Approximate Policy Iteration and Policy-Based Learning Methods

Chen-Yu Wei

Approximate Policy Iteration (API)

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

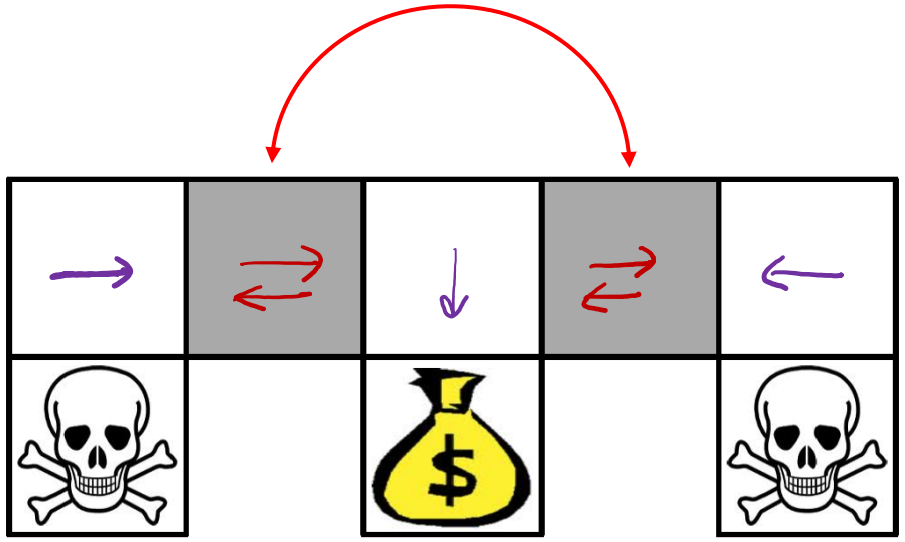
Evaluate $\hat{Q}_k \approx Q^{\pi_k}$

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

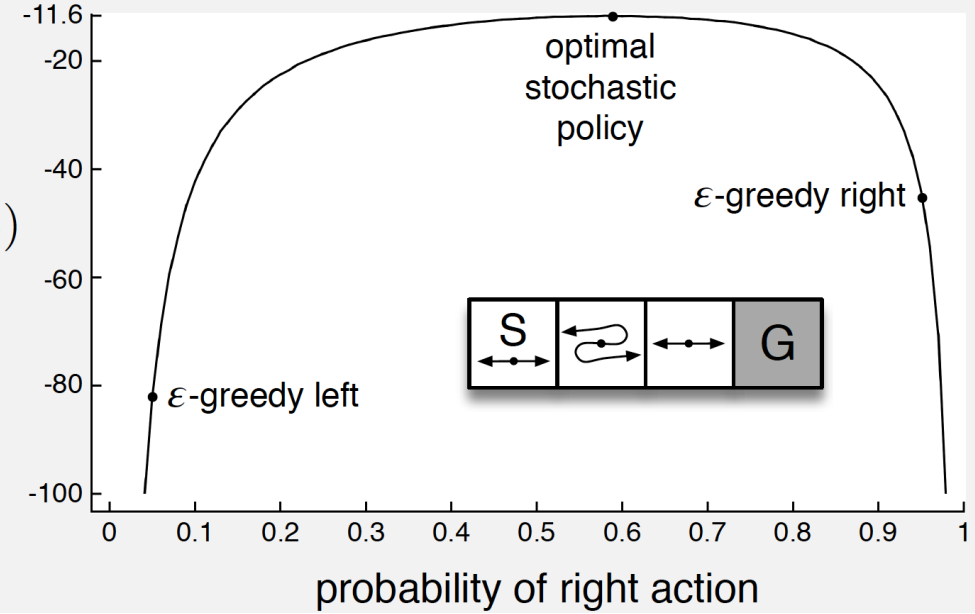
Value-based : $Q^*, V^\pi, V^* \approx V_\theta$

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

Limitation of Value Function Approximation



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



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 $k = 1,\ 2, \ldots$

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

$V^{\pi_\theta}(\rho) \triangleq \sum_s \rho(s) V^{\pi_\theta}(s)$

V^{π_θ}

How are exponential weights and policy gradient related?

Policy Learning in the Expert Setting

Policy Gradient for Softmax Policy in Expert Problem

Exponential weight

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

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

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

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

??

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

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

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

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

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

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

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

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

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

$$\pi_{k+1}(i) = \frac{\exp(\theta_{k+1}(i))}{\sum_b \exp(\theta_{k+1}(b))} = \frac{\exp(\theta_k(i))\exp(\eta\, \pi_{\theta_k}(i) A_{\theta_k}(i))}{\sum_b \exp(\theta_k(b))\exp(\eta\, \pi_{\theta_k}(b) A_{\theta_k}(b))} = \frac{\pi_k(i)\exp(\eta\, \pi_{\theta_k}(i) A_{\theta_k}(i))}{\sum_b \pi_k(b)\exp(\eta\, \pi_{\theta_k}(b) A_{\theta_k}(b))}$$

$$A_{\pi_k}(i) = R(i) - \sum_a \pi_k(a) R(a)$$ → constant for i

Exponential weights:

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

$$= \frac{\pi_k(i)\exp(\eta R(i) - c)}{\sum_b \pi_k(b)\exp(\eta R(b) - c)} \qquad \exp(-c)$$

PG over Softmax

$$\pi_{k+1}(i) = \frac{\pi_k(i)\exp\left(\eta \pi_k(i) A_{\pi_k}(i)\right)}{\sum_b \pi_k(b)\exp\left(\eta \pi_k(b) A_{\pi_k}(b)\right)}$$

