

# CS 8803

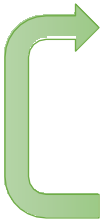
## Deep Reinforcement Learning

Lec 8: Advanced Policy Gradients  
Fall 2024

Animesh Garg  
Slides from Sergey Levine

# Recap: policy gradients

REINFORCE algorithm:

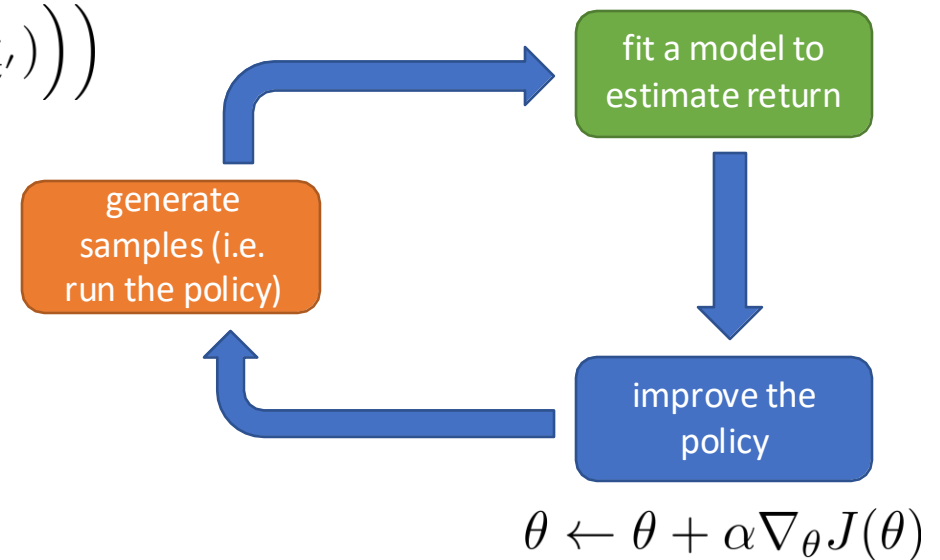
- 
1. sample  $\{\tau^i\}$  from  $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$  (run the policy)
  2.  $\nabla_\theta J(\theta) \approx \sum_i \left( \sum_{t=1}^T \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i) \left( \sum_{t'=t}^T r(\mathbf{s}_{t'}^i, \mathbf{a}_{t'}^i) \right) \right)$
  3.  $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \nabla_\theta \log \pi_\theta(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \underbrace{\hat{Q}_{i,t}^\pi}_{\text{"reward to go"}}$$

“reward to go”


can also use function approximation here

$$\hat{Q}^\pi(\mathbf{x}_t, \mathbf{u}_t) = \sum_{t'=t}^T r(\mathbf{x}_{t'}, \mathbf{u}_{t'})$$




# Why does policy gradient work?

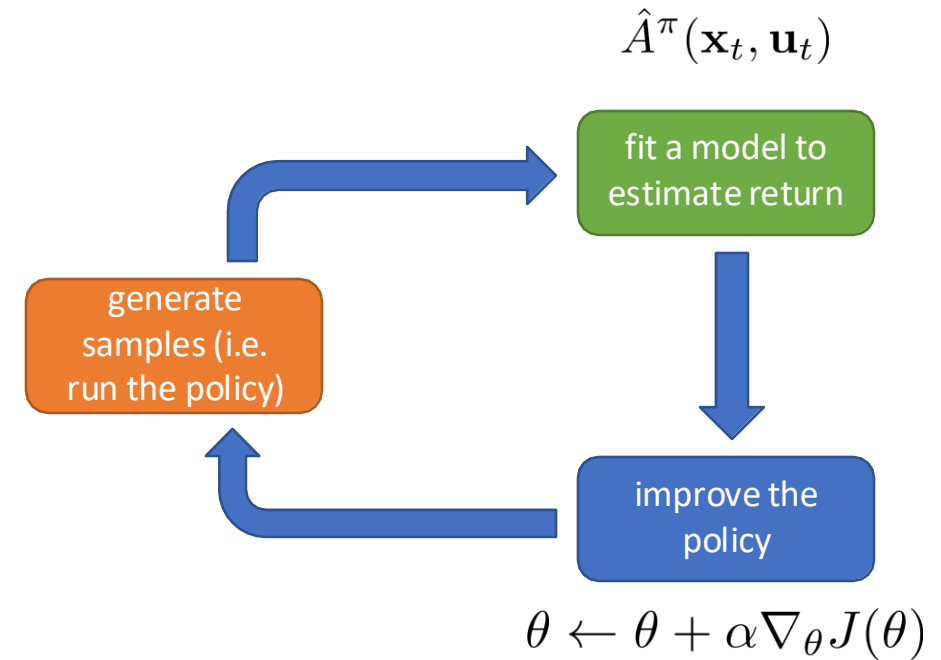
$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\theta} \log \pi_{\theta}(\mathbf{a}_{i,t} | \mathbf{s}_{i,t}) \hat{A}_{i,t}^{\pi}$$

