

Approximate Policy Iteration and Policy-Based Learning Methods

Chen-Yu Wei

Approximate Policy Iteration (API)

For $k = 1, 2, \dots$

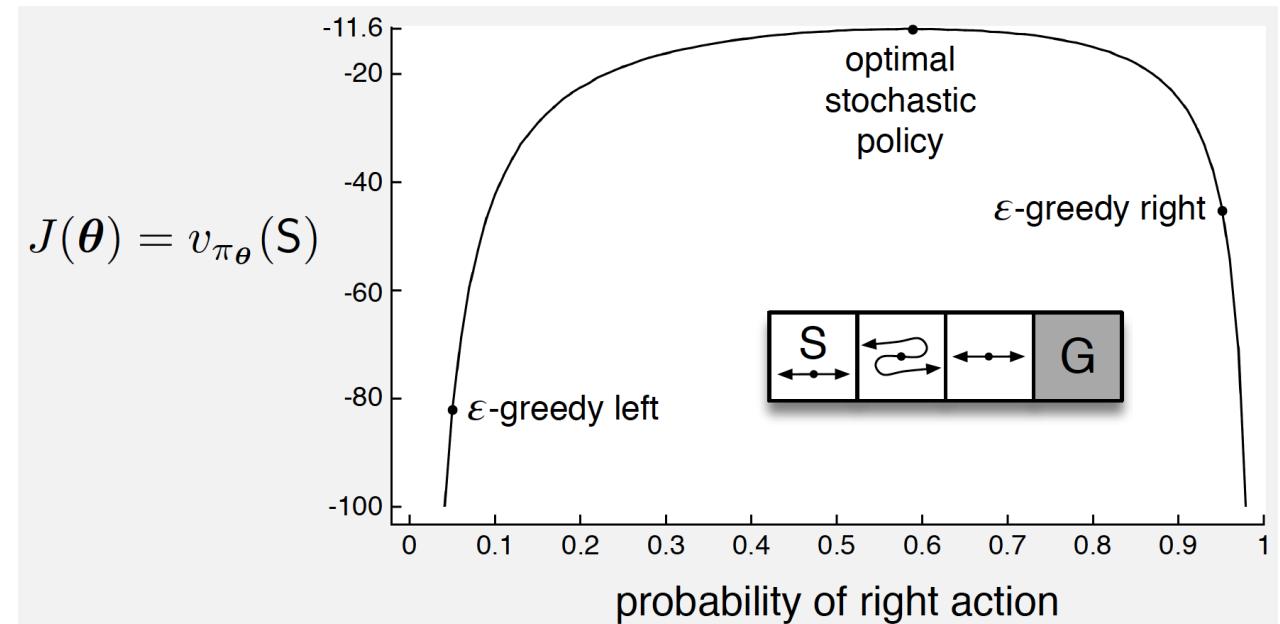
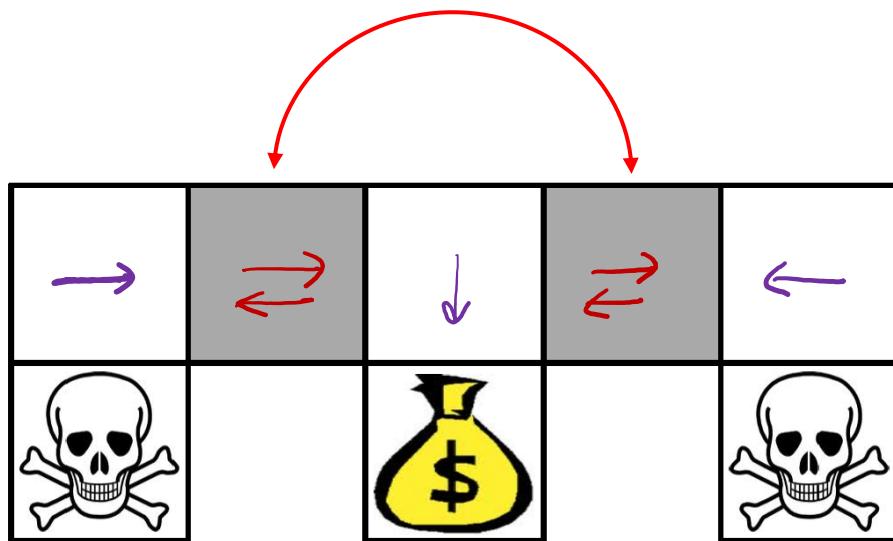
Evaluate $\hat{Q}_k \approx Q^{\pi_k}$

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

Value-based : $\overset{Q^{\pi}}{\underset{Q}{\approx}} V^*, v^* \approx \boxed{V_0}$

Policy-based : $\underline{\pi_0(a|s)}$

Limitation of Value Function Approximation



Idea 1: Exponential Weights

For $k = 1, 2, \dots$

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

Perform incremental policy update such as

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

Idea 2: Policy Gradient

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

For $k = 1, 2, \dots$

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

$$V^{\pi_\theta}(\rho) \stackrel{\Delta}{=} \underbrace{\sum_s \rho(s) V^{\pi_\theta}(s)}_{V^{\pi_\theta}}$$

How are exponential weights and policy gradient related?

