

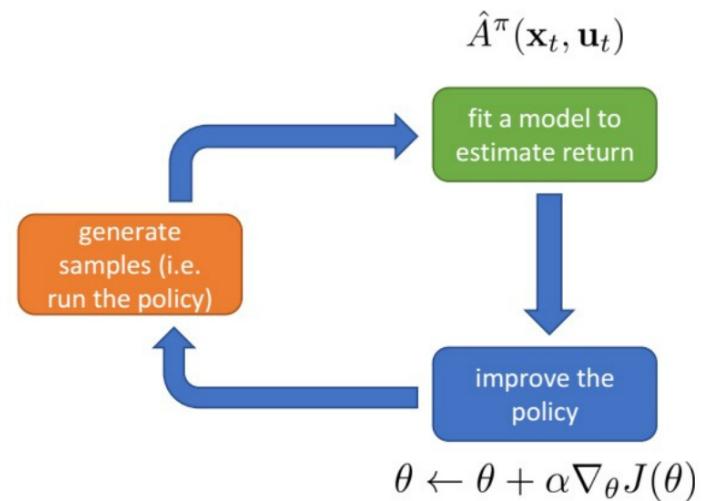
و π میں کسی s کے لئے a کا انتخاب $\pi(a|s)$ کے مطابق ہے تو π On-policy π کے نام سے جانا جاتا ہے۔

Policy Gradient as Policy Iteration

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

main steps of policy gradient algorithm:

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



Familiar to policy iteration algorithm:

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

Policy Gradient as Policy Iteration

$$\theta \rightarrow \theta'$$

μ, ν $J(\theta) = E_{\tau \sim p_\theta(\tau)} \left[\sum_t \gamma^t r(s_t, a_t) \right]$

return

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

could be interpreted as policy improvement!

$$\max_{\theta'} J(\theta') - J(\theta) \quad \text{def} \quad E_{\tau \sim P_{\theta'}} \left[\sum_t \underbrace{\gamma^t A^{\pi_\theta}(s_t, a_t)}_{r(s_t, a_t) + \gamma V^\pi(s_{t+1})} \right] - V^{\pi_\theta}(s_t)$$

کاریکاتوریستیک رلیوں کے لئے

نهایی مکانیزم برای آنچه که در پیش از آن مذکور شد
باشد که در اینجا اینجا را با $\text{PI} \leftarrow \text{PG}$ نویسید

Policy Gradient as Policy Iteration

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

proof:
$$\begin{aligned} J(\theta') - J(\theta) &= J(\theta') - E_{s_0 \sim p(s_0)} [V^{\pi_\theta}(s_0)] \\ &= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)} [V^{\pi_\theta}(s_0)] \\ &= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t V^{\pi_\theta}(s_t) - \sum_{t=1}^{\infty} \gamma^t V^{\pi_\theta}(s_t) \right] \\ &= J(\theta') + E_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t (\gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t)) \right] \\ &= E_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] + E_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t (\gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t)) \right] \\ &= E_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + \gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t)) \right] \\ &= E_{\tau \sim p_{\theta'}(\tau)} \left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] = E_{\tau \sim P_\theta(\tau)} \left[\frac{P_\theta(\tau)}{P_\theta(\tau)} \sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \end{aligned}$$

Objective RL

$J(\theta)$ با $E_{s \sim p(s)}$ بازگشتی و غیر قابل تنبیه

با $E_{s \sim p(s)}$ بازگشتی و غیر قابل تنبیه

با $E_{s \sim p(s)}$ بازگشتی و غیر قابل تنبیه

با $E_{s \sim p(s)}$ بازگشتی و غیر قابل تنبیه

با $E_{s \sim p(s)}$ بازگشتی و غیر قابل تنبیه

با $E_{s \sim p(s)}$ بازگشتی و غیر قابل تنبیه

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 π_θ

$\pi_\theta = \pi_{\theta'} \Rightarrow P_\theta \equiv P_{\theta'}$

$J(\theta') - J(\theta)$

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

P_θ

importance weight

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

توضیحات مربوطه
متوجه شدن از فرق میان P_θ و $P_{\theta'}$

تحلیل $\pi_\theta = \pi_{\theta'}$

Policy Gradient as Policy Iteration

$$J(\theta') - J(\theta)$$

Can we ignore distribution mismatch?