- 
1. Estimate  $\hat{A}^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$  for current policy  $\pi$
  2. Use  $\hat{A}^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$  to get *improved* policy  $\pi'$

look familiar?

policy iteration algorithm:

- 
1. evaluate  $A^{\pi}(\mathbf{s}, \mathbf{a})$
  2. set  $\pi \leftarrow \pi'$



# Policy gradient as policy iteration

$$J(\theta) = E_{\tau \sim p_{\theta}(\tau)} \left[ \sum_t \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

$$\begin{aligned} J(\theta') - J(\theta) &= J(\theta') - E_{\mathbf{s}_0 \sim p(\mathbf{s}_0)} [V^{\pi_{\theta}}(\mathbf{s}_0)] \\ &= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)} [V^{\pi_{\theta}}(\mathbf{s}_0)] \end{aligned}$$

$$\text{claim: } J(\theta') - J(\theta) = E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_t \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right]$$

$$\begin{aligned} &= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_{t=0}^{\infty} \gamma^t V^{\pi_{\theta}}(\mathbf{s}_t) - \sum_{t=1}^{\infty} \gamma^t V^{\pi_{\theta}}(\mathbf{s}_t) \right] \\ &= J(\theta') + E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_{t=0}^{\infty} \gamma^t (\gamma V^{\pi_{\theta}}(\mathbf{s}_{t+1}) - V^{\pi_{\theta}}(\mathbf{s}_t)) \right] \\ &= E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right] + E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_{t=0}^{\infty} \gamma^t (\gamma V^{\pi_{\theta}}(\mathbf{s}_{t+1}) - V^{\pi_{\theta}}(\mathbf{s}_t)) \right] \\ &= E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_{t=0}^{\infty} \gamma^t (r(\mathbf{s}_t, \mathbf{a}_t) + \gamma V^{\pi_{\theta}}(\mathbf{s}_{t+1}) - V^{\pi_{\theta}}(\mathbf{s}_t)) \right] \\ &= E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_{t=0}^{\infty} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \end{aligned}$$

# Policy gradient as policy iteration

$$J(\theta') - J(\theta) = E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_t \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right]$$

expectation under  $\pi_{\theta'}$

advantage under  $\pi_\theta$

$$E_{\tau \sim p_{\theta'}(\tau)} \left[ \sum_t \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] = \sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)} \left[ \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]$$

$$= \sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]$$

is it OK to use  $p_\theta(\mathbf{s}_t)$  instead?

importance sampling

$$\begin{aligned} E_{x \sim p(x)}[f(x)] &= \int p(x) f(x) dx \\ &= \int \frac{q(x)}{q(x)} p(x) f(x) dx \\ &= \int q(x) \frac{p(x)}{q(x)} f(x) dx \\ &= E_{x \sim q(x)} \left[ \frac{p(x)}{q(x)} f(x) \right] \end{aligned}$$

# Ignoring distribution mismatch?

$$\sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] \stackrel{?}{\approx} \underbrace{\sum_t E_{\mathbf{s}_t \sim p_{\theta}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]}_{\bar{A}(\theta')}$$

why do we want this to be true?

$$J(\theta') - J(\theta) \approx \bar{A}(\theta') \quad \Rightarrow \quad \theta' \leftarrow \arg \max_{\theta'} \bar{A}(\theta)$$

2. Use  $\hat{A}^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$  to get *improved* policy  $\pi'$

is it true? and when?