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

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

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

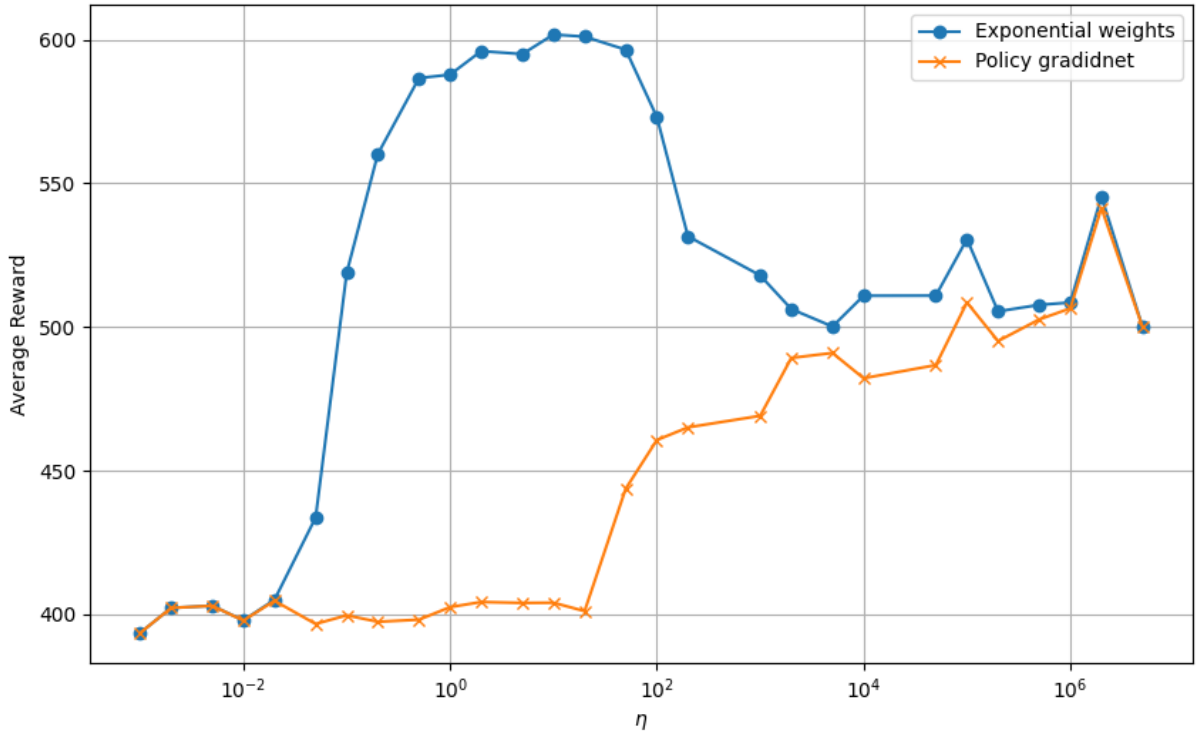
Plot total reward in 1000 rounds

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

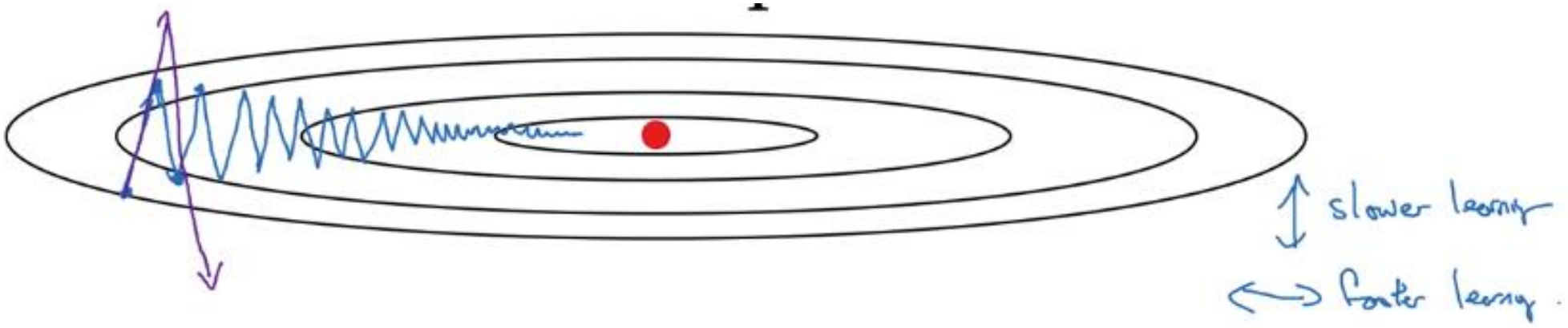
PG: $\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta \pi_{\theta_k}(a) A_{\theta_k}(a)$

small eta: too slow on action 1

larger eta: too fast on action 2

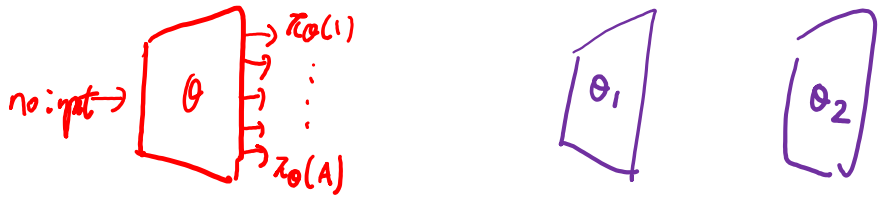


Optimization over ill-conditioned loss



https://math.stackexchange.com/questions/2285282/relating-condition-number-of-hessian-to-the-rate-of-convergence

Two Ideas of Policy Updates



Policy Gradient over softmax policies

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

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

$\nabla_\theta V^{\pi_\theta}\big|_{\theta=\theta_k} = \nabla_\theta V^{\pi_{\theta_k}}$

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

$\Leftrightarrow \theta_{k+1} = \operatorname{argmax}_\theta \{ \langle \theta, g_k \rangle - \frac{1}{2\eta} \|\theta - \theta_k\|^2 \}$

$= \operatorname{argmax}_\theta \{ \langle \theta - \theta_k, g_k \rangle - \frac{1}{2\eta} \|\theta - \theta_k\|^2 \}$