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

why do we want this to be true?

initial Goal

$$J(\theta') - J(\theta) \approx \bar{A}(\theta') \Rightarrow \underline{\theta'} \leftarrow \arg \max_{\theta'} \bar{A}(\theta') \quad 2.$$

$\bar{A}(\theta')$ متوسط انتاج Advantageal

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

is it true? and when?

$p_\theta(s_t)$ is close to $p_{\theta'}(s_t)$ when π_θ is close to $\pi_{\theta'}$

Bounding the distribution change

Claim: $p_{\theta}(s_t)$ is close to $p_{\theta'}(s_t)$ when π_{θ} is close to $\pi_{\theta'}$

Simple case: assume π_{θ} is a deterministic policy $a_t = \pi_{\theta}(s_t)$

$\pi_{\theta'}$ is close to π_{θ} if $\pi_{\theta'}(a_t \neq \pi_{\theta}(s_t) | s_t) \leq \epsilon$

$$p_{\theta'}(s_t) = (1 - \epsilon)^t p_{\theta}(s_t) + (1 - (1 - \epsilon)^t) p_{\text{mistake}}(s_t)$$

probability we made no mistakes

$$P_{\theta}(s_t) = (1 - \epsilon)^t P_{\theta}(s_t) + \text{some other distribution}$$

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

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

Total Variation distance ($P_{\theta}, P_{\theta'}$)
= $\sum_{\tau} |P_{\theta}(\tau) - P_{\theta'}(\tau)|$
seem familiar?

cool stuff

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 π_θ is *close* to $\pi_{\theta'}$

General case: assume π_θ is an arbitrary distribution

$\pi_{\theta'}$ is *close* to π_θ 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 ϵ

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

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

$$\sum_{t=0}^{\infty} \epsilon \rho^t < \infty \quad \rho < 1$$

Bounding the objective value

$$x \geq -|x|$$

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

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

$$E_{p_{\theta'}(\mathbf{s}_t)}[f(\mathbf{s}_t)] = \sum_{\mathbf{s}_t} p_{\theta'}(\mathbf{s}_t) f(\mathbf{s}_t) \geq \sum_{\mathbf{s}_t} p_{\theta}(\mathbf{s}_t) f(\mathbf{s}_t) - |p_{\theta}(\mathbf{s}_t) - p_{\theta'}(\mathbf{s}_t)| \max_{\mathbf{s}_t} |f(\mathbf{s}_t)|$$

$$\sum_t J(\theta') - J(\theta) \geq E_{p_{\theta}(\mathbf{s}_t)}[f(\mathbf{s}_t)] - 2\epsilon t \max_{\mathbf{s}_t} |f(\mathbf{s}_t)|$$

$$\sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)} \left[E_{\mathbf{a}_t \sim \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \left[\frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_{\theta}}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] \geq O(T r_{\max}) \text{ or } O\left(\frac{r_{\max}}{1-\gamma}\right)$$

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

maximizing this maximizes a bound on the thing we want!

Some Useful Preliminaries: Taylor Series

Approximation of a differentiable function around a given point with sum of terms of the function's derivatives:

$$f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2} (x - x_0)^T H(x - x_0) + \dots$$

Some Useful Preliminaries: Constrained Optimization

- equality constraints: method of Lagrange multipliers

$$\begin{aligned} & \text{optimize } f(x) \\ & \text{subject to: } g(x) = 0 \end{aligned} \quad \longrightarrow \quad \mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$$

- Inequality constraints: KKT

$$\begin{aligned} & \text{optimize } f(x) \\ & \text{subject to:} \\ & \quad g_i(x) \leq 0, \\ & \quad h_j(x) = 0 \end{aligned} \quad \longrightarrow \quad \begin{aligned} & \mathcal{L}(x, \mu, \lambda) = f(x) + \mu^T g(x) + \lambda^T h(x) \\ & \text{subject to:} \\ & \quad \mu_i \geq 0, \\ & \quad \mu^T g(x) = 0 \end{aligned}$$

Some Useful Preliminaries: KL-Divergence

A Common distance measure for distributions:

$$D_{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$

Other useful distance measures:

- Total variation distance
- Wasserstein distance
- Jensen–Shannon divergence
- ...