Claim:  $p_{\theta}(\mathbf{s}_t)$  is *close* to  $p_{\theta'}(\mathbf{s}_t)$  when  $\pi_{\theta}$  is *close* to  $\pi_{\theta'}$

# Bounding the Distribution Change

# Ignoring distribution mismatch?

$$\sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] \stackrel{?}{\approx} \underbrace{\sum_t E_{\mathbf{s}_t \sim p_{\theta}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]}_{\bar{A}(\theta')}$$

why do we want this to be true?

$$J(\theta') - J(\theta) \approx \bar{A}(\theta') \quad \Rightarrow \quad \theta' \leftarrow \arg \max_{\theta'} \bar{A}(\theta)$$

2. Use  $\hat{A}^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$  to get *improved* policy  $\pi'$

is it true? and when?

Claim:  $p_{\theta}(\mathbf{s}_t)$  is *close* to  $p_{\theta'}(\mathbf{s}_t)$  when  $\pi_{\theta}$  is *close* to  $\pi_{\theta'}$



# Bounding the distribution change

Claim:  $p_\theta(\mathbf{s}_t)$  is *close* to  $p_{\theta'}(\mathbf{s}_t)$  when  $\pi_\theta$  is *close* to  $\pi_{\theta'}$

Simple case: assume  $\pi_\theta$  is a *deterministic* policy  $\mathbf{a}_t = \pi_\theta(\mathbf{s}_t)$

$\pi_{\theta'}$  is *close* to  $\pi_\theta$  if  $\pi_{\theta'}(\mathbf{a}_t \neq \pi_\theta(\mathbf{s}_t) | \mathbf{s}_t) \leq \epsilon$

$$p_{\theta'}(\mathbf{s}_t) = \underbrace{(1 - \epsilon)^t}_{\text{probability we made no mistakes}} p_\theta(\mathbf{s}_t) + (1 - (1 - \epsilon)^t) \underbrace{p_{\text{mistake}}(\mathbf{s}_t)}_{\text{some } \textit{other} \text{ distribution}}$$

**seem familiar?**

$$|p_{\theta'}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| = (1 - (1 - \epsilon)^t) |p_{\text{mistake}}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| \leq 2(1 - (1 - \epsilon)^t)$$

$$\text{useful identity: } (1 - \epsilon)^t \geq 1 - \epsilon t \text{ for } \epsilon \in [0, 1] \leq 2\epsilon t$$

**not a great bound, but a bound!**

# Bounding the distribution change

Claim:  $p_\theta(\mathbf{s}_t)$  is *close* to  $p_{\theta'}(\mathbf{s}_t)$  when  $\pi_\theta$  is *close* to  $\pi_{\theta'}$

General case: assume  $\pi_\theta$  is an arbitrary distribution

$\pi_{\theta'}$  is *close* to  $\pi_\theta$  if  $|\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) - \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)| \leq \epsilon$  for all  $\mathbf{s}_t$

Useful lemma: if  $|p_X(x) - p_Y(x)| = \epsilon$ , exists  $p(x, y)$  such that  $p(x) = p_X(x)$  and  $p(y) = p_Y(y)$  and  $p(x = y) = 1 - \epsilon$

$\Rightarrow p_X(x)$  “agrees” with  $p_Y(y)$  with probability  $\epsilon$

$\Rightarrow \pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)$  takes a different action than  $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$  with probability at most  $\epsilon$

$$\begin{aligned} |p_{\theta'}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| &= (1 - (1 - \epsilon)^t) |p_{\text{mistake}}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| \leq 2(1 - (1 - \epsilon)^t) \\ &\leq 2\epsilon t \end{aligned}$$

# Bounding the objective value

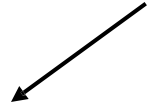
$\pi_{\theta'}$  is *close* to  $\pi_{\theta}$  if  $|\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) - \pi_{\theta}(\mathbf{a}_t|\mathbf{s}_t)| \leq \epsilon$  for all  $\mathbf{s}_t$

$$|p_{\theta'}(\mathbf{s}_t) - p_{\theta}(\mathbf{s}_t)| \leq 2\epsilon t$$

$$\begin{aligned} E_{p_{\theta'}(\mathbf{s}_t)}[f(\mathbf{s}_t)] &= \sum_{\mathbf{s}_t} p_{\theta'}(\mathbf{s}_t) f(\mathbf{s}_t) \geq \sum_{\mathbf{s}_t} p_{\theta}(\mathbf{s}_t) f(\mathbf{s}_t) - |p_{\theta}(\mathbf{s}_t) - p_{\theta'}(\mathbf{s}_t)| \max_{\mathbf{s}_t} f(\mathbf{s}_t) \\ &\geq E_{p_{\theta}(\mathbf{s}_t)}[f(\mathbf{s}_t)] - 2\epsilon t \max_{\mathbf{s}_t} f(\mathbf{s}_t) \end{aligned}$$