Exponential weights

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

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

$= \operatorname{argmax} \langle \pi_\theta - \pi_{\theta_k}, A_{\theta_k} \rangle - \frac{1}{\eta} KL(\pi_\theta, \pi_{\theta_k})$

$\sum_a (\pi_\theta(a) - \pi_{\theta_k}(a)) \cdot \text{cost} = 0$

$\tilde{R}(a) = R(a) - \text{const}$

$R(a) - \text{const}$

Two Ideas for Function Approximation over Policies

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

(Vanilla) Policy Gradient

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

Natural Policy Gradient

Approximating the NPG Update

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

$V^\pi = \langle \pi, R \rangle$
$= \sum_a \pi(a) R(a)$

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

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

$$\mathrm{KL}(\pi_\theta, \pi_{\theta_k}) \approx (\theta - \theta_k)^\top F_{\theta_k} (\theta - \theta_k) = \|\theta - \theta_k\|^2_{F_{\theta_k}}$$

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

(Fisher information matrix)

$$KL(\pi_\theta, \pi_{\theta+\Delta\theta}) \approx \frac{1}{2}(\Delta\theta)^T F_\theta (\Delta\theta) \quad \text{where} \quad F_\theta = \sum_a \pi_\theta(a) \left(\nabla_\theta \log \pi_\theta(a)\right)\left(\nabla_\theta \log \pi_\theta(a)\right)^T$$

$\Delta\theta \to 0$

$$KL(\pi_\theta, \pi_{\theta+\Delta\theta}) = \sum_a \pi_\theta(a) \ln \frac{\pi_\theta(a)}{\pi_{\theta+\Delta\theta}(a)}$$

$$= \sum_a \pi_\theta(a) \ln\left(\pi_\theta(a)\right) - \sum_a \pi_\theta(a) \ln\left(\pi_{\theta+\Delta\theta}(a)\right)$$

$$\approx \sum_a \pi_\theta(a) \ln\left(\pi_\theta(a)\right) - \sum_a \pi_\theta(a)\left(\ln \pi_\theta(a) + \nabla_\theta\left(\ln \pi_\theta(a)\right)^T \Delta\theta + \frac{1}{2}(\Delta\theta)^T \left(\nabla_\theta^2 \ln \pi_\theta(a)\right)\Delta\theta\right)$$

$$= -\sum_a \pi_\theta(a) \cdot \frac{\nabla \pi_\theta(a)^T \Delta\theta}{\pi_\theta(a)} - \sum_a \pi_\theta(a) \cdot \frac{1}{2}(\Delta\theta)^T \left(\frac{\left(\nabla^2 \pi_\theta(a)\right)\pi_\theta(a) - \left(\nabla \pi_\theta(a)\right)\left(\nabla \pi_\theta(a)\right)^T}{\left(\pi_\theta(a)\right)^2}\right)\Delta\theta$$

$$\nabla_\theta\left(\ln \pi_\theta(a)\right) = \frac{\nabla \pi_\theta(a)}{\pi_\theta(a)}$$

$$\nabla_\theta^2\left(\ln \pi_\theta(a)\right) = \frac{\left(\nabla^2 \pi_\theta(a)\right)\pi_\theta(a) - \left(\nabla \pi_\theta(a)\right)\left(\nabla \pi_\theta(a)\right)^T}{\left(\pi_\theta(a)\right)^2}$$

$$-\sum_a \nabla \pi_\theta(a)^T \Delta\theta = -\nabla\left(\sum_a \pi_\theta(a)\right)^T \Delta\theta = -\nabla(1)^T \Delta\theta = 0$$

$$\frac{1}{2}(\Delta\theta)^T \left(\nabla^2 \sum_a \pi_\theta(a)\right)\Delta\theta$$

For any θ, $\sum_a \pi_\theta(a) = 1$

$$f(\theta+\Delta\theta) \approx f(\theta) + \left(\nabla_\theta f(\theta)\right)^T \Delta\theta + \frac{1}{2}(\Delta\theta)^T \nabla_\theta^2 f(\theta)(\Delta\theta)$$

Hessian

NPG Updates

$$\frac{1}{2}\sum_a \pi_\theta(a)\,(\Delta\theta)^T\left(\frac{(\nabla_\theta \pi_\theta(a))(\nabla_\theta \pi_\theta(a))^T}{(\pi_\theta(a))^2}\right)\Delta\theta = \frac{1}{2}(\Delta\theta)^T F_\theta(\Delta\theta)$$

$$= (\nabla_\theta \log \pi_\theta(a))(\nabla_\theta \log \pi_\theta(a))^T$$

$$\nabla_\theta(\log \pi_\theta(a)) = \frac{\nabla_\theta \pi_\theta(a)}{\pi_\theta(a)}$$

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

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

NPG: $\theta_{k+1} = \underset{\theta}{\operatorname{argmax}}\left\{\sum_a\left(\pi_\theta(a) - \pi_{\theta_k}(a)\right)R(a) - \frac{1}{\eta}KL(\pi_\theta, \pi_{\theta_k})\right\}$

$\approx \underset{\theta}{\operatorname{argmax}}\left\{\langle \theta-\theta_k, \nabla_\theta V^{\pi_{\theta_k}}\rangle - \frac{1}{2\eta}(\theta-\theta_k)^T F_{\theta_k}(\theta-\theta_k)\right\} \to W(\theta)$

$\nabla_\theta W(\theta) = \nabla_\theta V^{\pi_{\theta_k}} - \frac{1}{\eta}F_{\theta_k}(\theta-\theta_k) = 0 \Rightarrow \theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1}\left(\nabla_\theta V^{\pi_{\theta_k}}\right)$

Summary: Policy Learning in the Expert Setting

PG	NPG
$\theta_{k+1} = \arg\max_{\theta} \langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}} \rangle - \frac{1}{2\eta} \|\theta - \theta_k\|^2$	$\theta_{k+1} = \arg\max_{\theta} \langle \pi_\theta - \pi_{\theta_k}, R \rangle - \frac{1}{\eta} \mathrm{KL}(\pi_\theta, \pi_{\theta_k})$
$\theta_{k+1} = \theta_k + \eta \nabla_\theta V^{\pi_{\theta_k}}$	$\theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1} \nabla_\theta V^{\pi_{\theta_k}}$ 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 Learning with Bandit Feedback

