q^π

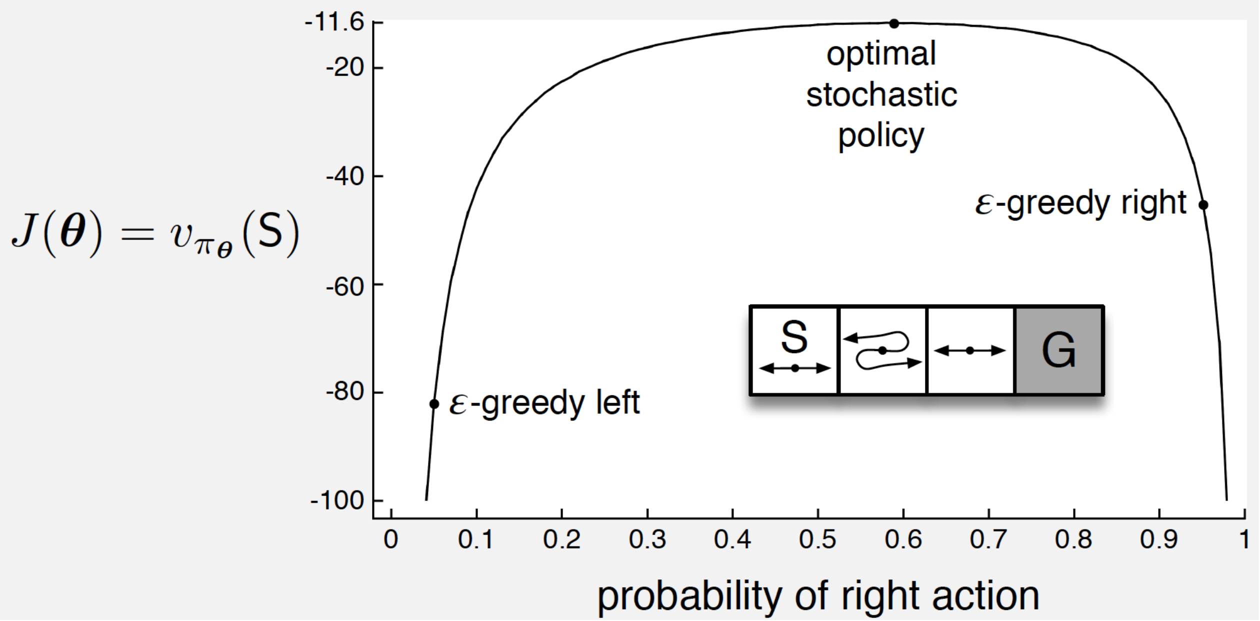
Policy-based : $\pi_\theta(a|s)$

Limitation (shared by all value-based methods):

- Vulnerable to representation error

Limitation of Value Function Approximation





Incremental Policy Updates through Exponential Weights

For $k = 1,\ 2, \ldots$

Evaluate $\hat{Q}_k \approx Q^{\pi_k}$

Perform incremental policy update such as

$$\pi_{k+1}(a|s) \propto \pi_k(a|s) \exp\left(\eta \hat{Q}_k(s,a)\right) \quad \Leftrightarrow \quad \pi_{\theta_{k+1}}(\cdot|s) = \arg\max_{\theta} \left\{ \sum_a \pi_\theta(a|s)\, \hat{Q}_k(s,a) - \frac{1}{\eta} KL\left(\pi_\theta(\cdot|s), \pi_{\theta_k}(\cdot|s)\right) \right\}$$

Function approximation for policies:

$\theta_{k+1} \leftarrow \text{IncrementalUpdate}_\eta(\theta_k)$

$\pi_{k+1} = \pi_{\theta_{k+1}}$

Another Idea: Policy Gradient

Policy-centric learning: maximize the return directly
c.f. Value-centric learning: minimize the Bellman error

For $k = 1,\ 2, \dots$

$$\theta_{k+1} \leftarrow \theta_k + \eta\, \nabla_\theta V^{\pi_\theta}(\rho)\Big|_{\theta=\theta_k}$$

$V^{\pi_\theta}(\rho) \overset{\Delta}{=} \sum_s \rho(s) V^{\pi_\theta}(s)$

What are the differences between exponential weights and policy gradient?

