

# **Approximate Policy Iteration and Policy-Based Learning Methods**

Chen-Yu Wei

# Approximate Policy Iteration (API)

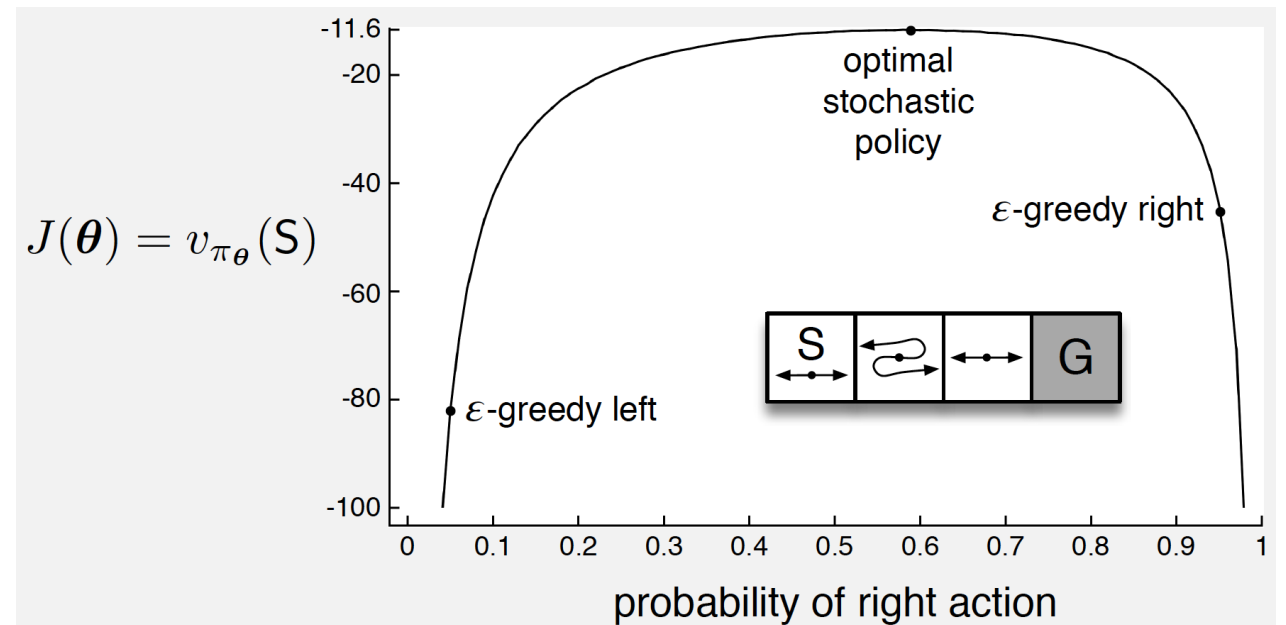
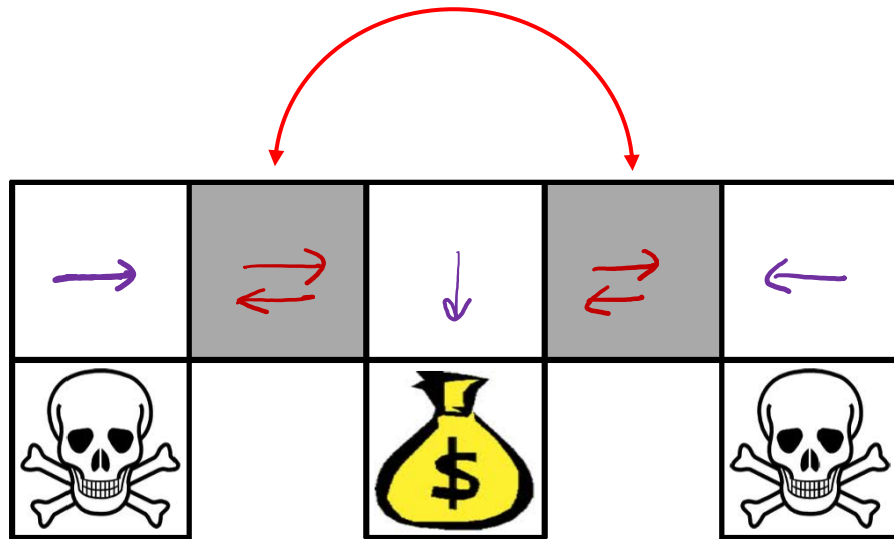
For  $k = 1, 2, \dots$