The design of EXP3

Full-information

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

→

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

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

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

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

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

NPG (regularization form) + Bandit Feedback

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

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

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

$\theta \leftarrow \theta_k$

Repeat m times:

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

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

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

$$F_{\theta_k} = \sum_a \pi_{\theta_k}(a) \left(\nabla_\theta \log \pi_\theta(a)\right)\left(\nabla_\theta \log \pi_\theta(a)\right)^T$$

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

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

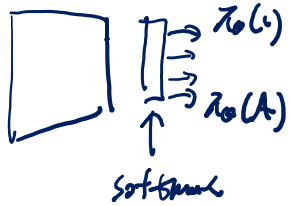
PG

$$\nabla_\theta V^{\pi_\theta} = \nabla_\theta \left[\sum_a \pi_\theta(a) R(a) \right] = \sum_a \left(\nabla_\theta \pi_\theta(a)\right) R(a) = \sum_a \pi_\theta(a) \cdot \frac{\nabla_\theta \pi_\theta(a)}{\pi_\theta(a)} R(a) = \mathbb{E}_{a_k \sim \pi_\theta} \left[\frac{\nabla_\theta \pi_\theta(a_k)}{\pi_\theta(a_k)} R(a_k) \right]$$

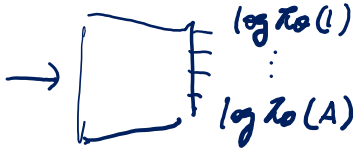
Suppose that $a_k \sim \pi_\theta$, get r_k

Define $g = \dfrac{\nabla_\theta \pi_\theta(a_k)}{\pi_\theta(a_k)} r_k \Rightarrow \mathbb{E}[g] = \sum_a \pi_\theta(a) \cdot \dfrac{\nabla_\theta \pi_\theta(a)}{\pi_\theta(a)} R(a) = \nabla_\theta V^{\pi_\theta}$

$$= \nabla_\theta \left(\log \pi_\theta(a_k)\right) r_k$$



weight $= \dfrac{r_k}{\pi_\theta(a_k)}$



weight $= r_k$

NPG

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

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

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

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

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

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

$$F_{\theta_k} = \sum_a \pi_\theta(a)\left(\nabla_\theta \log \pi_\theta(a)\right)\left(\nabla_\theta \log \pi_\theta(a)\right)^T$$

Summary: Policy Learning in Bandits

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

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

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

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

Policy Learning in MDPs

(Full-Information Case)

Exponential Weights

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

Perform individual exponential weight update **on all state s:**

$$\pi_{k+1}(a|s) = \frac{\pi_k(a|s)\exp(\eta Q^{\pi_k}(s,a))}{\sum_{a'}\pi_k(a'|s)\exp(\eta Q^{\pi_k}(s,a'))}$$

Analysis for Exponential Weights

$V^{\pi}(\rho) = \mathbb{E}_{s\sim\rho}\left[V^{\pi}(s)\right]$

Theorem.

The exponential weight algorithm guarantees

$$\sum_{k=1}^{K}\left(V^{\pi^\star}(\rho) - V^{\pi_k}(\rho)\right) \le \frac{1}{1-\gamma}\left(\frac{\ln A}{\eta} + \eta A K\right) = \frac{1}{(1-\gamma)^2}\sqrt{A(\ln A)K}$$

$V^{\pi_K}(\rho)$

for any initial state distribution ρ.

Remark. It is possible to show "last-iterate convergence"

$$V^{\pi^*}(\rho) - V^{\pi_k}(\rho) = \sum_s d_\rho^{\pi^*}(s) \sum_a \left(\pi^*(a|s) - \pi_k(a|s)\right) Q^{\pi_k}(s,a)$$

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

reward (a)

$$\sum_{k=1}^{K} \left(V^{\pi^*}(\rho) - V^{\pi_k}(\rho)\right) = \sum_s d_\rho^{\pi^*}(s) \sum_{k=1}^{K} \left[\sum_a \left(\pi^*(a|s) - \pi_k(a|s)\right) Q^{\pi_k}(s,a)\right]$$

$|Q^{\pi_k}(s,a)| \le \frac{1}{1-\gamma}$

$$\sum_{k=1}^{K} \sum_a \pi^*(a|s)\, Q^{\pi_k}(s,a) - \sum_{k=1}^{K} \sum_a \pi_k(a|s)\, Q^{\pi_k}(s,a)$$

$$\le \frac{1}{1-\gamma}\left(\frac{\ln A}{\eta} + \eta A K\right)$$

$$\le \left(\sum_s d_\rho^{\pi^*}(s)\right) \times \frac{1}{1-\gamma}\left(\frac{\ln A}{\eta} + 2AK\right) \le \frac{1}{(1-\gamma)^2}\left(\frac{\ln A}{\eta} + 2AK\right)$$

$\le \frac{1}{1-\gamma}$

Equivalent Forms of Exponential Weights

$$\forall s, \qquad \pi_{k+1}(\cdot\,|s) = \underset{\pi(\cdot|s)}{\operatorname{argmax}} \left\{ \sum_a \pi(a|s) Q^{\pi_k}(s,a) - \frac{1}{\eta} \mathrm{KL}(\pi(\cdot\,|s), \pi_k(\cdot\,|s)) \right\}$$

$$\sum_a \pi(a|s)(Q^{\pi_k}(s,a) - b(s)) \quad = \quad \square - b(s)$$

$V^{\pi_k}(s)$

$$\sum_a \pi(a|s) A^{\pi_k}(s,a)$$

$$\sum_a (\pi(a|s) - \pi_k(a|s)) Q^{\pi_k}(s,a)$$

...

