Approximate Policy Iteration and Policy-Based Learning Methods

Chen-Yu Wei

Approximate Policy Iteration (API)

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

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

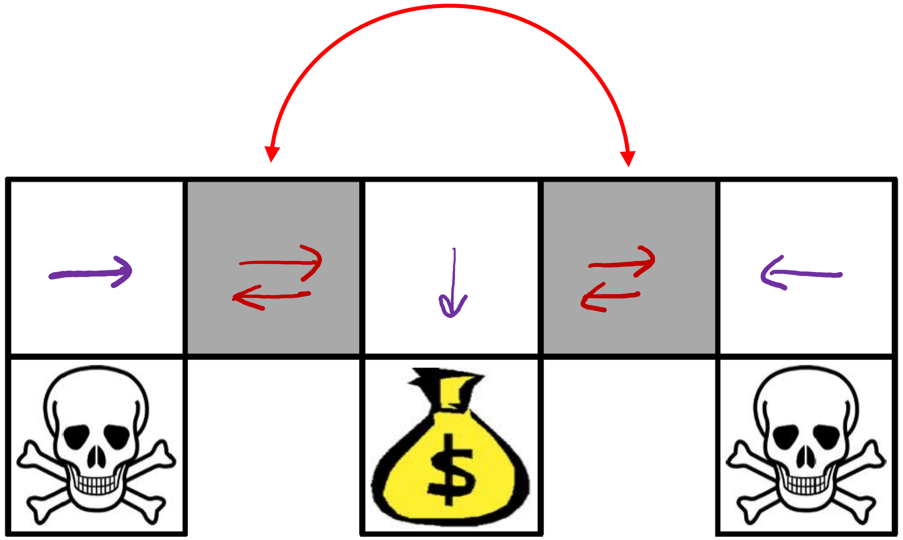
$$\pi_{k+1}(s) \leftarrow \underset{a}{\operatorname{argmax}}\, \hat{Q}_k(s, a)$$

Q^{π}

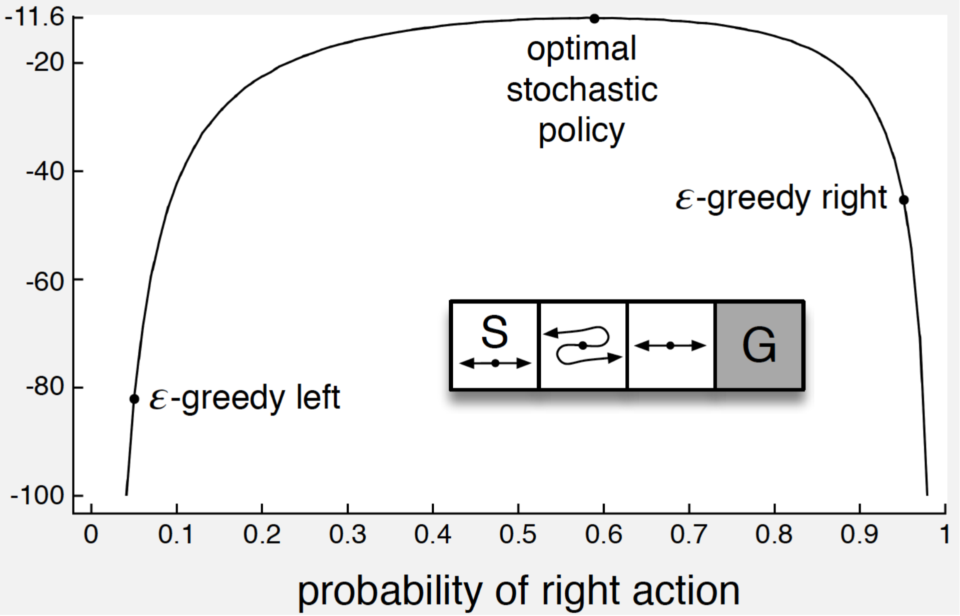
Value-based : $Q^*, V^\pi, V^* \approx \boxed{V_\theta}$

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

Limitation of Value Function Approximation



$$J(\boldsymbol{\theta}) = v_{\pi_{\boldsymbol{\theta}}}(\mathsf{S})$$



Idea 1: Exponential Weights

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

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

Perform incremental policy update such as

$$\pi_{k+1}(a|s) \propto \pi_k(a|s) \exp\left(\eta \hat{Q}_k(s,a)\right)$$

Idea 2: Policy Gradient

Parameterize policy by $\pi = \pi_\theta$

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

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

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

How are exponential weights and policy gradient related?

Policy Gradient in the Expert Setting

Policy Gradient for Softmax Policy in Expert Problem

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

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

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

Exponential weight

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

??