Some Useful Preliminaries: Fisher Information

likelihood function: $p_\theta(x)$

score function: $\nabla_\theta \log p_\theta(x)$

Fisher information: measuring the amount of information that a random variable (x) carries about likelihood parameters (θ):

$$\begin{aligned}[I(\theta)]_{i,j} &= E_{x \sim p_\theta(x)} \left[\left(\frac{\partial}{\partial \theta_i} \log p_\theta(x) \right) \left(\frac{\partial}{\partial \theta_j} \log p_\theta(x) \right) \right] \\ &= -E_{x \sim p_\theta(x)} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(x) \right]\end{aligned}$$

variance (covariance)
of score function

curvature of score function

Where are we at so far?

$$\approx J(\theta') - J(\theta)$$

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

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

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

gradients in $J(\theta')$ w.r.t. $J(\theta)$ are

A more convenient bound

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

$\pi_{\theta'}$ is *close* to π_θ 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: $|\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) - \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)| \leq \sqrt{\frac{1}{2}D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)\|\pi_\theta(\mathbf{a}_t|\mathbf{s}_t))}$

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

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

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

How do we optimize the objective?

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

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 ϵ , this is guaranteed to improve $J(\theta') - J(\theta)$

Dual, Primal $\left\{ \begin{array}{l} \text{Dual} \\ \text{Primal} \end{array} \right\} \leftarrow \text{Natural Gradient } \textcircled{1}$
Gradient Descent $\left\{ \begin{array}{l} \text{Gradient Descent} \\ \text{Primal} \end{array} \right\} \leftarrow \text{Natural Gradient } \textcircled{1}$
Natural Gradient $\left\{ \begin{array}{l} \text{Natural Gradient} \\ \text{Dual} \end{array} \right\} \leftarrow \text{Natural Gradient } \textcircled{2}$
closed form solution $\left\{ \begin{array}{l} \text{closed form solution} \\ \text{Dual} \end{array} \right\} \leftarrow \text{Natural Gradient } \textcircled{2}$

How do we enforce the constraint?

$$\left\{ \begin{array}{l} \theta' \leftarrow \arg \max_{\theta'} \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[\frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] \\ \text{such that } D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)) \leq \epsilon \\ \mathcal{L}(\theta', \lambda) = \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \left[\frac{\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] - \lambda(D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)) - \epsilon) \end{array} \right.$$

$= J(\theta') - J(\theta)$

→ 1. Maximize $\mathcal{L}(\theta', \lambda)$ with respect to θ' ← can do this incompletely (for a few grad steps)

→ 2. $\lambda \leftarrow \lambda + \alpha(D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)) - \epsilon)$ $\lambda \geq 0$

Intuition: raise λ if constraint violated too much, else lower it

an instance of dual gradient descent (more on this later!)

How (else) do we optimize the objective?

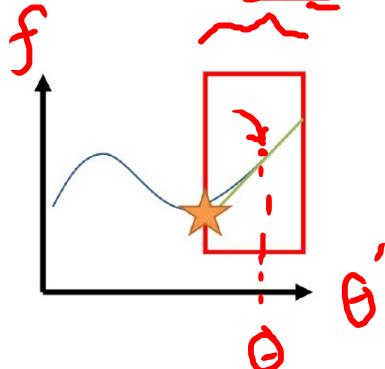
$$\bar{A}(\theta') = \arg \max_{\theta'} \sum_t E_{s_t \sim p_\theta(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t A^{\pi_\theta}(s_t, a_t) \right] \right]$$

such that $D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t)) \leq \epsilon$

$\max f(\theta)$
 $\text{s.t. } g(\theta) \geq 0$

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

Trust Region



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

such that $D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t)) \leq \epsilon$

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

$$f(\theta') = f(\theta) + \nabla f(\theta)^T (\theta' - \theta) + \dots$$

How (else) do we optimize the objective?

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

such that $D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t)) \leq \epsilon$

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

such that $D_{KL}(\pi_{\theta'}(a_t|s_t) \| \pi_\theta(a_t|s_t)) \leq \epsilon$

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

(see policy gradient lecture for derivation)

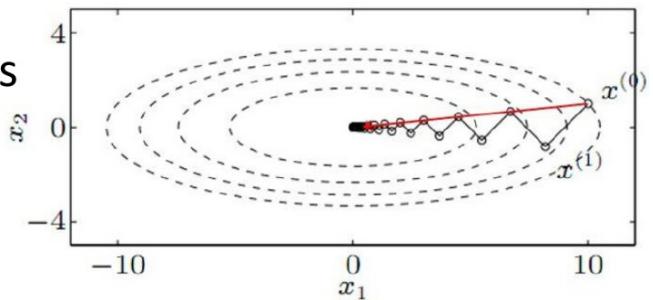
$$\nabla_{\theta} \bar{A}(\theta) = \sum_t E_{s_t \sim p_\theta(s_t)} \left[E_{a_t \sim \pi_\theta(a_t|s_t)} \left[\frac{\pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \gamma^t \nabla_{\theta} \log \pi_\theta(a_t|s_t) A^{\pi_\theta}(s_t, a_t) \right] \right]$$

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

exactly the normal policy gradient!