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

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

Value-based :  $\overset{Q^z}{Q^*}, V^z, V^* \approx \boxed{V_\theta}$   
Policy-based :  $\tau_\theta(a/s)$

# 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

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

## Idea 2: Policy Gradient

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

For  $k = 1, 2, \dots$

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

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

$V^{\pi_\theta}$

How are exponential weights and policy gradient related?

# **Policy Learning in the Expert Setting**

# Policy Gradient for Softmax Policy in Expert Problem

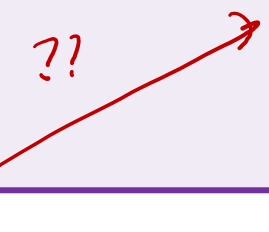
Assume full-information and fixed reward  $R = (R(1), \dots, R(A))$

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

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

*Exponential weight*

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



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

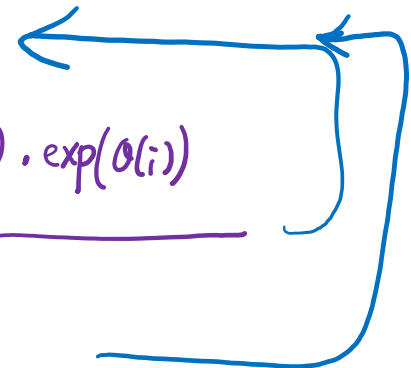
PG:  $\theta_{k+1} = \theta_k + \eta \nabla_\theta V^{\pi_{\theta_k}} \Big|_{\theta=\theta_k}$

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

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

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



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

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

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



Exponential weights:

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

$$\begin{aligned} \pi_{k+1}(i) &= \frac{\pi_k(i) \exp(\eta R(i))}{\sum_b \pi_k(b) \exp(\eta R(b))} \equiv \frac{\pi_k(i) \exp(\eta A_{\pi_k}(i))}{\sum_b \pi_k(b) \exp(\eta A_{\pi_k}(b))} \\ &= \frac{\pi_k(i) \exp(\eta R(i) - c)}{\sum_b \pi_k(b) \exp(\eta R(b) - c)} \frac{\exp(-c)}{\exp(-c)} \end{aligned}$$

