Approximate Policy Iteration and Policy-Based Learning Methods

Chen-Yu Wei

Approximate Policy Iteration (API)

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

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

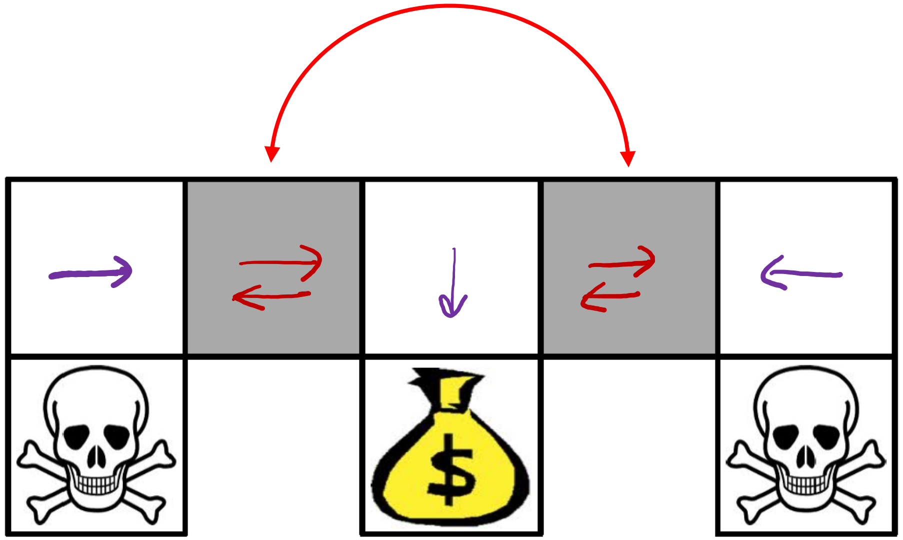
$$\pi_{k+1}(s) \leftarrow \underset{a}{\operatorname{argmax}}\, \hat{Q}_k(s, a)$$

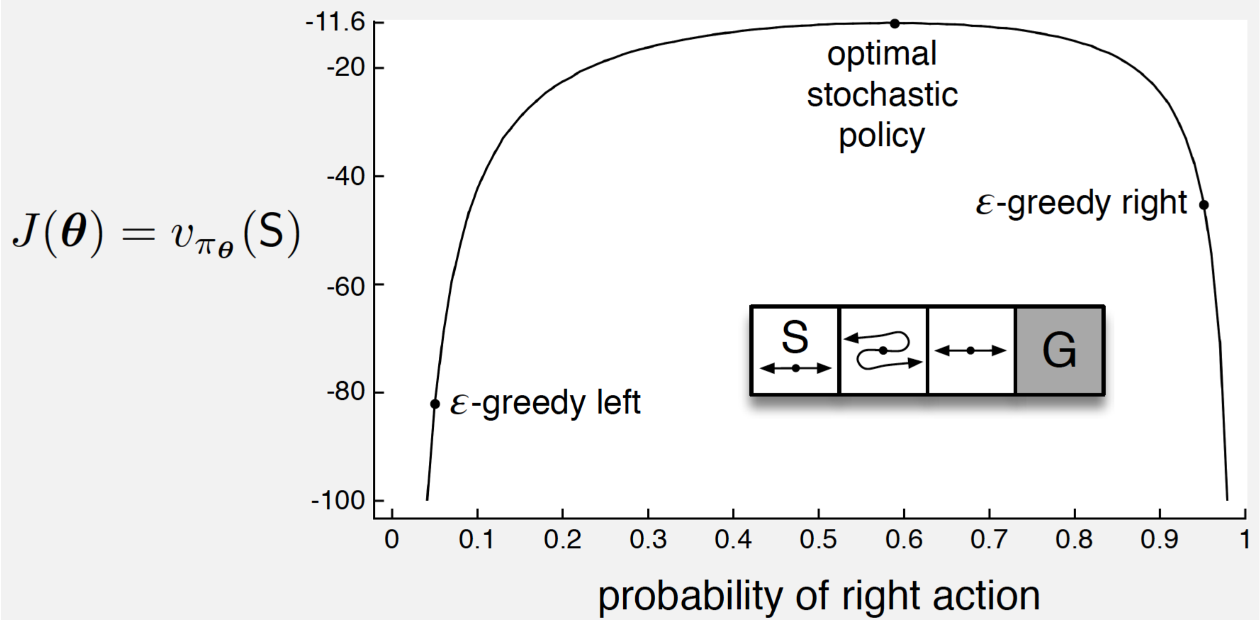
Q^{π}

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

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

Limitation of Value Function Approximation





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 $k = 1,\ 2, \ldots$

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

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

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

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

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

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

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

??