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

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

Exponentiation weight

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

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

$$(\nabla_{\theta} V^{\pi_{\theta}})_i = \sum_a \frac{\partial}{\partial \theta_i} (\pi_{\theta}(a)) \cancel{R(a)} =$$

$\left. \frac{\partial}{\partial \theta_i} \pi_{\theta}(a) \right|_{a=i}$

$$\text{when } a=i : \frac{\partial}{\partial \theta_i} \pi_{\theta}(a) =$$

$$\text{PG: } \underline{\theta}_{k+1} = \underline{\theta}_k + \gamma \nabla_{\theta} V^{\pi_{\theta}} \Big|_{\theta=\underline{\theta}_k}$$

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

$$= \frac{\exp(\theta(i)) \left(\sum_b \exp(\theta(b)) \right) - \exp(\theta(i)) \cdot \exp(\theta(i))}{\left(\sum_b \exp(\theta(b)) \right)^2}$$

$$\text{when } a \neq i : \frac{\partial}{\partial \theta_i} \pi_{\theta}(a) =$$

$$= \frac{\frac{\partial}{\partial \theta_i} \left(\frac{\exp(\theta(a))}{\sum_b \exp(\theta(b))} \right)}{\left(\sum_b \exp(\theta(b)) \right)^2}$$

$$\begin{aligned}
 \underline{\pi_{\theta}^{(k)}}_i &= \frac{\exp(\theta(i)) R(i)}{\sum_b \exp(\theta(b))} - \sum_a \frac{\exp(\theta(a)) \exp(\theta(i)) R(a)}{\left(\sum_b \exp(\theta(b))\right)^2} \\
 &= \frac{\exp(\theta(i))}{\sum_b \exp(\theta(b))} \left(R(i) - \sum_a \frac{\exp(\theta(a))}{\sum_b \exp(\theta(b))} R(a) \right) \\
 &= \pi_{\theta}(i) \left(R(i) - \sum_a \pi_{\theta}(a) R(a) \right)
 \end{aligned}$$

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

$$\pi_{k+1}(i) = \frac{\exp(\theta_{k+1}(i))}{\sum_b \exp(\theta_{k+1}(b))} = \frac{\underbrace{\exp(\theta_k(i))}_{\text{circled}} \exp(\gamma \pi_{\theta_k}(i) A_{\theta_k}(i))}{\sum_b \exp(\theta_k(b)) \exp(\gamma \pi_{\theta_k}(b) A_{\theta_k}(b))} = \frac{\pi_k(i) \exp(\gamma \pi_{\theta_k}(i) A_{\theta_k}(i))}{\sum_b \pi_k(b) \exp(\gamma \pi_{\theta_k}(b) A_{\theta_k}(b))}$$

Exponential weights :

$$A_{\pi_k}(i) = R(i) \left(- \sum_a \pi_k(a) R(a) \right) \rightarrow \text{constant for } i$$

$$\begin{aligned} \pi_{k+1}(i) &= \frac{\pi_k(i) \exp(\gamma R(i))}{\sum_b \pi_k(b) \exp(\gamma R(b))} \stackrel{?}{=} \frac{\pi_k(i) \exp(\gamma A_{\pi_k}(i))}{\sum_b \pi_k(b) \exp(\gamma A_{\pi_k}(b))} \\ &\quad \parallel \\ &\quad \frac{\pi_k(i) \exp(\gamma R(i) - c)}{\sum_b \pi_k(b) \exp(\gamma R(b) - c)} \quad \underline{\exp(-c)} \end{aligned}$$

PG over softmax

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

Comparison between EW and PG over softmax policies

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

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 = [Ber(0.6), Ber(0.4)]