$$\begin{aligned} \sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_{\theta}(\mathbf{a}_t|\mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta}(\mathbf{a}_t|\mathbf{s}_t)} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] &\geq \\ \sum_t E_{\mathbf{s}_t \sim p_{\theta}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_{\theta}(\mathbf{a}_t|\mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta}(\mathbf{a}_t|\mathbf{s}_t)} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] &- \sum_t 2\epsilon t C \end{aligned}$$

$O(T r_{\max})$  or  $O\left(\frac{r_{\max}}{1-\gamma}\right)$



maximizing this maximizes a bound on the thing we want!

# Where are we at so far?

$$\theta' \leftarrow \arg \max_{\theta'} \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]$$

such that  $|\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) - \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)| \leq \epsilon$

for small enough  $\epsilon$ , this is guaranteed to improve  $J(\theta') - J(\theta)$

# Policy Gradients with Constraints

# A more convenient bound

Claim:  $p_\theta(\mathbf{s}_t)$  is *close* to  $p_{\theta'}(\mathbf{s}_t)$  when  $\pi_\theta$  is *close* to  $\pi_{\theta'}$

$\pi_{\theta'}$  is *close* to  $\pi_\theta$  if  $|\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) - \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)| \leq \epsilon$  for all  $\mathbf{s}_t$

$$|p_{\theta'}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| \leq 2\epsilon t$$

a more convenient bound:  $|\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) - \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)| \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) \parallel \pi_\theta(\mathbf{a}_t|\mathbf{s}_t))}$

$\Rightarrow D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) \parallel \pi_\theta(\mathbf{a}_t|\mathbf{s}_t))$  bounds state marginal difference

$$D_{\text{KL}}(p_1(x) \parallel p_2(x)) = E_{x \sim p_1(x)} \left[ \log \frac{p_1(x)}{p_2(x)} \right]$$

KL divergence has some very convenient properties that make it much easier to approximate!

# How do we optimize the objective?

$$\theta' \leftarrow \arg \max_{\theta'} \sum_t E_{\mathbf{s}_t \sim p_{\theta}(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]$$

$$\text{such that } D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)) \leq \epsilon$$

for small enough  $\epsilon$ , this is guaranteed to improve  $J(\theta') - J(\theta)$

# How do we enforce the constraint?

$$\theta' \leftarrow \arg \max_{\theta'} \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]$$

such that  $D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) \parallel \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)) \leq \epsilon$

$$\mathcal{L}(\theta', \lambda) = \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] - \lambda (D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) \parallel \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)) - \epsilon)$$

1. Maximize  $\mathcal{L}(\theta', \lambda)$  with respect to  $\theta'$   **can do this incompletely (for a few grad steps)**
2.  $\lambda \leftarrow \lambda + \alpha (D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) \parallel \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)) - \epsilon)$

Intuition: raise  $\lambda$  if constraint violated too much, else lower it  
an instance of *dual gradient descent* (more on this later!)



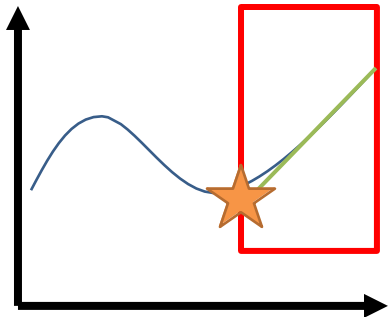
# Natural Gradient

# How (else) do we optimize the objective?

$$\theta' \leftarrow \arg \max_{\theta'} \overbrace{\sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]}^{\bar{A}(\theta')}$$

such that  $D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)) \leq \epsilon$

for small enough  $\epsilon$ , this is guaranteed to improve  $J(\theta') - J(\theta)$



$$\theta' \leftarrow \arg \max_{\theta'} \nabla_\theta \bar{A}(\theta)^T (\theta' - \theta)$$

such that  $D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)) \leq \epsilon$

**Use first order Taylor approximation for objective (a.k.a., linearization)**

# How do we optimize the objective?

$$\theta' \leftarrow \arg \max_{\theta'} \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]$$

such that  $D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)) \leq \epsilon$

$$\theta' \leftarrow \arg \max_{\theta'} \nabla_\theta \bar{A}(\theta)^T (\theta' - \theta)$$

such that  $D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)) \leq \epsilon$