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

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

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

$\frac{\partial}{\partial \theta_i} \pi_\theta(a)$

when $a = i$: $\frac{\partial}{\partial \theta_i} \pi_\theta(a) = \frac{\partial}{\partial \theta(i)} \left[\frac{\exp(\theta(i))}{\sum_b \exp(\theta(b))} \right] = \frac{\exp(\theta(i)) \left(\sum_b \exp(\theta(b))\right) - \exp(\theta(i)) \cdot \exp(\theta(i))}{\left(\sum_b \exp(\theta(b))\right)^2}$

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

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

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

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

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

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

Policy Gradient over softmax policies

For $k = 1,2, \dots$

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

Exponential weights

For $k = 1,2, \dots$

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

Experiments

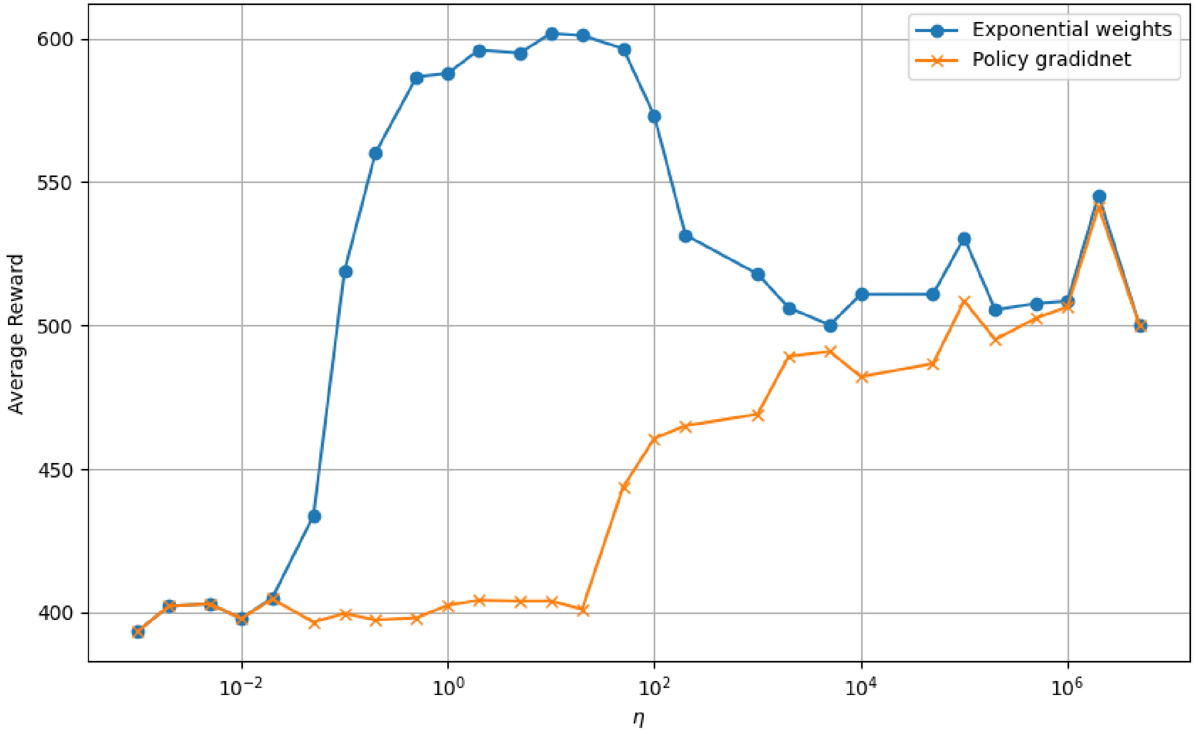
EW: $\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta A_{\theta_k}(a)$

PG: $\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta \pi_{\theta_k}(a) A_{\theta_k}(a)$

Reward = $[\mathrm{Ber}(0.6), \mathrm{Ber}(0.4)]$

Initial policy $\pi = [0.0001, 0.9999]$

Plot total reward in 1000 rounds



Two Ideas of Policy Updates

Policy Gradient over softmax policies

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

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

Exponential weights

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

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

Two Ideas for Function Approximation over Policies

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

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

(Vanilla) Policy Gradient

Natural Policy Gradient

Approximating the NPG Update

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

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

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

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

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)

NPG Updates

$$\boxed{\theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1} \left(\nabla_\theta V^{\pi_\theta} \Big|_{\theta=\theta_k} \right)}$$

cf. vanilla PG: $\theta_{k+1} = \theta_k + \eta \left(\nabla_\theta V^{\pi_\theta} \Big|_{\theta=\theta_k} \right)$

Summary: Policy Learning in the Expert Setting

PG	NPG
$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \left\langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}} \right\rangle - \frac{1}{2\eta} \lVert \theta - \theta_k \rVert^2$	$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \left\langle \pi_\theta - \pi_{\theta_k}, R \right\rangle - \frac{1}{\eta} \mathrm{KL}\left(\pi_\theta, \pi_{\theta_k}\right)$
$\theta_{k+1} = \theta_k + \eta \nabla_\theta V^{\pi_{\theta_k}}$	$\theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1} \nabla_\theta V^{\pi_{\theta_k}}$ where $F_\theta = \mathbb{E}_{a \sim \pi_\theta}[(\nabla_\theta \log \pi_\theta(a))(\nabla_\theta \log \pi_\theta(a))^\top]$
$\theta_{k+1}(a) = \theta_k(a) + \eta \pi_{\theta_k}(a) A_{\theta_k}(a)$ (under direct softmax parameterization)	$\theta_{k+1}(a) = \theta_k(a) + \eta A_{\theta_k}(a)$ (under direct softmax parameterization)