$\theta : (\theta_{s,a})|_{s,a \in S \times A}$ $\pi_\theta : \pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$

Natural Policy Gradient (Regularization Form)

$$\theta_{k+1} = \operatorname*{argmax}_{\theta} \sum_{s} d_{\rho}^{\pi_{\theta_k}}(s) \left(\sum_{a} \pi_\theta(a|s) Q^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta} \mathrm{KL}(\pi_\theta(\cdot|s), \pi_{\theta_k}(\cdot|s)) \right)$$

Policy Gradient

$$V^{\pi_\theta}(\rho) = \mathbb{E}\left[\sum_{h=1}^{\infty} \gamma^{h-1} r_h \;\middle|\; s_1 \sim \rho,\ a_h \sim \pi_\theta(\cdot|s_h)\right]$$

$$\theta_{k+1} = \theta_k + \eta \nabla_\theta V^{\pi_\theta}(\rho)\Big|_{\theta=\theta_k} = \theta_k + \eta \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s)\Big|_{\theta=\theta_k}\right) Q^{\pi_{\theta_k}}(s,a)$$

$$V^{\pi_{\theta'}}(\rho) - V^{\pi_\theta}(\rho) = \sum_s d_\rho^{\pi_{\theta'}}(s) \sum_a \left(\pi_{\theta'}(a|s) - \pi_\theta(a|s)\right) Q^{\pi_\theta}(s,a)$$

$$\theta' = \theta + \Delta\theta \Rightarrow V^{\pi_{\theta+\Delta\theta}}(\rho) - V^{\pi_\theta}(\rho) = \sum_s d_\rho^{\pi_{\theta+\Delta\theta}}(s) \sum_a \left(\pi_{\theta+\Delta\theta}(a|s) - \pi_\theta(a|s)\right) Q^{\pi_\theta}(s,a)$$

divide both sides by $\Delta\theta$

$$\Rightarrow (\Delta\theta)^T \left(\nabla_\theta V^{\pi_\theta}(\rho)\right) = \sum_s d_\rho^{\pi_{\theta+\Delta\theta}}(s) \sum_a (\Delta\theta)^T \left(\nabla_\theta \pi_\theta(a|s)\right) Q^{\pi_\theta}(s,a)$$

$$\Rightarrow \nabla_\theta V^{\pi_\theta}(\rho) = \sum_s d_\rho^{\pi_\theta}(s) \sum_a \left(\nabla_\theta \pi_\theta(a|s)\right) \left(Q^{\pi_\theta}(s,a) - b(s)\right)$$

$$f(\theta+\Delta\theta) - f(\theta) \approx (\Delta\theta)^T \nabla_\theta f(\theta)$$

$$\Rightarrow \sum_a \left(\nabla_\theta \pi_\theta(a|s)\right) b(s) = \nabla_\theta \left(\sum_a \pi_\theta(a|s)\, b(s)\right)$$

$$\sum_a \pi_\theta(a|s) = 1 \quad = 0$$

Policy Gradient

$$\theta_{k+1} = \theta_k + \eta \nabla_\theta V^{\pi_\theta}(\rho)\Big|_{\theta=\theta_k} = \theta_k + \eta \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s)\Big|_{\theta=\theta_k} \right) Q^{\pi_{\theta_k}}(s,a)$$

Policy Gradient Theorem

$$\nabla_\theta V^{\pi_\theta}(\rho) = \sum_{s,a} d_\rho^{\pi_\theta}(s) \big(\nabla_\theta \pi_\theta(a|s) \big) Q^{\pi_\theta}(s,a)$$

Proof By the value difference lemma, for any $\Delta\theta = \epsilon \mathrm{e}_i$ we have

$$V^{\pi_{\theta+\epsilon \mathrm{e}_i}}(\rho) - V^{\pi_\theta}(\rho) = \sum_{s,a} d_\rho^{\pi_{\theta+\epsilon \mathrm{e}_i}}(s) \left(\pi_{\theta+\epsilon \mathrm{e}_i}(a|s) - \pi_\theta(a|s) \right) Q^{\pi_\theta}(s,a)$$

Dividing both sides by ϵ and taking $\epsilon \to 0$, we get

$$\frac{\partial}{\partial \theta_i} V^{\theta}(\rho) = \sum_{s,a} d_\rho^{\pi_\theta}(s) \left(\frac{\partial}{\partial \theta_i} \pi_\theta(a|s) \right) Q^{\pi_\theta}(s,a)$$

NPG vs. PG

Natural Policy Gradient

$$\theta_{k+1} = \operatorname*{argmax}_{\theta} \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\, \pi_\theta(a|s) Q^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta} \sum_{s} d_\rho^{\pi_{\theta_k}}(s)\, \mathrm{KL}\big(\pi_\theta(\cdot\,|s), \pi_{\theta_k}(\cdot\,|s)\big)$$

Policy Gradient

$$\theta_{k+1} = \theta_k + \eta \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s)\Big|_{\theta=\theta_k} \right) Q^{\pi_{\theta_k}}(s,a)$$

$$\theta_{k+1} = \operatorname*{argmax}_{\theta} \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\, (\theta - \theta_k)^\top \left(\nabla_\theta \pi_{\theta_k}(a|s) \right) Q^{\pi_{\theta_k}}(s,a) - \frac{1}{2\eta} \|\theta - \theta_k\|^2$$

NPG (Gradient-Update Form)