Policy Gradient for Softmax Policy in Expert Problems

Exponential weight

Assume full-information and fixed reward $R = (R(1), \dots, R(A))$

$$\pi_{k+1}(a) = \frac{\pi_k(a)\exp(\eta R(a))}{\sum_b \pi_k(b)\exp(\eta R(b))}$$

Let $\theta = (\theta(1), \dots, \theta(A))$ and $\pi_\theta(a) = \frac{\exp(\theta(a))}{\sum_{b=1}^{A}\exp(\theta(b))}$

$\Rightarrow \nabla_\theta V^{\pi_\theta} = ?$

??

$$V^{\pi_\theta} = \sum_a \pi_\theta(a) R(a)$$

$$\text{PG: } \theta_{k+1} = \theta_k + \eta \nabla_\theta V^{\pi_\theta}\Big|_{\theta=\theta_k}$$

$$\left(\nabla_\theta V^{\pi_\theta}\right)_i = \sum_a \frac{\partial}{\partial\theta_i}\left(\pi_\theta(a)\right) R(a) = \frac{\exp(\theta(i)) R(i)}{\sum_b \exp(\theta(b))} - \sum_a \frac{\exp(\theta(a))\exp(\theta(i)) R(a)}{\left(\sum_b \exp(\theta(b))\right)^2} \quad \checkmark$$

$$\frac{\partial}{\partial\theta_i}\pi_\theta(a) \begin{cases} \text{when } a = i: & \frac{\partial}{\partial\theta_i}\pi_\theta(a) = \frac{\partial}{\partial\theta(i)}\left[\frac{\exp(\theta(i))}{\sum_b \exp(\theta(b))}\right] = \frac{\exp(\theta(i))\left(\sum_b \exp(\theta(b))\right) - \exp(\theta(i))\cdot\exp(\theta(i))}{\left(\sum_b \exp(\theta(b))\right)^2} \\ \text{when } a \neq i: & \frac{\partial}{\partial\theta_i}\pi_\theta(a) = \frac{\partial}{\partial\theta(i)}\left[\frac{\exp(\theta(a))}{\sum_b \exp(\theta(b))}\right] = \frac{0 - \exp(\theta(a))\exp(\theta(i))}{\left(\sum_b \exp(\theta(b))\right)^2} \end{cases}$$

$$\underline{\left(\nabla_\theta V^{\pi_\theta}\right)_i} = \frac{\exp(\theta(i))\,R(i)}{\sum_b \exp(\theta(b))} - \sum_a \frac{\exp(\theta(a))\exp(\theta(i))\,R(a)}{\left(\sum_b \exp(\theta(b))\right)^2}$$

$$= \frac{\exp(\theta(i))}{\sum_b \exp(\theta(b))}\left(R(i) - \sum_a \frac{\exp(\theta(a))}{\sum_b \exp(\theta(b))} R(a)\right)$$

$$= \pi_\theta(i)\left(R(i) - \sum_a \pi_\theta(a) R(a)\right)$$

PG: $\theta_{k+1}(i) \leftarrow \theta_k(i) + \eta\, \pi_{\theta_k}(i)\left(R(i) - \sum_a \pi_{\theta_k}(a) R(a)\right)$

Comparison between EW and PG over softmax policies

$$\theta = (\theta(a), \ldots, \theta(A)), \qquad \pi_\theta(a) = \frac{\exp(\theta(a))}{\sum_b \exp(\theta(b))}, \qquad V^{\pi_\theta} = \sum_a \pi_\theta(a)\, R(a)$$

Policy Gradient over softmax policies

For $k = 1,2, \ldots$

$$\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta \pi_{\theta_k}(a) A(a)$$

Exponential weights

For $k = 1,2, \ldots$

$$\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta A(a)$$

Two Types of Policy Gradients

Policy Gradient over softmax policies

$$\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta \pi_{\theta_k}(a) R(a)$$

||

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \left\langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}} \right\rangle - \frac{1}{2\eta} \|\theta - \theta_k\|^2$$

(Vanilla) Policy Gradient

Exponential weights

$$\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta R(a)$$

||

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \left\langle \pi_\theta - \pi_{\theta_k}, R \right\rangle - \frac{1}{\eta} \mathrm{KL}\left(\pi_\theta, \pi_{\theta_k}\right)$$

Natural Policy Gradient

Approximating the NPG Update

$$\boxed{\theta_{k+1} = \operatorname*{argmax}_{\theta} \langle \pi_\theta - \pi_{\theta_k}, R \rangle - \frac{1}{\eta} \mathrm{KL}\left(\pi_\theta, \pi_{\theta_k}\right)}$$

When $\theta_{k+1} \approx \theta_k$ (i.e., when η is small), the following hold:

$$\langle \pi_\theta - \pi_{\theta_k}, R \rangle = V^{\pi_\theta} - V^{\pi_{\theta_k}} \approx (\theta - \theta_k)^\top \nabla_\theta V^{\pi_\theta}\Big|_{\theta=\theta_k}$$

$$\mathrm{KL}\left(\pi_\theta, \pi_{\theta_k}\right) \approx (\theta - \theta_k)^\top F_{\theta_k} (\theta - \theta_k)$$

where $F_{\theta_k} := \sum_a \pi_\theta(a) \left(\nabla_\theta \log \pi_\theta(a)\right)\left(\nabla_\theta \log \pi_\theta(a)\right)^\top\Big|_{\theta=\theta_k}$

(Fisher information matrix)