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

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

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

✓

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

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

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

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

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

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

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

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

Exponential weights:

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

$$\| $$

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

PG over softmax

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

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

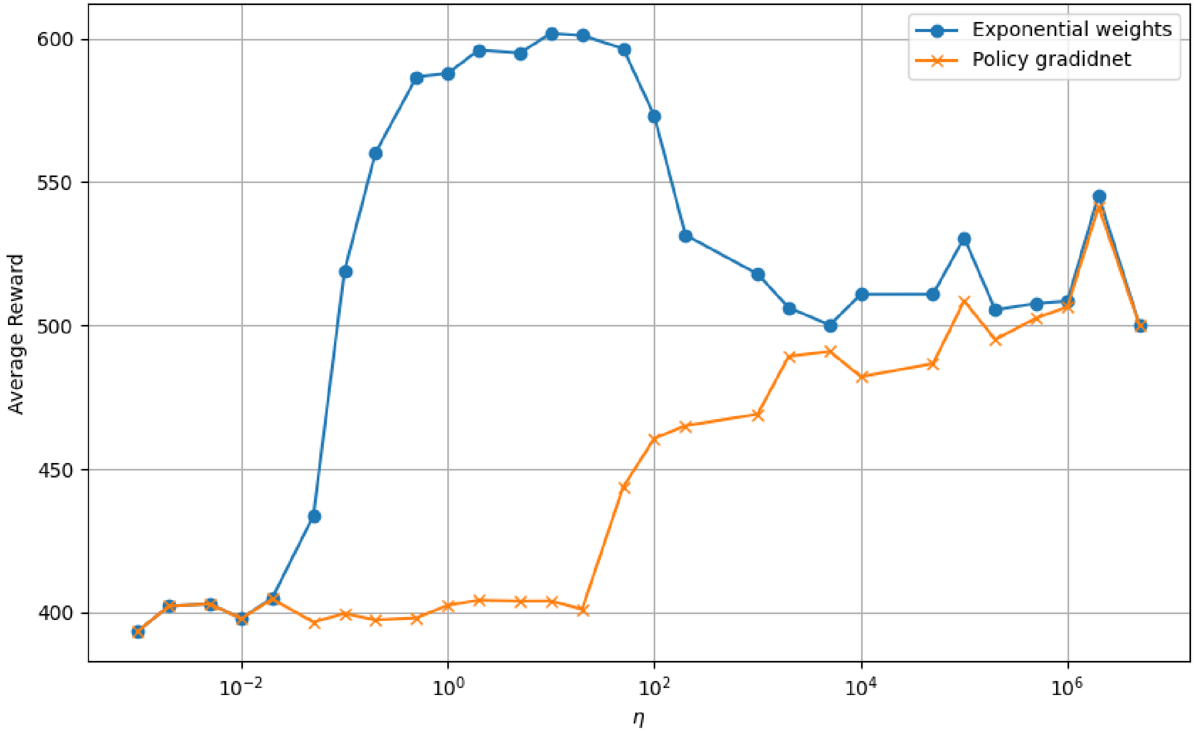
small eta: too slow on action 1

larger eta: too fast on action 2

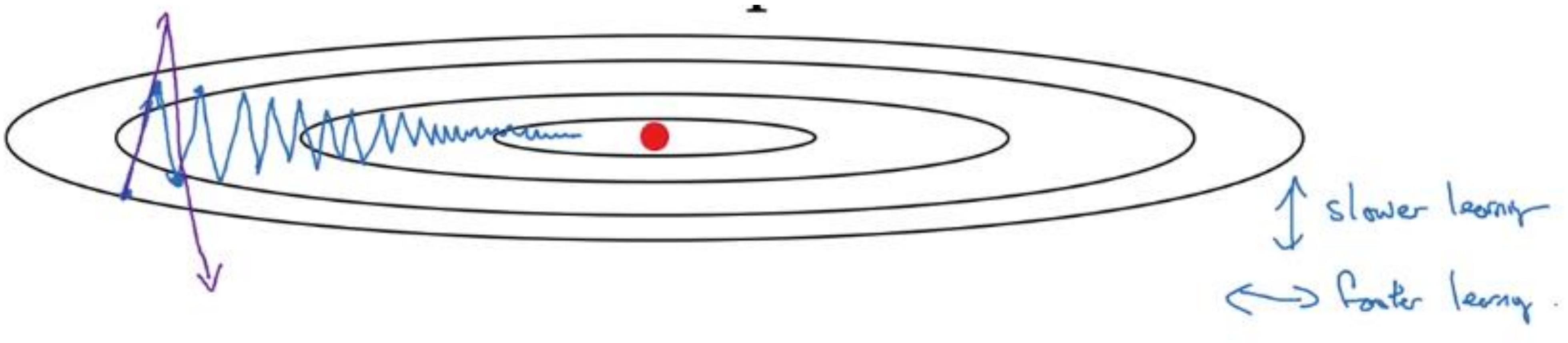
Reward = $[\mathrm{Ber}(0.6), \mathrm{Ber}(0.4)]$