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

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

## 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)]

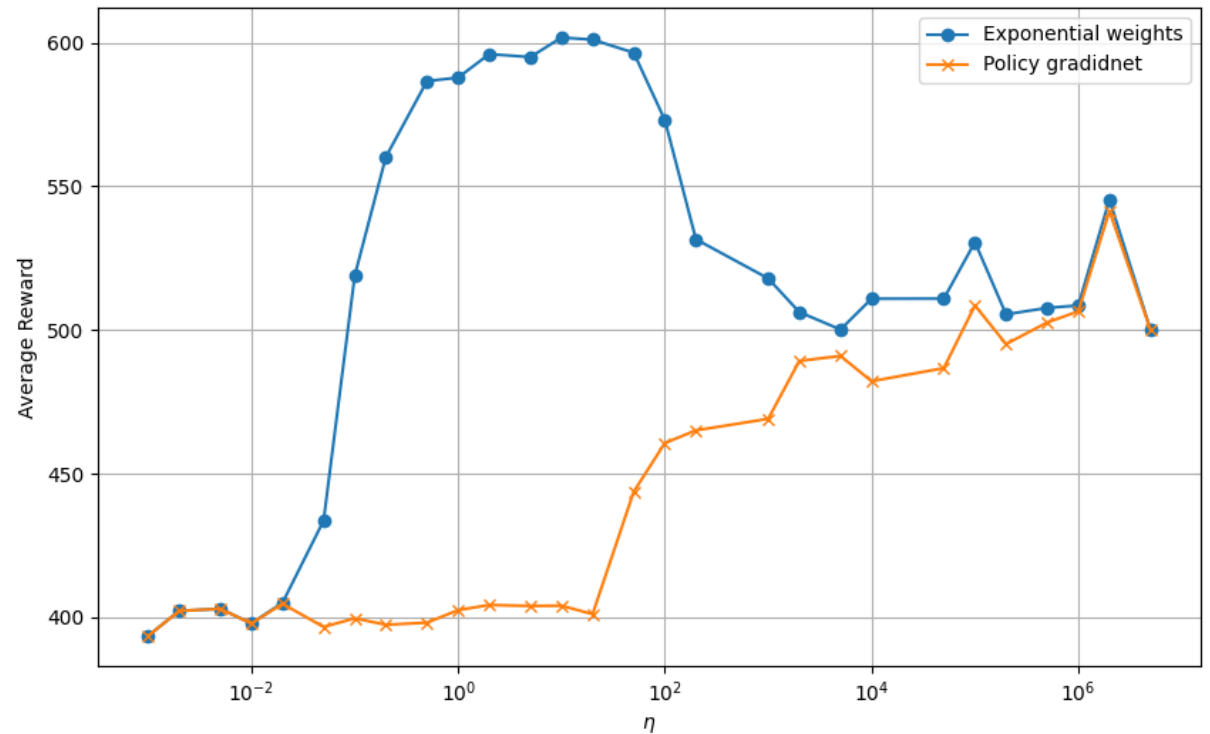
Initial policy  $\pi = [0.0001, 0.9999]$

Plot total reward in 1000 rounds

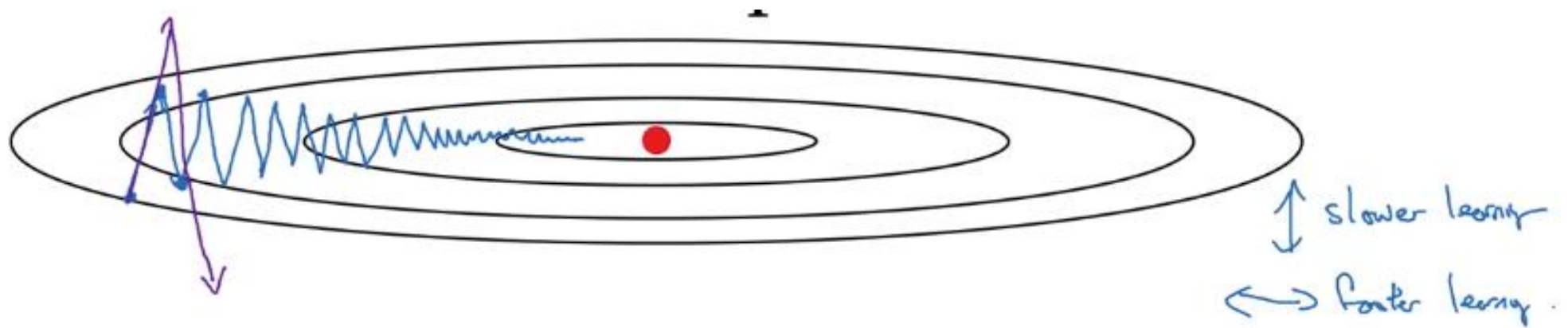
**EW:**  $\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta A_{\theta_k}(a)$

**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*

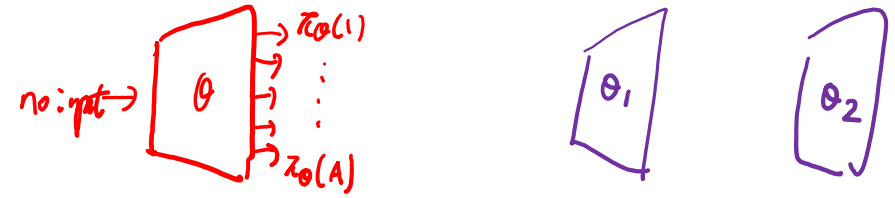


# Optimization over ill-conditioned loss



<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

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

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

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

## Exponential weights

$$\theta_{k+1}(a) \leftarrow \theta_k(a) + \eta A_{\theta_k}(a) \left( R(a) - \text{const} \right)$$