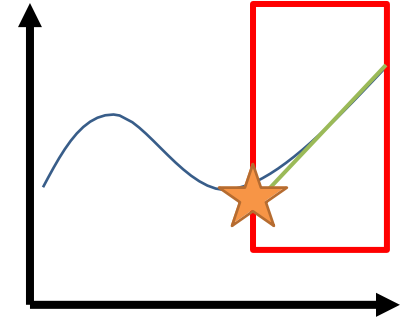
$$\nabla_{\theta'} \bar{A}(\theta') = \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[ \frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t \nabla_{\theta'} \log \pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]$$

**(see policy gradient lecture for derivation)**

$$\nabla_\theta \bar{A}(\theta) = \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[ \cancel{\frac{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)}} \gamma^t \nabla_\theta \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t) A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right]$$

$$\nabla_\theta \bar{A}(\theta) = \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[ E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[ \gamma^t \nabla_\theta \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t) A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] = \nabla_\theta J(\theta)$$

**exactly the normal policy gradient!**



# Can we just use the gradient then?

$$\theta' \leftarrow \arg \max_{\theta'} \nabla_{\theta} J(\theta)^T (\theta' - \theta)$$

$$\text{such that } D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)) \leq \epsilon$$

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta)$$

$$\pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)$$

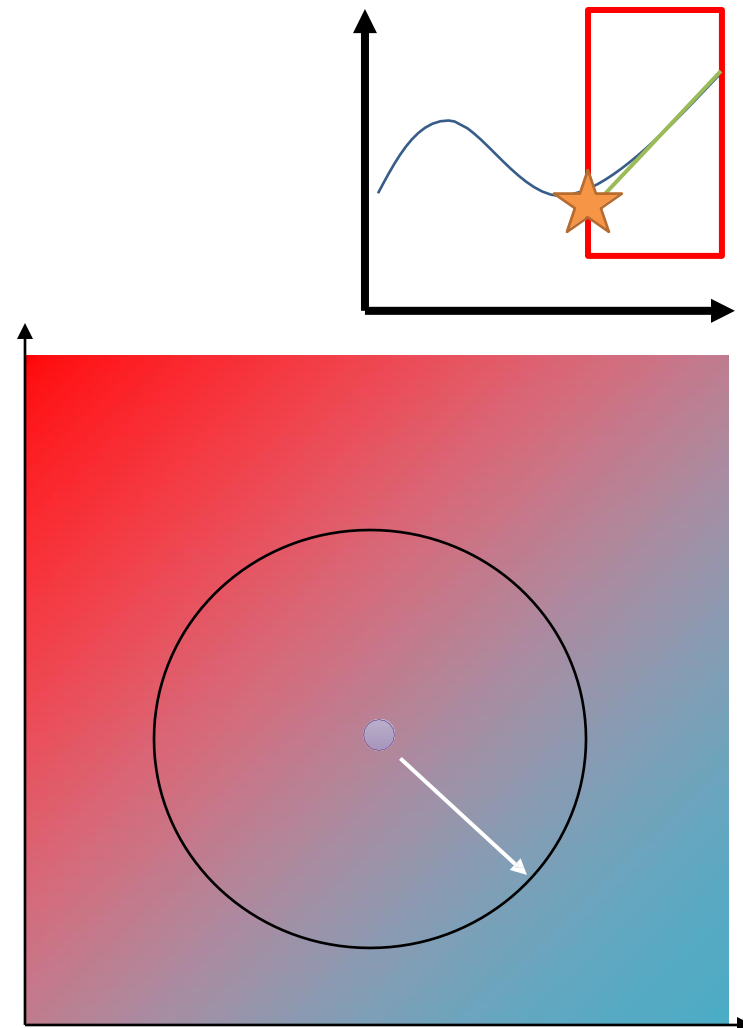
some parameters change probabilities a lot more than others!