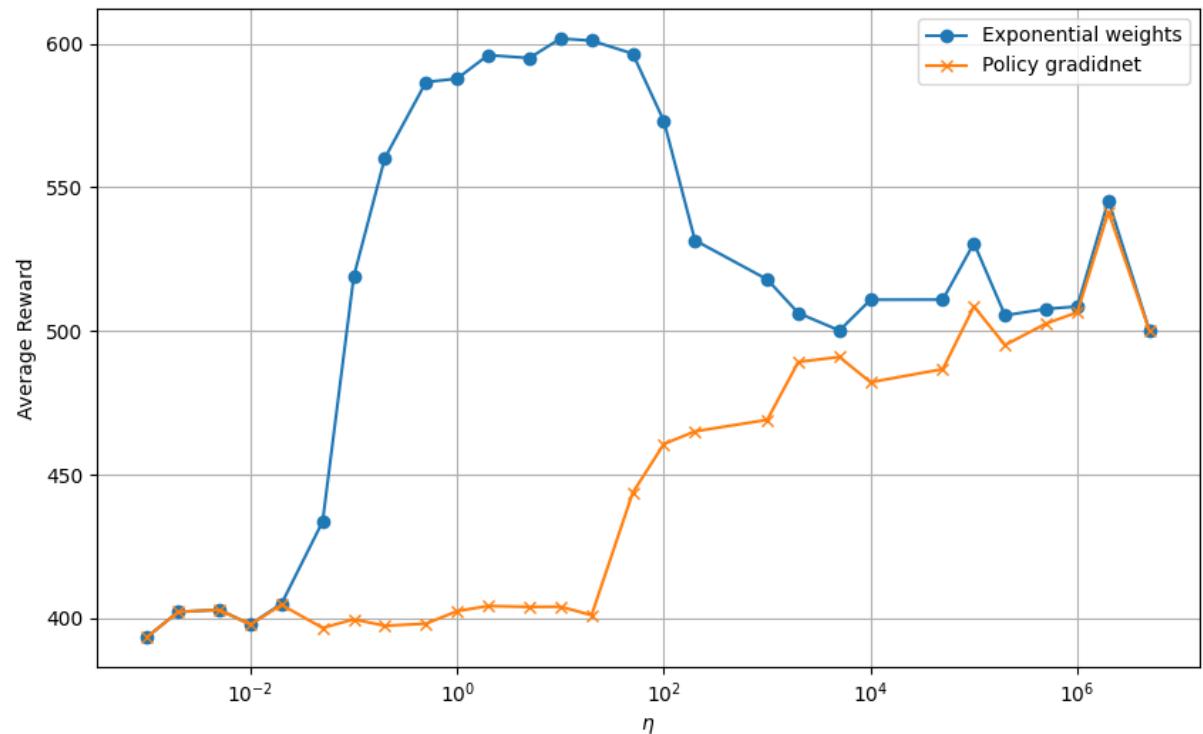
Initial policy $\pi = [0.0001, 0.9999]$

Plot total reward in 1000 rounds

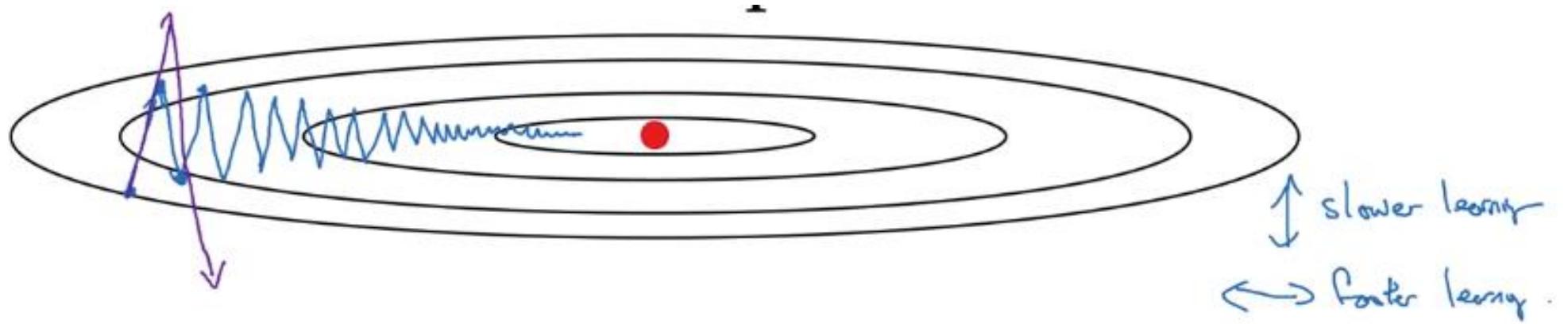
$$\text{EW: } \theta_{k+1}(a) \leftarrow \theta_k(a) + \eta A_{\theta_k}(a)$$

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

small η : too slow on action 1
larger η : 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

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



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

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

Exponential weights

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



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

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

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

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

$$\stackrel{\Leftarrow}{=} \operatorname{argmax}_{\theta} \left\langle \pi_{\theta} - \pi_{\theta_k}, A_{\theta_k} \right\rangle - \frac{1}{\eta} \text{KL}(\pi_{\theta}, \pi_{\theta_k})$$

$\sum_a (\pi_{\theta}(a) - \pi_{\theta_k}(a)) R(a) = 0$
 $R(a) = R(a) - \underline{\text{const}}$

Two Ideas for Function Approximation over Policies

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

(Vanilla) Policy Gradient

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

Natural Policy Gradient

Approximating the NPG Update

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

$$\begin{aligned} V^{\pi} &= \langle \pi, R \rangle \\ &= \sum_a \pi(a) R(a) \end{aligned}$$

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

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

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

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

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

\downarrow
 $\Delta\theta \rightarrow 0$

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

$$f(\theta + \Delta\theta) \approx f(\theta) + (\nabla_\theta f(\theta))^T \Delta\theta + \frac{1}{2} (\Delta\theta)^T \underbrace{\nabla_\theta^2 f(\theta)}_{\text{Hessian}} (\Delta\theta)$$

$$\begin{aligned} &= \sum_a \pi_\theta(a) \ln(\pi_\theta(a)) - \sum_a \pi_\theta(a) \ln(\pi_{\theta+\Delta\theta}(a)) \\ &\approx \sum_a \pi_\theta(a) \cancel{\ln(\pi_\theta(a))} - \sum_a \pi_\theta(a) \left(\cancel{\ln \pi_\theta(a)} + \cancel{\nabla_\theta(\ln \pi_\theta(a))^T \Delta\theta} + \frac{1}{2} (\Delta\theta)^T \cancel{\nabla_\theta^2 \ln \pi_\theta(a) \Delta\theta} \right) \end{aligned}$$

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

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

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

$$\begin{aligned} &\downarrow \\ &- \sum_a \nabla \pi_\theta(a)^T \Delta\theta \\ &= - \nabla \left(\sum_a \pi_\theta(a) \right)^T \Delta\theta \\ &= 0 \end{aligned}$$

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

For any θ . $\sum_a \pi_\theta(a) = 1$

NPG Updates

$$\frac{1}{2} \sum_a \bar{\pi}_\theta(a) (\Delta \theta)^T \left(\frac{(\nabla_\theta \bar{\pi}_\theta(a)) (\nabla_\theta \bar{\pi}_\theta(a))^T}{(\bar{\pi}_\theta(a))^2} \right) \Delta \theta = \frac{1}{2} (\Delta \theta)^T \bar{F}_\theta(\Delta \theta)$$

$$= (\nabla_\theta \log \bar{\pi}_\theta(a)) (\nabla_\theta \log \bar{\pi}_\theta(a))^T$$

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

$$\nabla_\theta (\log \bar{\pi}_\theta(a)) = \frac{\nabla_\theta \bar{\pi}_\theta(a)}{\bar{\pi}_\theta(a)}$$

cf. vanilla PG:

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

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

$$\approx \underset{\theta}{\operatorname{argmax}} \left\{ \langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}} \rangle - \frac{1}{2\gamma} (\theta - \theta_k)^T \bar{F}_{\theta_k} (\theta - \theta_k) \right\} \rightarrow W(\theta)$$

$$\nabla_\theta W(\theta) = \nabla_\theta V^{\pi_{\theta_k}} - \frac{1}{\gamma} \bar{F}_{\theta_k} (\theta - \theta_k) = 0 \Rightarrow \theta_{k+1} = \theta_k + \eta \bar{F}_{\theta_k}^{-1} (\nabla_\theta V^{\pi_{\theta_k}})$$