Importance of step size in RL

- Supervised learning: Step too far \rightarrow next updates will fix it
- Reinforcement learning: Policy is determining data collection!
 - Step too far \rightarrow bad policy
 - Next batch: collected under bad policy
 - May not be able to recover from a bad choice, collapse in performance!
- Learning rate tuning is hard
 - Poor conditioning could be more dangerous in RL settings
 - More sophisticated optimizers can reduce numerical issues
 - Need for advanced learning rate adjustment methods!



Natural Policy Gradient

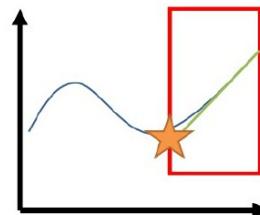
$$\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta)$$

$$\nabla_{\theta} D_{KL}(\pi_\theta \| \pi^*) = 0$$

Could be shown as a constraint problem:

$$\theta' \leftarrow \arg \max_{\theta'} (\theta' - \theta)^T \nabla_{\theta} J(\theta) \text{ s.t. } \|\theta' - \theta\|^2 \leq \epsilon$$

controls how far we go

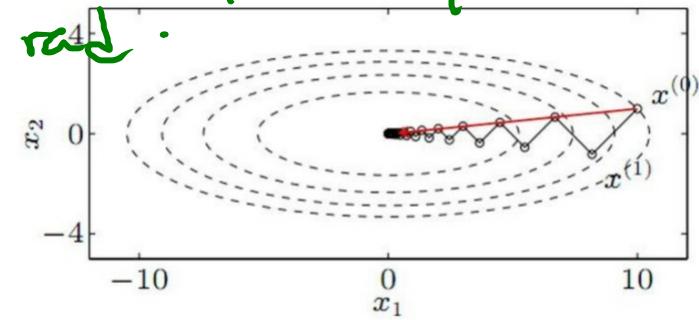


Can we rescale the gradients to reduce the problem with poor conditioning?

$$\theta' \leftarrow \arg \max_{\hat{\theta}'} (\hat{\theta}' - \theta)^T \nabla_{\theta} J(\theta) \text{ s.t. } D(\pi_{\theta'}, \pi_{\theta}) \leq \epsilon$$

parameterization-independent divergence measure

این سیاست را با این تغیرات که این قدر در policy Grad-Lines داریم



Now $f(\theta')$ \rightarrow we want to find $\nabla_{\theta'} f(\theta')$ \equiv the PG w.r.t θ'
 and finally

$$\begin{aligned} \nabla_{\theta'} \overbrace{D_{KL}(\pi_{\theta'} || \pi_{\theta})}^{f(\theta')} &= \nabla_{\theta'} \sum_a \pi_{\theta'}(a) \log \frac{\pi_{\theta'}(a)}{\pi_{\theta}(a)}, \\ &= \sum_a \underbrace{\nabla \pi_{\theta'}(a)}_{0} \underbrace{\log \frac{\pi_{\theta'}}{\pi_{\theta}}} + \underbrace{\pi_{\theta'} \frac{\nabla \pi_{\theta'}}{\pi_{\theta'}}}_{\theta'=\theta} \\ &= \sum_a \nabla \pi_{\theta'} \Big|_{\theta'=\theta} = \nabla_{\theta'} \sum_a \pi_{\theta'}(a) = 0 \end{aligned}$$

$$\begin{aligned} \text{grad w.r.t } \theta' &= \sum_a H_{\pi_{\theta'}} \cdot \log \frac{\pi_{\theta'}}{\pi_{\theta}} + (\nabla \pi_{\theta'}) \cdot \frac{\nabla \pi_{\theta'}}{\pi_{\theta'}} + H_{\pi_{\theta'}} \Big|_{\theta'=\theta}, \\ F &\stackrel{?}{=} \sum_a \frac{\nabla \pi_{\theta'} \nabla \pi_{\theta'}}{\pi_{\theta'}} = H \sum_a \pi_{\theta'}(a) \end{aligned}$$