Claim: gradient ascent does this:

$$\theta' \leftarrow \arg \max_{\theta'} \nabla_{\theta} J(\theta)^T (\theta' - \theta)$$

$$\text{such that } \|\theta - \theta'\|^2 \leq \epsilon$$

$$\theta' = \theta + \sqrt{\frac{\epsilon}{\|\nabla_{\theta} J(\theta)\|^2}} \nabla_{\theta} J(\theta)$$



# Can we just use the gradient then?

$$\theta' \leftarrow \arg \max_{\theta'} \nabla_{\theta} J(\theta)^T (\theta' - \theta)$$

such that  $D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) \parallel \pi_{\theta}(\mathbf{a}_t|\mathbf{s}_t)) \leq \epsilon$



not the same!

$$\theta' \leftarrow \arg \max_{\theta'} \nabla_{\theta} J(\theta)^T (\theta' - \theta)$$

such that  $\|\theta - \theta'\|^2 \leq \epsilon$

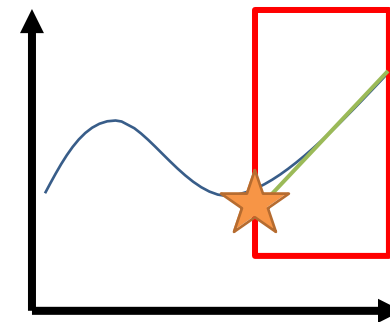
second order Taylor expansion

$$D_{\text{KL}}(\pi_{\theta'} \parallel \pi_{\theta}) \approx \frac{1}{2} (\theta' - \theta)^T \mathbf{F} (\theta' - \theta)$$

Fisher-information matrix

$$\mathbf{F} = E_{\pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(\mathbf{a}|\mathbf{s}) \nabla_{\theta} \log \pi_{\theta}(\mathbf{a}|\mathbf{s})^T]$$

can estimate with samples



# Can we just use the gradient then?

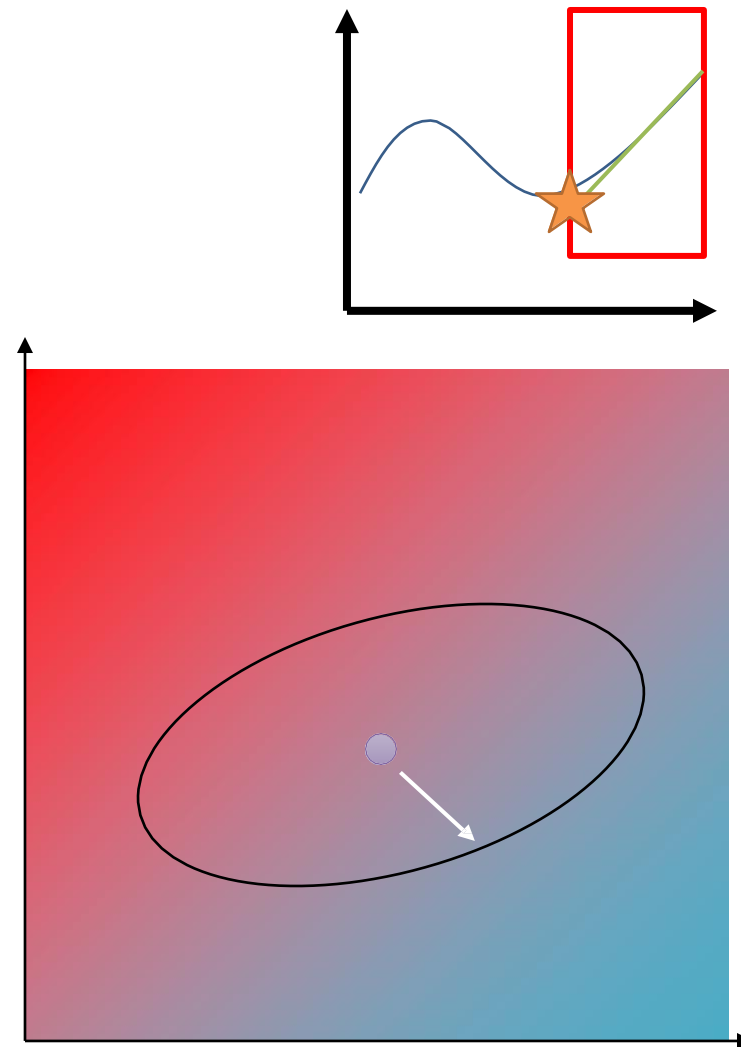
$$\theta' \leftarrow \arg \max_{\theta'} \nabla_{\theta} J(\theta)^T (\theta' - \theta)$$

$$\text{such that } D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)) \leq \epsilon$$