Summary: Policy Learning in the Expert Setting

$$V^2 = \sum_a \pi(a) R(a)$$

PG	NPG
$\theta_{k+1} = \operatorname{argmax}_{\theta} \langle \theta - \theta_k, \nabla_{\theta} V^{\pi_{\theta_k}} \rangle - \frac{1}{2\eta} \ \theta - \theta_k\ ^2$	$\theta_{k+1} = \operatorname{argmax}_{\theta} \langle \pi_{\theta} - \pi_{\theta_k}, R \rangle - \frac{1}{\eta} \text{KL}(\pi_{\theta}, \pi_{\theta_k})$ <p style="margin-left: 100px;"> $V^{\pi_{\theta_k}}$ - $V^{\pi_{\theta}}$ $\approx \frac{1}{2} \ \theta - \theta_k\ _F^2$ </p>
$\theta_{k+1} = \theta_k + \eta \nabla_{\theta} V^{\pi_{\theta_k}}$	$\theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1} \nabla_{\theta} V^{\pi_{\theta_k}}$ <p>where $F_{\theta} = \mathbb{E}_{a \sim \pi_{\theta}} [(\nabla_{\theta} \log \pi_{\theta}(a)) (\nabla_{\theta} \log \pi_{\theta}(a))^T]$</p>
$\theta_{k+1}(a) = \theta_k(a) + \eta \pi_{\theta_k}(a) A_{\theta_k}(a)$ <p>(under direct softmax parameterization)</p>	$\theta_{k+1}(a) = \theta_k(a) + \eta A_{\theta_k}(a)$ <p>(under direct softmax parameterization)</p>



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

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

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

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

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 $r_{k1}, r_{k2}, \dots, r_{kn}$

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

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

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 $r_{k1}, r_{k2}, \dots, r_{kn}$

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

$$\theta \leftarrow \theta_k$$

Repeat m times:

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

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

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

$$\bar{F}_{\theta_k} = \sum_a \pi_{\theta_k}(a) (\nabla_{\theta} \log \pi_{\theta}(a)) (\nabla_{\theta} \log \pi_{\theta}(a))^T$$

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

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

PG

$$\nabla_{\theta} V^{\pi_{\theta}} = \nabla_{\theta} \left[\sum_a \pi_{\theta}(a) R(a) \right] = \sum_a (\underbrace{\nabla_{\theta} \pi_{\theta}(a)}_{\pi_{\theta}(a)}) R(a)$$

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

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

Softmax

$$\text{weight} = \frac{r_k}{\pi_{\theta}(a_k)} = \frac{\nabla_{\theta}(\log \pi_{\theta}(a_k))}{\pi_{\theta}(a_k)} r_k$$

NPG

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

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

weight = r_k

PG + Bandit Feedback

For $k = 1, 2, \dots$

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

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

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

PG + Bandit Feedback

For $k = 1, 2, \dots$

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

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

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

NPG (Gradient-Update Form) + Bandit Feedback

For $k = 1, 2, \dots$

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

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

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

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

Summary: Policy Learning in Bandits

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

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

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

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

Policy Learning in MDPs

(Full-Information Case)

Exponential Weights

For $k = 1, 2, \dots$

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

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

Theorem.

The exponential weight algorithm guarantees

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

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

for any initial state distribution ρ .

Remark. It is possible to show “last-iterate convergence”

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

$\pi_{\text{real}}(a|s) \propto \pi_k(a|s) \exp(\eta \underline{Q^{\pi_k}(s, a)})$
reward(a)

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

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

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

$$\leq \frac{1}{1-\gamma} \left(\frac{\ln A}{\eta} + 2AK \right)$$

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

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

Equivalent Forms of Exponential Weights

$$\forall s, \quad \pi_{k+1}(\cdot | s) = \operatorname{argmax}_{\pi(\cdot | s)} \left\{ \underbrace{\sum_a \left(\pi(a | s) Q^{\pi_k}(s, a) \right)}_{\text{blue bracket}} - \frac{1}{\eta} \text{KL}(\pi(\cdot | s), \pi_k(\cdot | s)) \right\}$$

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

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

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

...

$$\theta: (\theta_{s,a}) \Big|_{s,a \in S \times A}$$

$$\pi_\theta : \left(\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_a \exp(\theta_{s,a})} \right)$$

Natural Policy Gradient (Regularization Form)

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

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

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

Proof By the value difference lemma, for any $\Delta\theta = \epsilon e_i$ we have

$$V^{\pi_{\theta+\epsilon e_i}}(\rho) - V^{\pi_\theta}(\rho) = \sum_{s,a} d_\rho^{\pi_{\theta+\epsilon e_i}}(s) (\pi_{\theta+\epsilon e_i}(a|s) - \pi_\theta(a|s)) Q^{\pi_\theta}(s, a)$$

Dividing both sides by ϵ and taking $\epsilon \rightarrow 0$, 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

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

Policy Gradient

$$\begin{aligned} \theta_{k+1} &= \theta_k + \gamma g \\ \Leftrightarrow \theta_{k+1} &= \operatorname{argmax}_{\theta} (\theta - \theta_k)^T g - \frac{1}{2\eta} \|\theta - \theta_k\|^2 \end{aligned}$$

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

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

NPG (Gradient-Update Form)

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

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

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

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

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

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

Summary: Full-Information Policy Learning in MDPs

Unified Idea:

$$\begin{aligned}\theta_{k+1} &= \operatorname{argmax}_{\theta} \left(V^{\pi_\theta}(\rho) - V^{\pi_{\theta_k}}(\rho) - \frac{1}{\eta} D(\theta, \theta_k) \right) \\ &\approx \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) (\pi_\theta(a|s) - \pi_{\theta_k}(a|s)) \underbrace{Q^{\pi_{\theta_k}}(s, a) - b(s)}_{\text{Performance Difference (area)}} \quad (\theta \rightarrow \theta_k) \\ &\approx (\theta - \theta_k)^\top \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s) \Big|_{\theta=\theta_k} \right) \underbrace{Q^{\pi_{\theta_k}}(s, a) - b(s)}_{\text{Performance Difference (area)}} \\ &= (\theta - \theta_k)^\top \left(\nabla_\theta V^{\pi_\theta}(\rho) \Big|_{\theta=\theta_k} \right)\end{aligned}$$