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

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

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

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

$$\begin{aligned} &= \operatorname{argmax}_{\theta} \langle \pi_{\theta} - \pi_{\theta_k}, A_{\theta_k} \rangle - \frac{1}{\eta} \text{KL}(\pi_{\theta}, \pi_{\theta_k}) \\ &\stackrel{a}{\approx} \sum_a (\pi_{\theta}(a) - \pi_{\theta_k}(a)) \text{cost} = 0 \\ &R(a) = R(a) - \text{const} \end{aligned}$$

# 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**

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

**Natural Policy Gradient**

# Approximating the NPG Update

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

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

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

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

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

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

**(Fisher information matrix)**

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

$\Delta\theta \rightarrow 0$

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

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

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

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

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

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

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

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

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

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



# NPG Updates

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

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

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

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

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

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

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

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

# Summary: Policy Learning in the Expert Setting

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

| PG                                                                                                                                                              | NPG                                                                                                                                                                                                                                                                                                                                                                 |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\theta_{k+1} = \operatorname{argmax}_{\theta} \langle \theta - \theta_k, \nabla_{\theta} V^{\pi_{\theta_k}} \rangle - \frac{1}{2\eta} \ \theta - \theta_k\ ^2$ | $\theta_{k+1} = \operatorname{argmax}_{\theta} \langle \pi_{\theta} - \pi_{\theta_k}, R \rangle - \frac{1}{\eta} \operatorname{KL}(\pi_{\theta}, \pi_{\theta_k})$ <p style="text-align: right; color: red; font-size: small;"> <math>V^{\pi_{\theta}} - V^{\pi_{\theta_k}}</math><br/> <math>\approx \frac{1}{2} \ \theta - \theta_k\ _{F_{\theta_k}}^2</math> </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 <math>F_{\theta} = \mathbb{E}_{a \sim \pi_{\theta}} [(\nabla_{\theta} \log \pi_{\theta}(a))(\nabla_{\theta} \log \pi_{\theta}(a))^{\top}]</math></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>                                                                                                                                                                                                                                                               |

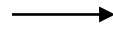
# Policy Learning with Bandit Feedback



# 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

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

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

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

# NPG (regularization form) + Bandit Feedback

For  $k = 1, 2, \dots$

Use  $\pi_{\theta_k}$  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  $\pi_{\theta_k}$  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

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

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

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

PG

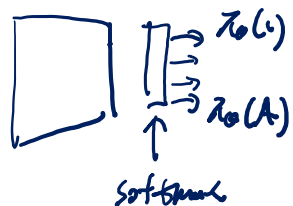
NPG

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

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

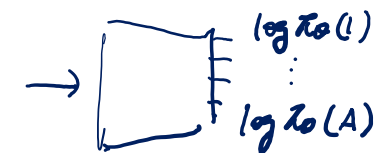
Suppose that  $a_k \sim \pi_{\theta}$ , get  $r_k$

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



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

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



weight =  $r_k$



# PG + Bandit Feedback

For  $k = 1, 2, \dots$

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

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

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

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

# Summary: Policy Learning in Bandits

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

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

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

# **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

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

## Theorem.

The exponential weight algorithm guarantees

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

for any initial state distribution  $\rho$ .

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

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

$$\pi_{\text{opt}}(a|s) \propto \pi_k(a|s) \exp\left(\frac{\eta \pi_k(s,a)}{\text{reward}(a)}\right)$$

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

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

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

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

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



# Equivalent Forms of Exponential Weights

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

$$\sum_a \pi(a|s) \underbrace{(Q^{\pi_k}(s, a) - b(s))}_{V^{\pi_k}(s)} = \boxed{\phantom{\sum_a \pi(a|s) Q^{\pi_k}(s, a)}} - b(s)$$

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

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

...