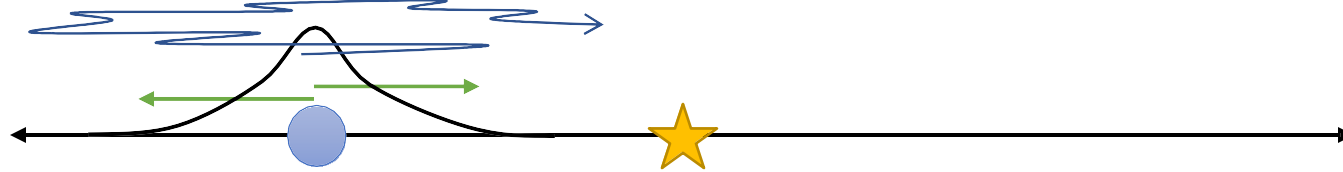
$$D_{\text{KL}}(\pi_{\theta'} \| \pi_{\theta}) \approx \frac{1}{2} (\theta' - \theta)^T \mathbf{F} (\theta' - \theta)$$

$$\theta' = \theta + \alpha \mathbf{F}^{-1} \nabla_{\theta} J(\theta) \quad \leftarrow \text{natural gradient}$$

$$\alpha = \sqrt{\frac{2\epsilon}{\nabla_{\theta} J(\theta)^T \mathbf{F} \nabla_{\theta} J(\theta)}}$$



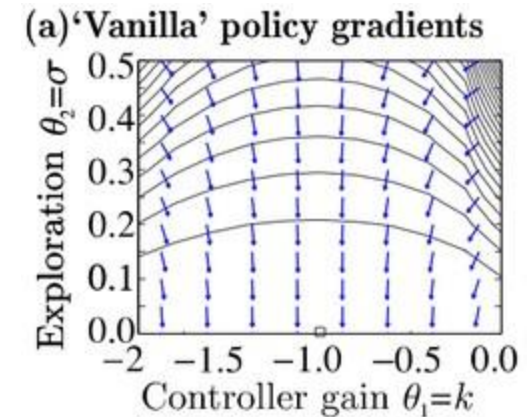
# Is this even a problem in practice?



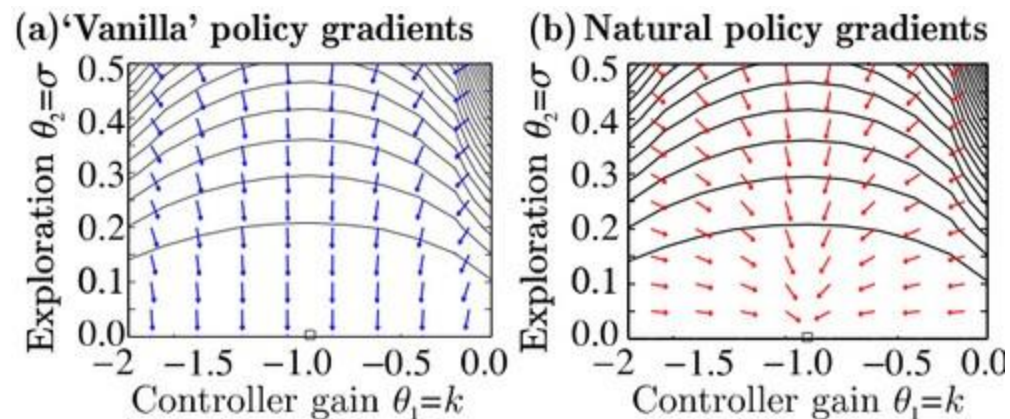
$$r(s_t, a_t) = -s_t^2 - a_t^2$$

$$\log \pi_{\theta}(a_t|s_t) = -\frac{1}{2\sigma^2}(ks_t - a_t)^2 + \text{const}$$