Policy Gradient with Bandit Feedback

Recall how we design the EXP3 algorithm

Full-information

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

$\longrightarrow$

Bandit

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

Inverse propensity weighting

$$\hat{r}_k(a) = \frac{r_k(a)\mathbb{I}\{a_k = a\}}{\pi_k(a)}$$

$$\hat{r}_k(a) = \frac{(r_k(a) - b - c(a))\mathbb{I}\{a_k = a\}}{\pi_k(a)} + c(a)$$

NPG (regularization form) + Bandit Feedback

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

Use π_{θ_k} to draw $a_{k1}, a_{k2}, \ldots, a_{kn}$, and get rewards $r_{k1}, r_{k2}, \ldots, r_{kn}$

Approximate $\quad R(a) \approx \sum_{i=1}^{n} \frac{(r_{ki} - b)\, \mathbb{I}\{a_{ki} = a\}}{\pi_{\theta_k}(a_{ki})}$ $\qquad$ ($n = 1$ recovers EXP3)

NPG (regularization form) + Bandit Feedback

For $k = 1, 2, \ldots$

Use π_{θ_k} to draw $a_{k1}, a_{k2}, \ldots, a_{kn}$, and get rewards $r_{k1}, r_{k2}, \ldots, r_{kn}$

Let $\hat{R}_k(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{(r_{ki} - b)\, \mathbb{I}\{a_{ki} = a\}}{\pi_{\theta_k}(a_{ki})}$

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

NPG (regularization form) + Bandit Feedback

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

Use π_{θ_k} to draw $a_{k1}, a_{k2}, \dots, a_{kn}$, and get rewards $r_{k1}, r_{k2}, \dots, r_{kn}$

Let $\quad \hat{R}_k(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{(r_{ki} - b)\ \mathbb{I}\{a_{ki} = a\}}{\pi_{\theta_k}(a_{ki})}$

$\theta \leftarrow \theta_k$

Repeat m times:

$$\theta \leftarrow \theta + \nabla_\theta \left(\langle \pi_\theta - \pi_{\theta_k}, \hat{R}_k \rangle - \frac{1}{\eta} \mathrm{KL}(\pi_\theta, \pi_{\theta_k}) \right)$$

$\theta_{k+1} \leftarrow \theta$

PG / NPG (Gradient-Update Form) + Bandit Feedback

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

PG

$$\theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1} \left(\nabla_\theta V^{\pi_\theta} \Big|_{\theta=\theta_k} \right)$$

NPG

PG + Bandit Feedback

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

Use π_{θ_k} to draw $a_{k1}, a_{k2}, \dots, a_{kn}$, and get rewards $r_{k1}, r_{k2}, \dots, r_{kn}$

Let $$g_k = \frac{1}{n} \sum_{i=1}^{n} (r_{ki} - b) \nabla_\theta \log \pi_\theta(a_{ki}) \Bigg|_{\theta=\theta_k}$$

$$\theta_{k+1} = \theta_k + \eta g_k$$

NPG (Gradient-Update Form) + Bandit Feedback

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

Use π_{θ_k} to draw $a_{k1}, a_{k2}, \ldots, a_{kn}$, and get rewards $r_{k1}, r_{k2}, \ldots, r_{kn}$

Let $$g_k = \frac{1}{n}\sum_{i=1}^{n}(r_{ki} - b)\nabla_\theta \log \pi_\theta(a_{ki})\Big|_{\theta=\theta_k}$$

$$\theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1} g_k$$

Summary: Policy Learning in Bandits

PG	NPG
$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}} \rangle - \frac{1}{2\eta} \lVert \theta - \theta_k \rVert^2$	$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \langle \pi_\theta - \pi_{\theta_k}, \boxed{R} \rangle - \frac{1}{\eta} \mathrm{KL}(\pi_\theta, \pi_{\theta_k})$
$\theta_{k+1} = \theta_k + \eta \boxed{\nabla_\theta V^{\pi_{\theta_k}}}$	$\theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1} \nabla_\theta V^{\pi_{\theta_k}}$ where $F_\theta = \mathbb{E}_{a \sim \pi_\theta}[(\nabla_\theta \log \pi_\theta(a))(\nabla_\theta \log \pi_\theta(a))^\top]$

$$\nabla_\theta V^{\pi_{\theta_k}} \approx \frac{1}{n} \sum_{i=1}^{n} (r_{ki} - b) \nabla_\theta \log \pi_\theta(a_{ki}) \bigg|_{\theta = \theta_k}$$

$$R(a) \approx \frac{1}{n} \sum_{i=1}^{n} \frac{(r_{ki} - b)\, \mathbb{I}\{a_{ki} = a\}}{\pi_{\theta_k}(a_{ki})}$$