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

$$\theta_{k+1} = \operatorname*{argmax}_\theta \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\left(\pi_\theta(a|s) - \pi_{\theta_k}(a|s)\right)Q^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta}\sum_s d_\rho^{\pi_{\theta_k}}(s)\,\mathrm{KL}\left(\pi_\theta(\cdot\,|s), \pi_{\theta_k}(\cdot\,|s)\right)$$

$$\approx \operatorname*{argmax}_\theta \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\,(\theta-\theta_k)^\top\left(\nabla_\theta \pi_\theta(a|s)\Big|_{\theta=\theta_k}\right)Q^{\pi_{\theta_k}}(s,a) - \frac{1}{2\eta}\sum_s d_\rho^{\pi_{\theta_k}}(s)\,(\theta-\theta_k)^\top F_{\theta_k}(s)(\theta-\theta_k)$$

$$= \operatorname*{argmax}_\theta\ (\theta-\theta_k)^\top\left(\nabla_\theta V^{\pi_\theta}(\rho)\Big|_{\theta=\theta_k}\right) - \frac{1}{2\eta}(\theta-\theta_k)^\top F_{\theta_k}(\theta-\theta_k)$$

$$F_{\theta_k} = \sum_s d_\rho^{\pi_{\theta_k}}(s)\,F_{\theta_k}(s)$$

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

Summary: Full-Information Policy Learning in MDPs

Unified Idea:

$$\theta_{k+1} = \operatorname*{argmax}_{\theta} \left(V^{\pi_\theta}(\rho) - V^{\pi_{\theta_k}}(\rho) - \frac{1}{\eta} D(\theta, \theta_k) \right)$$

$$\approx \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\pi_\theta(a|s) - \pi_{\theta_k}(a|s)\right) Q^{\pi_{\theta_k}}(s,a) \quad - b(s)$$

Performance Difference Lemma $(\theta \to \theta_k)$

$$\approx (\theta - \theta_k)^\top \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s)\Big|_{\theta=\theta_k} \right) Q^{\pi_{\theta_k}}(s,a) \quad - b(s)$$

$$= (\theta - \theta_k)^\top \left(\nabla_\theta V^{\pi_\theta}(\rho)\Big|_{\theta=\theta_k} \right)$$

Summary: Full-Information Policy Learning in MDPs

PG	NPG
$\operatorname{argmax}_{\theta} \langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}}(\rho) \rangle - \frac{1}{2\eta} \lVert \theta - \theta_k \rVert^2$	$\operatorname{argmax}_{\theta} \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\, \pi_\theta(a\lvert s) Q^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta} \sum_{s} d_\rho^{\pi_{\theta_k}}(s)\, \mathrm{KL}\big(\pi_\theta(\cdot\lvert s), \pi_{\theta_k}(\cdot\lvert s)\big)$
$\theta_k + \eta \nabla_\theta V^{\pi_{\theta_k}}(\rho)$	$\theta_k + \eta F_{\theta_k}^{-1} \nabla_\theta V^{\pi_{\theta_k}}(\rho)$ where $F_\theta = \mathbb{E}_{s \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta} [(\nabla_\theta \log \pi_\theta(a\lvert s))(\nabla_\theta \log \pi_\theta(a\lvert s))^\top]$
	Tabular Case: $\pi_{k+1}(a\lvert s) = \frac{\pi_k(a\lvert s) \exp(\eta Q^{\pi_k}(s,a))}{\sum_{a'} \pi_k(a'\lvert s) \exp(\eta Q^{\pi_k}(s,a'))}$

Policy Learning in MDPs

(Bandit Feedback Case)

NPG (Regularization Form)

$$\theta_{k+1} = \operatorname*{argmax}_{\theta} \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s)\pi_{\theta}(a|s)\big(Q^{\pi_{\theta_k}}(s,a) - b(s)\big) - \frac{1}{\eta}\sum_{s} d_{\rho}^{\pi_{\theta_k}}(s)\mathrm{KL}\big(\pi_{\theta}(\cdot\,|s), \pi_{\theta_k}(\cdot\,|s)\big)$$

$$d_{\rho}^{\pi}(s) = \mathbb{E}\left[\sum_{h=1}^{\infty} \gamma^{h-1}\, \mathbb{I}\{s_h = s\} \;\middle|\; s_1 \sim \rho, a_h \sim \pi(\cdot\,|s_h)\right]$$

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{h=1}^{\infty} \gamma^{h-1}\, R(s_h, a_h) \;\middle|\; (s_1, a_1) = (s,a), a_h \sim \pi(\cdot\,|s_h) \quad \text{for } h \geq 2\right]$$

NPG (Regularization Form)

$$\theta_{k+1} = \operatorname{argmax}_{\theta} \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s)\pi_{\theta}(a|s)\big(Q^{\pi_{\theta_k}}(s,a) - b(s)\big) - \frac{1}{\eta}\sum_{s} d_{\rho}^{\pi_{\theta_k}}(s)\mathrm{KL}\big(\pi_{\theta}(\cdot|s), \pi_{\theta_k}(\cdot|s)\big)$$

For a fixed θ, an estimator for $\sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s)\pi_{\theta}(a|s)\big(Q^{\pi_{\theta_k}}(s,a) - b(s)\big)$ can be obtained as follows:

Sample a trajectory $(s_1 \sim \rho, a_1, r_1, s_2, a_2, r_2, \dots, s_\tau, a_\tau, r_\tau)$ using policy π_{θ_k}

Define $R_h = \sum_{i=h}^{\tau} \gamma^{i-h} r_i$

Define the estimator as $\sum_{h=1}^{\tau} \gamma^{h-1} \frac{\pi_{\theta}(a_h|s_h)}{\pi_{\theta_k}(a_h|s_h)}(R_h - b(s_h))$

Similarly, $\sum_s d_{\rho}^{\pi_{\theta_k}}(s)\mathrm{KL}\big(\pi_{\theta}(\cdot|s), \pi_{\theta_k}(\cdot|s)\big)$ can be estimated as $\sum_{h=1}^{\tau} \gamma^{h-1}\mathrm{KL}\big(\pi_{\theta}(\cdot|s_h), \pi_{\theta_k}(\cdot|s_h)\big)$