$$\theta = (k, \sigma)$$

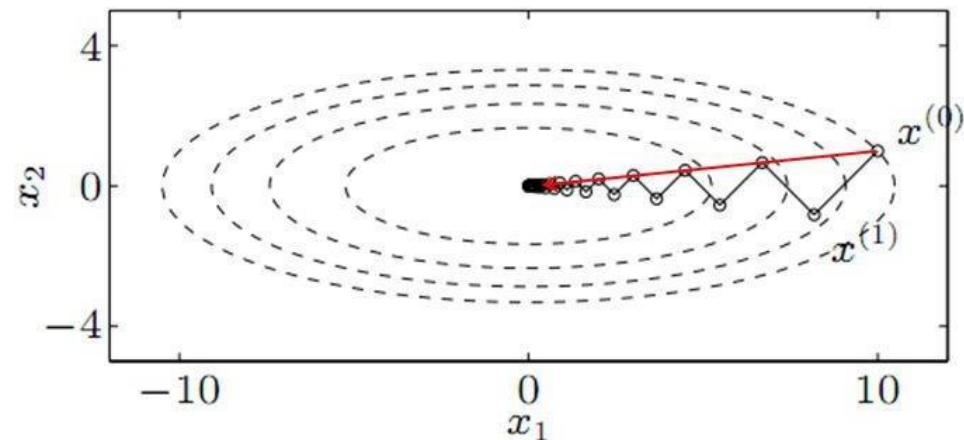


(image from Peters & Schaal 2008)



(figure from Peters & Schaal 2008)

Essentially the same problem as this:



# Practical methods and notes

- Natural policy gradient

$$\theta' = \theta + \alpha \mathbf{F}^{-1} \nabla_{\theta} J(\theta)$$

- Generally a good choice to stabilize policy gradient training
- See this paper for details:
  - Peters, Schaal. Reinforcement learning of motor skills with policy gradients.
- Practical implementation: requires efficient Fisher-vector products, a bit non-trivial to do without computing the full matrix
  - See: Schulman et al. Trust region policy optimization

- Trust region policy optimization

$$\alpha = \sqrt{\frac{2\epsilon}{\nabla_{\theta} J(\theta)^T \mathbf{F} \nabla_{\theta} J(\theta)}}$$

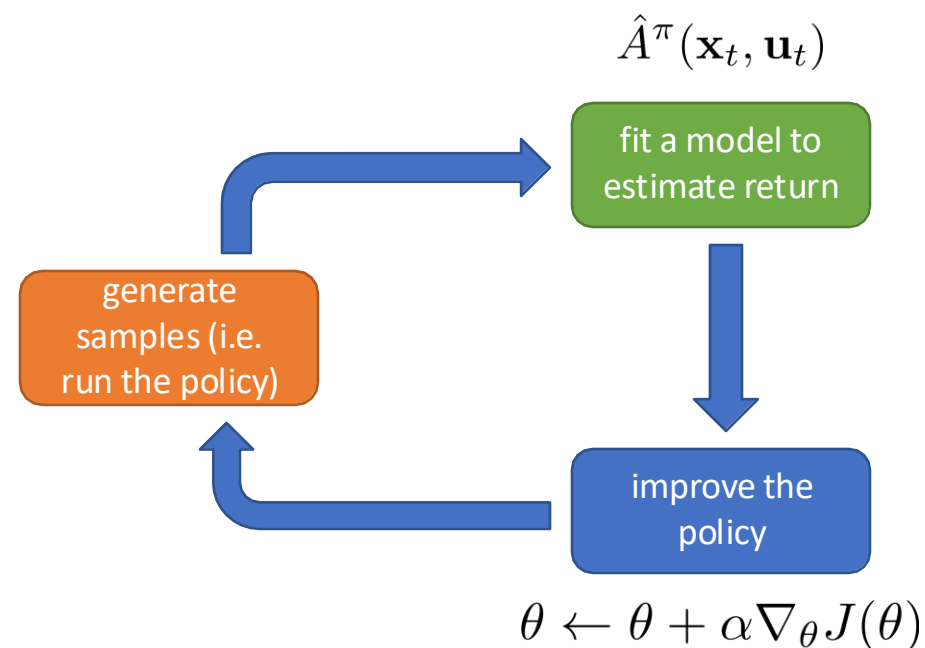
- Just use the IS objective directly

- Use regularization to stay close to old policy
- See: Proximal policy optimization



# Review

- Policy gradient = policy iteration
- Optimize advantage under new policy state distribution
- Using old policy state distribution optimizes a bound, *if* the policies are close enough
- Results in *constrained* optimization problem
- First order approximation to objective = gradient ascent
- Regular gradient ascent has the wrong constraint, use natural gradient
- Practical algorithms
  - Natural policy gradient
  - Trust region policy optimization
  - Proximal Policy Optimization



# Acknowledgements

Slides adapted from

CS 188 UC Berkeley

Pieter Abbeel, Dan Klein et al.

CS 285 UC Berkeley

Sergey Levine

CSC 498 Univ of Toronto

Animesh Garg