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

Plot total reward in 1000 rounds

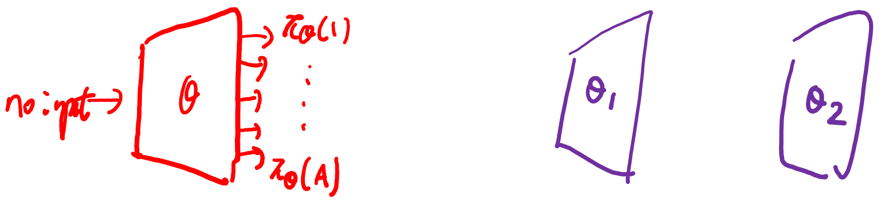


Optimization over ill-conditioned loss



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

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

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

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

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

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

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

Exponential weights

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

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

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

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

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

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

Two Ideas for Function Approximation over Policies

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

(Vanilla) Policy Gradient

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

Natural Policy Gradient

Approximating the NPG Update

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

$V^\pi = \langle \pi, R \rangle$

$= \sum_a \pi(a) R(a)$

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)

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

$\downarrow$

$\Delta\theta \to 0$

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

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

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

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

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

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

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

$\downarrow$

$$-\sum_a \nabla \pi_\theta(a)^T \Delta\theta$$

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

$$= 0$$

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

For any θ, $\sum_a \pi_\theta(a) = 1$

NPG Updates

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

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

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

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

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

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

$$\approx \underbrace{\arg\max_\theta \left\{ \langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}} \rangle - \frac{1}{2\eta}(\theta-\theta_k)^T F_{\theta_k}(\theta-\theta_k) \right\}}_{\to W(\theta)}$$

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

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 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

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

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

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

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

NPG (regularization form) + Bandit Feedback

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

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

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

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 $k = 1,\ 2, \ldots$

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

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

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

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

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

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

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

Summary: Policy Learning in Bandits

PG	NPG
$\theta_{k+1} = \underset{\theta}{\text{argmax}} \langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}} \rangle - \frac{1}{2\eta} \lVert \theta - \theta_k \rVert^2$	$\theta_{k+1} = \underset{\theta}{\text{argmax}} \langle \pi_\theta - \pi_{\theta_k}, \boxed{R} \rangle - \frac{1}{\eta} \text{KL}(\pi_\theta, \pi_{\theta_k})$
$\theta_{k+1} = \theta_k + \eta \boxed{\nabla_\theta V^{\pi_{\theta_k}}}$	$\theta_{k+1} = \theta_k + \eta F_{\theta_k}^{-1} \nabla_\theta V^{\pi_{\theta_k}}$ 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 $k = 1,\ 2, \ldots$

Perform individual exponential weight update **on all state $\boldsymbol{s}$:**

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

Analysis for Exponential Weights

Theorem.

The exponential weight algorithm guarantees

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

for any initial state distribution ρ.

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

Equivalent Forms of Exponential Weights

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

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

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

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

...

Natural Policy Gradient (Regularization Form)

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

Policy Gradient

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

PG vs. NPG (Regularization Form)

Policy Gradient

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

Natural Policy Gradient

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

Natural Policy Gradient (Gradient-Update Form)

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

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

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

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

Summary: Full-Information Policy Learning in MDPs

PG	NPG
$\underset{\theta}{\operatorname{argmax}} \left\langle \theta - \theta_k, \nabla_\theta V^{\pi_{\theta_k}}(\rho) \right\rangle - \frac{1}{2\eta} \|\theta - \theta_k\|^2$	$\underset{\theta}{\operatorname{argmax}} \sum_{s,a} d_\rho^{\pi_{\theta_k}}(s)\, \pi_\theta(a\|s) A^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta} \sum_s d_\rho^{\pi_{\theta_k}}(s)\, \mathrm{KL}\big(\pi_\theta(\cdot\,\|s), \pi_{\theta_k}(\cdot\,\|s)\big)$
$\theta_k + \eta \nabla_\theta V^{\pi_{\theta_k}}(\rho)$	$\theta_k + \eta F_{\theta_k}^{-1} \nabla_\theta V^{\pi_{\theta_k}}(\rho)$ Type equation here. where $F_\theta = \mathbb{E}_{s \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta} [(\nabla_\theta \log \pi_\theta(a\|s))(\nabla_\theta \log \pi_\theta(a\|s))^\top]$
	Tabular Case: $\pi_{k+1}(a\|s) = \frac{\pi_k(a\|s) \exp(\eta A^{\pi_k}(s,a))}{\sum_{a'} \pi_k(a'\|s) \exp(\eta A^{\pi_k}(s,a'))}$

Policy Learning in MDPs

(Bandit Feedback Case)

NPG (Regularization Form)

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

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

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

NPG (Regularization Form)

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

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

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

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)

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

Use π_{θ_k} to collect n trajectories

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

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

Practical version will not include the discount factor at the front

NPG (Regularization Form)

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

Use π_{θ_k} to collect n trajectories

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

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

NPG (Regularization Form)

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

Use π_{θ_k} to collect n trajectories

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

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

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