NPG (Regularization Form) + Bandit Feedback

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

$Q(s,a) \le \frac{1}{1-\gamma}$

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \cdots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \ldots\ldots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \cdots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{\tau_n} \gamma^{h-1} \frac{\pi_\theta\left(a_h^{(i)} \middle| s_h^{(i)}\right)}{\pi_{\theta_k}\left(a_h^{(i)} \middle| s_h^{(i)}\right)} \left(R_h^{(i)} - b\left(s_h^{(i)}\right) \right) \right.$$

$$\left. - \frac{1}{\eta} \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{\tau_n} \gamma^{h-1} \mathrm{KL}\left(\pi_\theta\left(\cdot \middle| s_h^{(i)}\right), \pi_{\theta_k}\left(\cdot \middle| s_h^{(i)}\right)\right) \right\}$$

$R_h^{(i)} \le \frac{1}{1-\gamma}$

Practical version will not include the discount factor at the front

NPG (Regularization Form) + Bandit Feedback

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

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \cdots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \ldots \ldots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \cdots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

$$\theta_{k+1} = \underset{\theta}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{\tau_n} \frac{\pi_\theta\left(a_h^{(i)} \middle| s_h^{(i)}\right)}{\pi_{\theta_k}\left(a_h^{(i)} \middle| s_h^{(i)}\right)} \left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right) \right.$$

$$\left. - \frac{1}{\eta} \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{\tau_n} \mathrm{KL}\left(\pi_\theta\left(\cdot \middle| s_h^{(i)}\right), \pi_{\theta_k}\left(\cdot \middle| s_h^{(i)}\right)\right) \right\}$$

NPG (Regularization Form) + Bandit Feedback

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

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \cdots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \ldots \ldots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \cdots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

Let $W_k(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \frac{\pi_\theta\left(a_h^{(i)}\middle|s_h^{(i)}\right)}{\pi_{\theta_k}\left(a_h^{(i)}\middle|s_h^{(i)}\right)}\left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right) - \frac{1}{\eta}\frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \mathrm{KL}\left(\pi_\theta\left(\cdot\middle|s_h^{(i)}\right), \pi_{\theta_k}\left(\cdot\middle|s_h^{(i)}\right)\right)$

$\theta \leftarrow \theta_k$

Repeat m times:

$$\theta \leftarrow \theta + \alpha\nabla_\theta W_k(\theta)$$

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

PG

$$\theta_{k+1} = \theta_k + \eta \nabla_\theta V^{\pi_\theta}(\rho)\Big|_{\theta=\theta_k} = \theta_k + \eta \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s)\Big|_{\theta=\theta_k} \right) \left(Q^{\pi_{\theta_k}}(s,a) - b(s) \right)$$

$$\sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\, \pi_{\theta_k}(a|s) \left(\frac{\nabla_\theta \pi_\theta(a|s)\big|_{\theta=\theta_k}}{\pi_{\theta_k}(a|s)} \right) \left(Q^{\pi_{\theta_k}}(s,a) - b(s) \right)$$

PG + Bandit Feedback (REINFORCE)

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

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \cdots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \ldots \ldots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \cdots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

Define

$$g = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \frac{\nabla_\theta \pi_\theta\left(a_h^{(i)} \middle| s_h^{(i)}\right)\Big|_{\theta=\theta_k}}{\pi_{\theta_k}\left(a_h^{(i)} \middle| s_h^{(i)}\right)}\left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right)$$

Perform update

$$\theta_{k+1} \leftarrow \theta_k + \eta g$$

PG + Bandit Feedback (REINFORCE)

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

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \cdots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \ldots \ldots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \cdots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

Define

$$g = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \nabla_\theta \log \pi_\theta \left(a_h^{(i)} \middle| s_h^{(i)}\right)\Big|_{\theta=\theta_k} \left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right)$$

Perform update

$$\theta_{k+1} \leftarrow \theta_k + \eta g$$

NPG (Gradient-Update Form) + Bandit Feedback

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

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \cdots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \ldots\ldots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \cdots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

Define

$$g = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \nabla_\theta \log \pi_\theta\left(a_h^{(i)}\middle|s_h^{(i)}\right)\Big|_{\theta=\theta_k} \left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right)$$

Perform update

$$\theta_{k+1} \leftarrow \theta_k + \eta F_{\theta_k}^{-1} g$$

$$F_{\theta_k} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \left(\nabla_\theta \log \pi_{\theta_k}\left(a_h^{(i)}\middle|s_h^{(i)}\right)\right)\left(\nabla_\theta \log \pi_{\theta_k}\left(a_h^{(i)}\middle|s_h^{(i)}\right)\right)^\top$$

$d \times d$

$$\left(\sum_i x_i x_i^\top\right)^{-1}\left(\sum_i x_i y_i\right) = \text{the solution of } \operatorname*{argmin}_w \sum_i \left(x_i^\top w - y_i\right)^2$$

$$F_{\theta_k}^{-1} g = \operatorname*{argmin}_w \sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \left(\nabla_\theta \log \pi_{\theta_k}\left(a_h^{(i)}\middle|s_h^{(i)}\right)^\top w - \left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right)\right)^2$$

Summary: Policy Learning in MDPs

PG	NPG
$\operatorname{argmax}_{\theta} \langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}}(\rho) \rangle - \frac{1}{2\eta} \|\theta - \theta_k\|^2$	$\operatorname{argmax}_{\theta} \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\, \pi_\theta(a\|s) Q^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta} \sum_s d_\rho^{\pi_{\theta_k}}(s)\, \mathrm{KL}\big(\pi_\theta(\cdot\|s), \pi_{\theta_k}(\cdot\|s)\big)$
$\theta_k + \eta \nabla_\theta V^{\pi_{\theta_k}}(\rho)$	$\theta_k + \eta F_{\theta_k}^{-1} \nabla_\theta V^{\pi_{\theta_k}}(\rho)$ where $F_\theta = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta}[(\nabla_\theta \log \pi_\theta(a\|s))(\nabla_\theta \log \pi_\theta(a\|s))^\top]$

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{\tau_n} \frac{\nabla_\theta \pi_\theta \left(a_h^{(i)} \middle| s_h^{(i)}\right)\Big|_{\theta=\theta_k}}{\pi_{\theta_k}\left(a_h^{(i)} \middle| s_h^{(i)}\right)} \left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{\tau_n} \nabla_\theta \log \pi_\theta \left(a_h^{(i)} \middle| s_h^{(i)}\right)\Big|_{\theta=\theta_k} \left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right)$$