Summary: Full-Information Policy Learning in MDPs

PG	NPG
$\operatorname{argmax}_{\theta} \langle \theta - \theta_k, \nabla_{\theta} V^{\pi_{\theta_k}}(\rho) \rangle - \frac{1}{2\eta} \ \theta - \theta_k\ ^2$	$\operatorname{argmax}_{\theta} \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \pi_{\theta}(a s) Q^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta} \sum_s d_{\rho}^{\pi_{\theta_k}}(s) \text{KL}(\pi_{\theta}(\cdot s), \pi_{\theta_k}(\cdot s))$
$\theta_k + \eta \nabla_{\theta} V^{\pi_{\theta_k}}(\rho)$	$\theta_k + \eta F_{\theta_k}^{-1} \nabla_{\theta} V^{\pi_{\theta_k}}(\rho)$ where $F_{\theta} = \mathbb{E}_{s \sim d_{\rho}^{\pi_{\theta}}} \mathbb{E}_{a \sim \pi_{\theta}} [(\nabla_{\theta} \log \pi_{\theta}(a s)) (\nabla_{\theta} \log \pi_{\theta}(a s))^T]$
	Tabular Case: $\pi_{k+1}(a s) = \frac{\pi_k(a s) \exp(\eta Q^{\pi_k}(s,a))}{\sum_{a'} \pi_k(a' s) \exp(\eta Q^{\pi_k}(s,a'))}$

Policy Learning in MDPs

(Bandit Feedback Case)

NPG (Regularization Form)

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

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

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

NPG (Regularization Form)

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

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

For a fixed θ , an estimator for $\sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \pi_{\theta}(a|s) (Q^{\pi_{\theta_k}}(s,a) - b(s))$ 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}

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

$$\mathbb{E} \left[\sum_{h=1}^{\tau} \gamma^{h-1} \frac{\pi_{\theta}(a_h|s_h)}{\pi_{\theta_k}(a_h|s_h)} R_h \right] = \sum_{h=1}^{\tau} \sum_{s,a} \gamma^{h-1} \Pr\{s_h=s\} \pi_{\theta_k}(a|s) \cdot \frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} Q^{\pi_{\theta_k}}(s,a)$$

Define the estimator as

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

$d_{\rho}^{\pi_{\theta_k}}$

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

NPG (Regularization Form) + Bandit Feedback

For $k = 1, 2, \dots$

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

Use π_{θ_k} to collect n trajectories

$$(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}), \dots, (s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)})$$

$$\theta_{k+1} = \operatorname{argmax}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \gamma^h \cancel{\left[\frac{\pi_{\theta}(a_h^{(i)} | s_h^{(i)})}{\pi_{\theta_k}(a_h^{(i)} | s_h^{(i)})} \left(R_h^{(i)} - b(s_h^{(i)}) \right) \right]} \leq \frac{1}{1-\gamma} \right. \\ \left. - \frac{1}{\eta n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \gamma^h \cancel{\left[\text{KL} \left(\pi_{\theta}(\cdot | s_h^{(i)}), \pi_{\theta_k}(\cdot | s_h^{(i)}) \right) \right]} \right\}$$

Practical version will not include the discount factor at the front

NPG (Regularization Form) + Bandit Feedback

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

$$(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}), \dots, (s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)})$$

$$\theta_{k+1} = \operatorname{argmax}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \frac{\pi_{\theta}(a_h^{(i)} | s_h^{(i)})}{\pi_{\theta_k}(a_h^{(i)} | s_h^{(i)})} \left(R_h^{(i)} - b(s_h^{(i)}) \right) - \frac{1}{\eta n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \text{KL}(\pi_{\theta}(\cdot | s_h^{(i)}), \pi_{\theta_k}(\cdot | s_h^{(i)})) \right\}$$

NPG (Regularization Form) + Bandit Feedback

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

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

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

$$\theta \leftarrow \theta_k$$

Repeat m times:

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

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

PG

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

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

PG + Bandit Feedback (REINFORCE)

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

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

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 $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

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

Define

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

Perform update

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

NPG (Gradient-Update Form) + Bandit Feedback

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

$$(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}), \dots, (s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)})$$

Define

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

Perform update

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

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

$$\left(\sum_i x_i x_i^T \right) \left(\sum_i x_i y_i \right) = \text{the solution of} \\ \arg \min_w \sum_i (x_i^T w - y_i)^2$$

$$F_{\theta_k}^{-1} g = \arg \min_w \sum_{i=1}^n \sum_{h=1}^{\tau_n} \left(\nabla_{\theta} \log \pi_{\theta_k} (a_h^{(i)} | s_h^{(i)})^T w - (R_h^{(i)} - b(s_h^{(i)})) \right)^2$$

Summary: Policy Learning in MDPs

PG	NPG
$\operatorname{argmax}_{\theta} \langle \theta - \theta_k, \nabla_{\theta} V^{\pi_{\theta_k}}(\rho) \rangle - \frac{1}{2\eta} \ \theta - \theta_k\ ^2$	$\operatorname{argmax}_{\theta} \left[\sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \pi_{\theta}(a s) Q^{\pi_{\theta_k}}(s,a) \right] - \frac{1}{\eta} \sum_s d_{\rho}^{\pi_{\theta_k}}(s) \text{KL}(\pi_{\theta}(\cdot s), \pi_{\theta_k}(\cdot s))$
$\theta_k + \eta \boxed{\nabla_{\theta} V^{\pi_{\theta_k}}(\rho)}$	$\theta_k + \eta F_{\theta_k}^{-1} \nabla_{\theta} V^{\pi_{\theta_k}}(\rho)$ <p>where $F_{\theta} = \mathbb{E}_{s \sim d_{\rho}^{\pi_{\theta}}} \mathbb{E}_{a \sim \pi_{\theta}} [(\nabla_{\theta} \log \pi_{\theta}(a s)) (\nabla_{\theta} \log \pi_{\theta}(a s))^T]$</p>

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

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

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

Actor-Critic Methods

Review: Full-Information Policy Learning in MDPs

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

}

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

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

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

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 $V_\phi(s)$:

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

Use $Q_\phi(s, a)$:

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

Possible estimators for $A^{\pi_{\theta_k}}(s, a)$:

Let $(s_1, a_1, r_1, s_2, a_2, r_2 \dots)$ be a trajectory starting from $s_1 = s, a_1 = a$

$$r_1 + \gamma V_\phi(s_2) - V_\phi(s_1)$$

$$r_1 + \gamma r_2 + \gamma^2 V_\phi(s_3) - V_\phi(s_1)$$

⋮

$$Q_\phi(s_1, a_1) - \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s)} [Q_\phi(s_1, a')]$$

$$r_1 + \gamma Q_\phi(s_2, a_2) - \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s)} [Q_\phi(s_1, a')]$$

$$r_1 + \gamma r_2 + \gamma^2 Q_\phi(s_3, a_3) - \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s)} [Q_\phi(s_1, a')]$$

⋮

Pure Policy-Based Methods vs. Actor-Critic Methods

Actor-Critic with Q_ϕ

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

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

Define

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

Perform updates

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

Advantage Actor-Critic (A2C) = PG + V_ϕ

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

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

Define

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

or any other advantage estimator in the previous slide

Perform updates

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

Proximal Policy Optimization (PPO) = NPG + V_ϕ

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

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

Perform updates

$$\theta_{k+1} \leftarrow \operatorname{argmax}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \frac{\pi_{\theta}\left(a_h^{(i)} \mid s_h^{(i)}\right)}{\pi_{\theta_k}\left(a_h^{(i)} \mid s_h^{(i)}\right)} \left(r_h^{(i)} + \gamma V_{\phi_k}\left(s_{h+1}^{(i)}\right) - V_{\phi_k}\left(s_h^{(i)}\right) \right) \right.$$

or any other advantage estimator in the previous slide

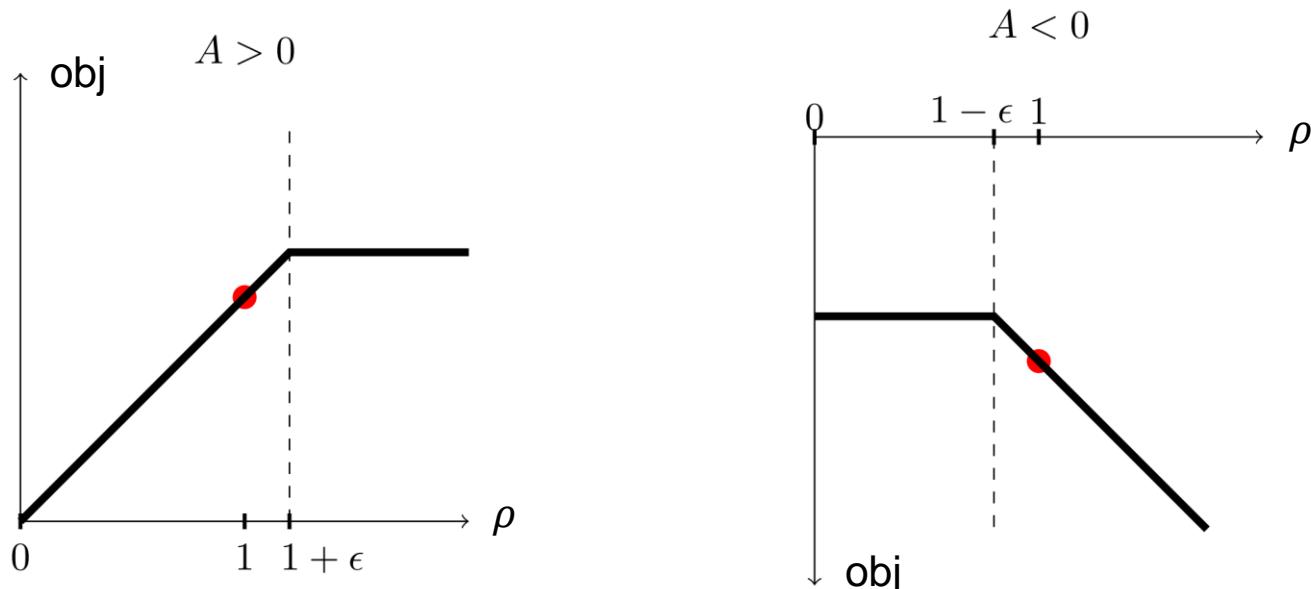
$$\left. - \frac{1}{\eta n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \text{KL}\left(\pi_{\theta}\left(\cdot \mid s_h^{(i)}\right), \pi_{\theta_k}\left(\cdot \mid s_h^{(i)}\right)\right) \right\}$$

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

Additional Technique 1: Clipped Objective (for PPO)

$$\rho := \frac{\pi_{\theta} \left(a_h^{(i)} \middle| s_h^{(i)} \right)}{\pi_{\theta_k} \left(a_h^{(i)} \middle| s_h^{(i)} \right)}$$
$$A := \left(r_h^{(i)} + \gamma V_{\phi_k} \left(s_{h+1}^{(i)} \right) - V_{\phi_k} \left(s_h^{(i)} \right) \right)$$

Instead of using ρA as the objective, use $\min\{\rho A, \text{clip}_{[1-\epsilon, 1+\epsilon]}(\rho)A\}$



algorithm	avg. normalized score
No clipping or penalty	-0.39
Clipping, $\epsilon = 0.1$	0.76
Clipping, $\epsilon = 0.2$	0.82
Clipping, $\epsilon = 0.3$	0.70
Adaptive KL $d_{\text{targ}} = 0.003$	0.68
Adaptive KL $d_{\text{targ}} = 0.01$	0.74
Adaptive KL $d_{\text{targ}} = 0.03$	0.71
Fixed KL, $\beta = 0.3$	0.62
Fixed KL, $\beta = 1.$	0.71
Fixed KL, $\beta = 3.$	0.72
Fixed KL, $\beta = 10.$	0.69

Additional Technique 2: Entropy Bonus

In the objective of policy update, add a bonus term

$$H(\pi_\theta(\cdot | s)) = \sum_a \pi_\theta(a|s) \ln \frac{1}{\pi_\theta(a|s)}$$