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

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

# Natural Policy Gradient (Regularization Form)

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

# 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

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

Dividing both sides by  $\epsilon$  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

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

## Natural Policy Gradient

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

## Policy Gradient

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

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

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

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

# NPG (Gradient-Update Form)

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

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

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

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

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

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

# Summary: Full-Information Policy Learning in MDPs

Unified Idea:

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

# Summary: Full-Information Policy Learning in MDPs

| PG                                                                                                                                                     | NPG                                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\operatorname{argmax}_{\theta} \langle \theta - \theta_k, \nabla_{\theta} V^{\pi_{\theta_k}}(\rho) \rangle - \frac{1}{2\eta} \ \theta - \theta_k\ ^2$ | $\operatorname{argmax}_{\theta} \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \pi_{\theta}(a s) Q^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta} \sum_s d_{\rho}^{\pi_{\theta_k}}(s) \operatorname{KL}(\pi_{\theta}(\cdot s), \pi_{\theta_k}(\cdot s))$                                                   |
| $\theta_k + \eta \nabla_{\theta} V^{\pi_{\theta_k}}(\rho)$                                                                                             | $\theta_k + \eta F_{\theta_k}^{-1} \nabla_{\theta} V^{\pi_{\theta_k}}(\rho)$ <p>where <math>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}]</math></p> |
|                                                                                                                                                        | <p>Tabular Case:</p> $\pi_{k+1}(a s) = \frac{\pi_k(a s) \exp(\eta Q^{\pi_k}(s,a))}{\sum_{a'} \pi_k(a' s) \exp(\eta Q^{\pi_k}(s,a'))}$                                                                                                                                                       |

# **Policy Learning in MDPs**

(Bandit Feedback Case)



# NPG (Regularization Form)

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

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

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

# NPG (Regularization Form)

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

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

For a fixed  $\theta$ , 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  $\pi_{\theta_k}$

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

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

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

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

# NPG (Regularization Form) + Bandit Feedback

For  $k = 1, 2, \dots$

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

Use  $\pi_{\theta_k}$  to collect  $n$  trajectories

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

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

Practical version will not include the discount factor at the front

# NPG (Regularization Form) + Bandit Feedback

For  $k = 1, 2, \dots$

Use  $\pi_{\theta_k}$  to collect  $n$  trajectories

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

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

# NPG (Regularization Form) + Bandit Feedback

For  $k = 1, 2, \dots$

Use  $\pi_{\theta_k}$  to collect  $n$  trajectories

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

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

$$\theta \leftarrow \theta_k$$


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  $\pi_{\theta_k}$  to collect  $n$  trajectories

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

Define

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

Perform update

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

# PG + Bandit Feedback (REINFORCE)

For  $k = 1, 2, \dots$

Use  $\pi_{\theta_k}$  to collect  $n$  trajectories

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

Define

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

Perform update

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



# NPG (Gradient-Update Form) + Bandit Feedback

For  $k = 1, 2, \dots$

Use  $\pi_{\theta_k}$  to collect  $n$  trajectories

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

Define

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

Perform update

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

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

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

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

$d \times d$

# Summary: Policy Learning in MDPs

| PG                                                                                                                                                     | NPG                                                                                                                                                                                                                                                                                         |
|--------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\operatorname{argmax}_{\theta} \langle \theta - \theta_k, \nabla_{\theta} V^{\pi_{\theta_k}}(\rho) \rangle - \frac{1}{2\eta} \ \theta - \theta_k\ ^2$ | $\operatorname{argmax}_{\theta} \sum_{s,a} d_{\rho}^{\pi_{\theta_k}}(s) \pi_{\theta}(a s) Q^{\pi_{\theta_k}}(s,a) - \frac{1}{\eta} \sum_s d_{\rho}^{\pi_{\theta_k}}(s) \operatorname{KL}(\pi_{\theta}(\cdot s), \pi_{\theta_k}(\cdot s))$                                                   |
| $\theta_k + \eta \nabla_{\theta} V^{\pi_{\theta_k}}(\rho)$                                                                                             | $\theta_k + \eta F_{\theta_k}^{-1} \nabla_{\theta} V^{\pi_{\theta_k}}(\rho)$ <p>where <math>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}]</math></p> |

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

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

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