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{\tau_n} \frac{\pi_\theta \left(a_h^{(i)} \middle| s_h^{(i)}\right)}{\pi_{\theta_k}\left(a_h^{(i)} \middle| s_h^{(i)}\right)} \left(R_h^{(i)} - b\left(s_h^{(i)}\right)\right)$$

Actor-Critic Methods

Review: Full-Information Policy Learning in MDPs

$$\theta_{k+1} = \operatorname*{argmax}_{\theta} \left(V^{\pi_\theta}(\rho) - V^{\pi_{\theta_k}}(\rho) - \frac{1}{\eta} D(\theta, \theta_k) \right)$$

$$\approx \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\left(\pi_\theta(a|s) - \pi_{\theta_k}(a|s)\right) Q^{\pi_{\theta_k}}(s,a) = \mathbb{E}_{(s_i,a_i)}\left[\frac{\pi_\theta(a_i|s_i) - \pi_{\theta_k}(a_i|s_i)}{\pi_{\theta_k}(a_i|s_i)} Q^{\pi_{\theta_k}}(s_i,a_i)\right]$$

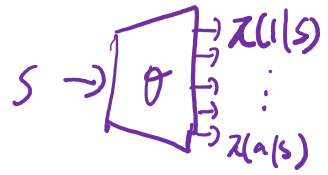
$$\approx (\theta - \theta_k)^\top \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s)\Big|_{\theta=\theta_k} \right) Q^{\pi_{\theta_k}}(s,a)$$

$$= \mathbb{E}_{(s_i,a_i)}\left[\frac{\nabla_\theta \pi_\theta(a_i|s_i)|_{\theta=\theta_k}}{\pi_{\theta_k}(a_i|s_i)} Q^{\pi_{\theta_k}}(s_i,a_i)\right]$$

PG/NPG: Estimate them using the empirical sum of reward in the trajectory (i.e., Monte Carlo estimator)

We can also use other estimators to balance bias and variance

Actor-Critic Methods



Use value function approximation to estimate $Q^{\pi_{\theta_k}}(s_i, a_i)$ or $A^{\pi_{\theta_k}}(s_i, a_i)$

Use $Q_\phi(s,a)$: $\min_\phi \mathbb{E}_{(s,a,r,s',a')\sim\pi_{\theta_k}}\left[\left(Q_\phi(s,a) - r - \gamma Q_{\phi_k}(s',a')\right)^2\right]$

e.g. $Q^{\pi_{\theta_k}}(s,a) \leftarrow Q_\phi(s,a)$

$A^{\pi_{\theta_k}}(s,a) \leftarrow Q_\phi(s,a) - \mathbb{E}_{a'\sim\pi_{\theta_k}(\cdot|s)}\left[Q_\phi(s,a')\right]$

Use $V_\phi(s)$: $\min_\phi \mathbb{E}_{(s,r,s')\sim\pi_{\theta_k}}\left[\left(V_\phi(s) - r - \gamma V_{\phi_k}(s')\right)^2\right]$

e.g. $Q^{\pi_{\theta_k}}(s,a) \leftarrow r + \gamma V_\phi(s')$

$A^{\pi_{\theta_k}}(s,a) \leftarrow r + \gamma V_\phi(s') - V_\phi(s)$

or $r_1 + \gamma r_2 + \gamma^2 V_\phi(s') - V_\phi(s)$

etc.

Actor-Critic with Q-function Approximator

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

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \cdots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \ldots\ldots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \cdots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

Define

$$g = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \frac{\nabla_\theta\, \pi_\theta\left(a_h^{(i)}\middle|s_h^{(i)}\right)\Big|_{\theta=\theta_k}}{\pi_{\theta_k}\left(a_h^{(i)}\middle|s_h^{(i)}\right)} Q_{\phi_k}\left(s_h^{(i)}, a_h^{(i)}\right) \text{ or } \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n}\sum_{a} \nabla_\theta\, \pi_\theta\left(a\middle|s_h^{(i)}\right)\Big|_{\theta=\theta_k} Q_{\phi_k}\left(s_h^{(i)}, a\right)$$

Perform updates

$$\theta_{k+1} \leftarrow \theta_k + \eta g \qquad \phi_{k+1} \leftarrow \phi_k - \lambda \nabla_\phi \sum_{i=1}^{n}\sum_{h=1}^{\tau_n}\left(Q_\phi\left(s_h^{(i)}, a_h^{(i)}\right) - r_h^{(i)} - \gamma Q_{\phi_k}\left(s_{h+1}^{(i)}, a_{h+1}^{(i)}\right)\right)^2\Big|_{\phi=\phi_k}$$

Advantage Actor-Critic (A2C)

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

Use π_{θ_k} to collect n trajectories

$$\left(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \cdots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}\right), \ldots\ldots, \left(s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \cdots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)}\right)$$

Define

$$g = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \nabla_\theta \log \pi_\theta\left(a_h^{(i)} \middle| s_h^{(i)}\right)\Big|_{\theta=\theta_k} \left(r_h^{(i)} + \gamma V_{\phi_k}\left(s_{h+1}^{(i)}\right) - V_{\phi_k}\left(s_h^{(i)}\right)\right)$$

$\approx Q^{\pi_{\theta_k}}(s_h^{(i)}, a_h^{(i)})$

Perform updates

$$\theta_{k+1} \leftarrow \theta_k + \eta g \qquad \phi_{k+1} \leftarrow \phi_k - \lambda \nabla_\phi \sum_{i=1}^{n}\sum_{h=1}^{\tau_n} \left(V_\phi\left(s_h^{(i)}\right) - r_h^{(i)} - \gamma V_{\phi_k}\left(s_{h+1}^{(i)}\right)\right)^2\Big|_{\phi=\phi_k}$$