For PPO:

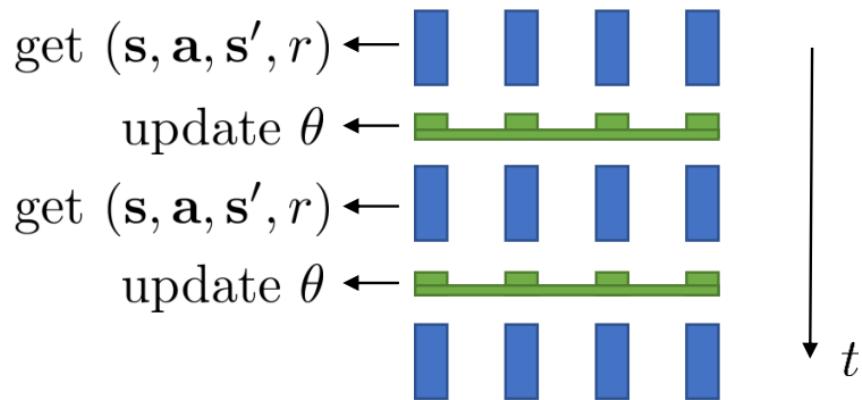
$$\operatorname{argmax}_\theta \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \frac{\pi_\theta(a_h^{(i)} | s_h^{(i)})}{\pi_{\theta_k}(a_h^{(i)} | s_h^{(i)})} A_h^{(i)} - \frac{1}{\eta} \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \text{KL}\left(\pi_\theta(\cdot | s_h^{(i)}), \pi_{\theta_k}(\cdot | s_h^{(i)})\right) + c \underbrace{\frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} H\left(\pi_\theta(\cdot | s_h^{(i)})\right)}_{\text{KL}\left(\pi_\theta(\cdot | s_h^{(i)}), \pi_{\text{unif}}(\cdot | s_h^{(i)})\right)} \right\}$$

For A2C:

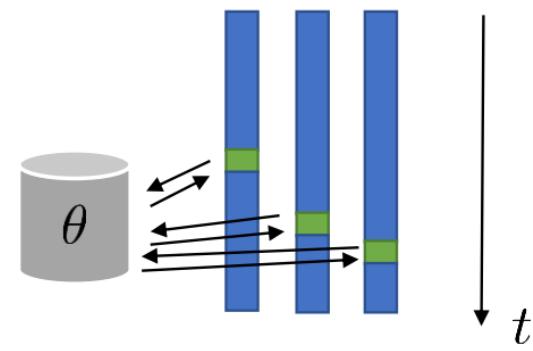
$$g = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \nabla_\theta \log \pi_\theta(a_h^{(i)} | s_h^{(i)}) \Big|_{\theta=\theta_k} A_h^{(i)} + c \nabla_\theta \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} H\left(\pi_\theta(\cdot | s_h^{(i)})\right)$$

Additional Technique 3: Parallel Experience Collection

synchronized parallel actor-critic



asynchronous parallel actor-critic



[Levine CS285 Lecture 6](#)

Actor-Critic Summary

Off-Policy Actor-Critic

- Leveraging **off-policy evaluation** → allow reusing data

Review: Full-Information Policy Learning in MDPs

$$\begin{aligned}\theta_{k+1} &= \operatorname{argmax}_{\theta} \left(V^{\pi_\theta}(\rho) - V^{\pi_{\theta_k}}(\rho) - \frac{1}{\eta} D(\theta, \theta_k) \right) \\ &\approx \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\pi_\theta(a|s) - \pi_{\theta_k}(a|s) \right) Q^{\pi_{\theta_k}}(s, a) \\ &\approx (\theta - \theta_k)^\top \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s) \Big|_{\theta=\theta_k} \right) Q^{\pi_{\theta_k}}(s, a)\end{aligned}$$

Use any off-policy policy evaluation methods to find ϕ_k such that $Q_{\phi_k}(s, a) \approx Q^{\pi_{\theta_k}}(s, a)$

Suppose that our (s_i, a_i) samples are obtained from $\hat{\pi}$

Off-Policy Actor-Critic

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

$$\approx \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\pi_\theta(a|s) - \pi_{\theta_k}(a|s) \right) Q_{\phi_k}(s, a) = \mathbb{E}_{s \sim \hat{\pi}} \left[\frac{d_\rho^{\pi_{\theta_k}}(s)}{d_\rho^\pi(s)} \sum_a \left(\pi_\theta(a|s) - \pi_{\theta_k}(a|s) \right) Q_{\phi_k}(s, a) \right]$$

$$\approx (\theta - \theta_k)^\top \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s) \left(\nabla_\theta \pi_\theta(a|s) \Big|_{\theta=\theta_k} \right) Q_{\phi_k}(s, a) = (\theta - \theta_k)^\top \mathbb{E}_{s \sim \hat{\pi}} \left[\frac{d_\rho^{\pi_{\theta_k}}(s)}{d_\rho^\pi(s)} \sum_a \nabla_\theta \pi_\theta(a|s) \Big|_{\theta=\theta_k} Q_{\phi_k}(s, a) \right]$$

Use any off-policy policy evaluation methods to find ϕ_k such that $Q_{\phi_k}(s, a) \approx Q^{\pi_{\theta_k}}(s, a)$

Suppose that our (s_i, a_i) samples are obtained from $\hat{\pi}$

Actor-Critic + Replay Buffer

For $k = 1, 2, \dots$

Collect samples using π_{θ_k} , and place them in the replay buffer

Sample a batch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ from replay buffer

Define

$$g = \frac{1}{n} \sum_{i=1}^n \sum_a \nabla_{\theta} \pi_{\theta}(a|s_i) \Big|_{\theta=\theta_k} Q_{\phi_k}(s_i, a) \quad \text{Note: not using } a_i \text{ here}$$