$$\underline{aa^T} = \underline{i} - \begin{bmatrix} & j_1 \\ & \dots \\ - & a_i a_j \end{bmatrix}$$

$$\left\{ \begin{array}{l} \max_{\theta'} \quad \nabla J^T(\theta' - \theta) \\ \text{s.t.} \quad (\theta' - \theta)^T F (\theta' - \theta) \leq E \end{array} \right.$$

$$\alpha \nabla J^T F^{-1} \nabla J \leq E$$

$$\alpha \leq \sqrt{\frac{E}{\nabla J^T F^{-1} \nabla J}}$$

$$\nabla J^T(\theta' - \theta) - \lambda [(\theta' - \theta)^T F (\theta' - \theta) - E]$$

$$\theta' - \theta = F^{-1} \nabla J$$

$$\theta' = \theta + \alpha F^{-1} \nabla J$$

$$\nabla J|_{\theta'=\theta} - \lambda F (\theta' - \theta) = 0$$

$$\begin{aligned} \underline{\underline{f(\theta')}} &= \underline{\underline{f(\theta)}} + \nabla_{\theta'} f \Big|_{\underline{\theta'} = \theta}^T \overset{\circ}{(}\theta' - \theta\overset{\circ}{)} \\ &+ \frac{1}{2} (\theta' - \theta)^T H_f (\theta' - \theta) + \dots \end{aligned}$$

Natural Policy Gradient

$$\theta' \leftarrow \arg \max_{\theta'} (\theta' - \theta)^T \nabla_{\theta} J(\theta) \text{ s.t. } D(\pi_{\theta'}, \pi_{\theta}) \leq \epsilon$$

parameterization-independent divergence measure

usually KL-divergence: $D_{\text{KL}}(\pi_{\theta'} \| \pi_{\theta}) = E_{\pi_{\theta'}} [\log \pi_{\theta} - \log \pi_{\theta'}]$

Taylor expansion:

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

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

$$\begin{aligned} \mathbf{F} &= \sum_a \frac{\nabla_{\theta} \pi_{\theta} \nabla_{\theta} \pi_{\theta}^T}{\pi_{\theta}} \nabla_{\theta} \log \pi_{\theta} \\ &= \sum_a \frac{(\nabla_{\theta} \pi_{\theta}) (\nabla_{\theta} \pi_{\theta})^T}{\pi_{\theta}} \cdot \frac{\nabla_{\theta} \log \pi_{\theta}}{\pi_{\theta}} \end{aligned}$$

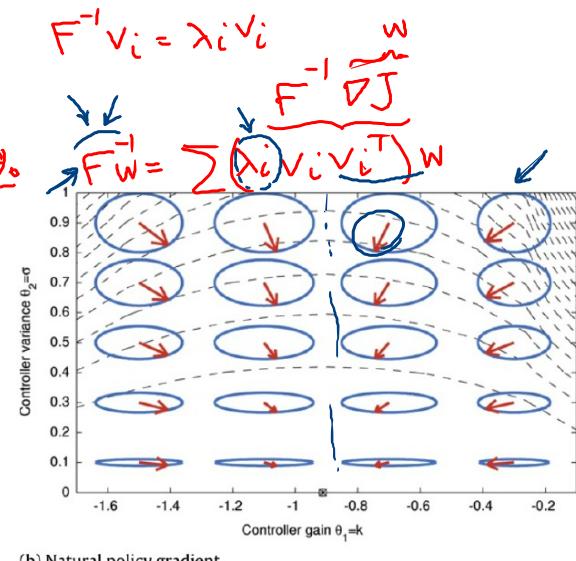
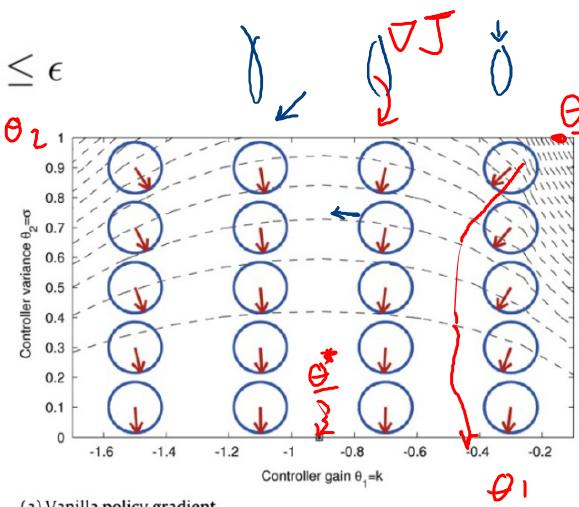
Natural Policy Gradient