Perform updates

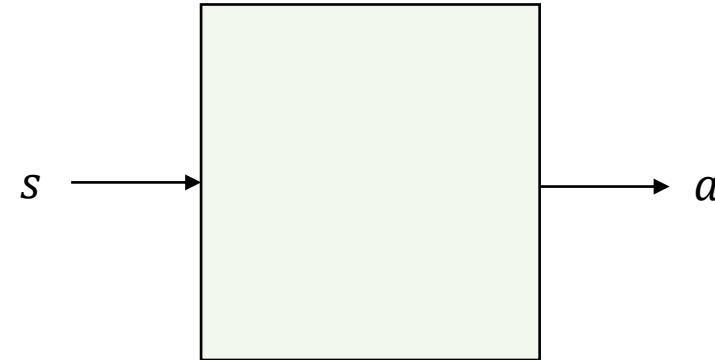
$$\theta_{k+1} \leftarrow \theta_k + \eta g$$

$$\phi_{k+1} \leftarrow \phi_k - \lambda \nabla_{\phi} \left(\frac{1}{n} \sum_{i=1}^n \left(Q_{\phi}(s_i, a_i) - r_i - \gamma \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s'_i)} [Q_{\phi}(s'_i, a')] \right)^2 \right) \Big|_{\phi=\phi_k}$$

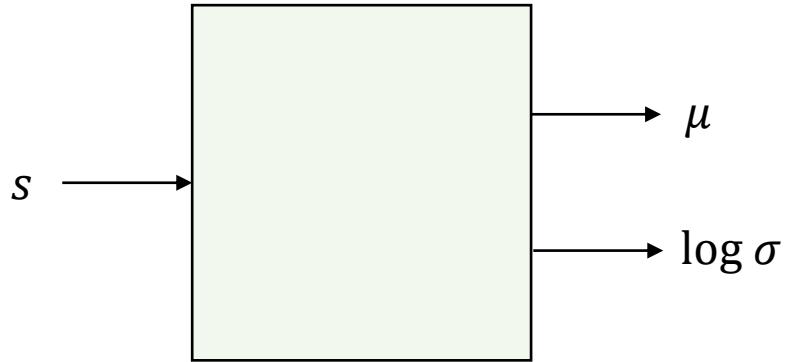
Dealing with Continuous Action Sets

Review: Linear Bandits and One-Point Gradient Estimator

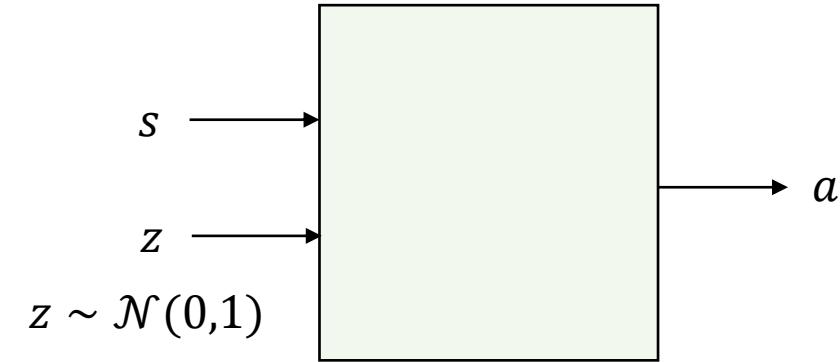
Policy Network for Continuous Action Sets



Policy Network for Continuous Action Sets



Explicitly models $\pi_\theta(a|s)$



Implicitly modeling $\pi_\theta(a|s)$

Option 1: making σ part of policy parameters

Option 2: making σ a hyper-parameters
(decreases over time)

can sample from it, but do not
know the function $\pi_\theta(\cdot | s)$

A2C / PPO with Continuous Action Sets

$$g = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \log \pi_{\theta}(a_i | s_i) \Big|_{\theta=\theta_k} A_i$$

$$\theta_{k+1} \leftarrow \operatorname{argmax}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\pi_{\theta}(a_i | s_i)}{\pi_{\theta_k}(a_i | s_i)} A_i - \frac{1}{\eta} \frac{1}{n} \sum_{i=1}^n \text{KL} \left(\pi_{\theta}(\cdot | s_i), \pi_{\theta_k}(\cdot | s_i) \right) \right\}$$

Recall: Actor-Critic without need for inverse weighting

Actor-critic with $Q_\phi(s, a)$ function approximation

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

$$(s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_{\tau_1}^{(1)}, a_{\tau_1}^{(1)}, r_{\tau_1}^{(1)}), \dots, (s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{\tau_n}^{(n)}, a_{\tau_n}^{(n)}, r_{\tau_n}^{(n)})$$

Define $g = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^{\tau_n} \sum_a \nabla_\theta \pi_\theta(a | s_h^{(i)}) \Big|_{\theta=\theta_k} Q_{\phi_k}(s_h^{(i)}, a)$

Perform updates

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

Deterministic Policy Gradient Theorem

Deterministic Policy Gradient Algorithm

For $k = 1, 2, \dots$

Use π_{θ_k} to collect n trajectories

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

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

Perform updates

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

Two Viewpoints for the Deterministic PG Algorithm

Deep Deterministic Policy Gradient (DDPG)

For $k = 1, 2, \dots$

Use π_θ to collect samples and place them in replay buffer

Sample a batch $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ from the replay buffer

$$\theta \leftarrow \theta + \eta \sum_{i=1}^n \nabla_\theta Q_\phi(s_i, \pi_\theta(s_i))$$

$$\phi \leftarrow \phi - \lambda \nabla_\phi \sum_{i=1}^n \left(Q_\phi(s_i, a_i) - r_i - \gamma Q_{\phi_{\text{tar}}}(s'_i, \pi_{\theta_{\text{tar}}}(s'_i)) \right)^2$$

$$\theta_{\text{tar}} \leftarrow \tau \theta + (1 - \tau) \theta_{\text{tar}}$$

$$\phi_{\text{tar}} \leftarrow \tau \phi + (1 - \tau) \phi_{\text{tar}}$$

Twin Delayed DDPG (TD3)

Soft Actor-Critic (SAC)