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

$$\theta' \leftarrow \arg \max_{\theta'} (\theta' - \theta)^T \nabla_{\theta} J(\theta) \text{ s.t. } D(\pi_{\theta'}, \pi_{\theta}) \leq \epsilon$$

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

natural gradient: pick α



trust region policy optimization: pick ϵ

can solve for optimal α while solving $\mathbf{F}^{-1} \nabla_{\theta} J(\theta)$

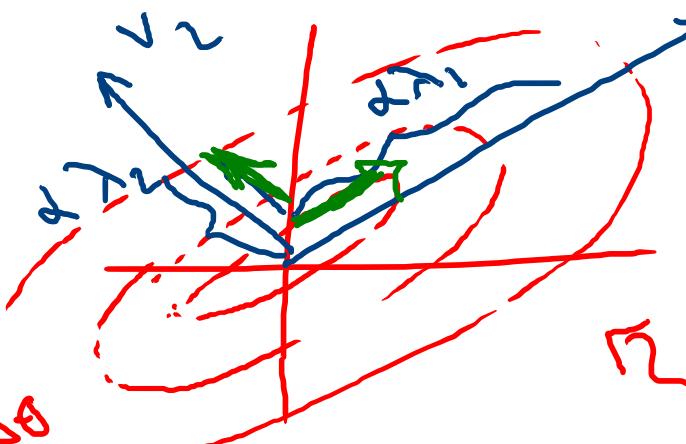
(figure from Peters & Schaal 2008)

جوابیه Conjugate prior $\propto \exp(-\frac{1}{2} \theta^T \mathbf{F} \theta)$

$$x^T A \propto$$

$$A \triangleleft_0$$

$$\Delta \theta F^{-1} = \sum \frac{1}{\lambda_i} v_i v_i^T \Delta \theta$$



$$\Delta \theta \propto \nabla_{\theta} J$$

$$D_{KL}(\pi_{\theta'} || \pi_{\theta})$$

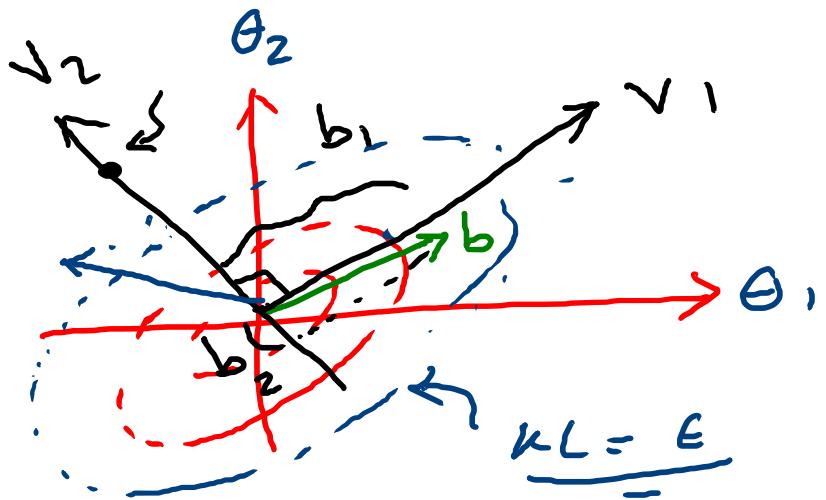
$$\max_{\theta'} \quad (\theta' - \theta)^T \nabla_{\theta} J$$

s.t.

$$(\theta' - \theta)^T F (\theta' - \theta) \leq \epsilon$$

assume that $F = I \Rightarrow \|\theta' - \theta\|^2 \leq \epsilon$

$$(\theta' \leftarrow \theta + \alpha \nabla J \quad \alpha = \sqrt{\frac{2\epsilon}{\|\nabla J\|^2}})$$



$$\langle \nabla J, \Delta \theta \rangle_{st. D_{KL}} \leq \epsilon \quad \frac{KL}{\epsilon} = \frac{E}{\lambda_1 + \frac{\lambda_2}{\lambda_2}}$$

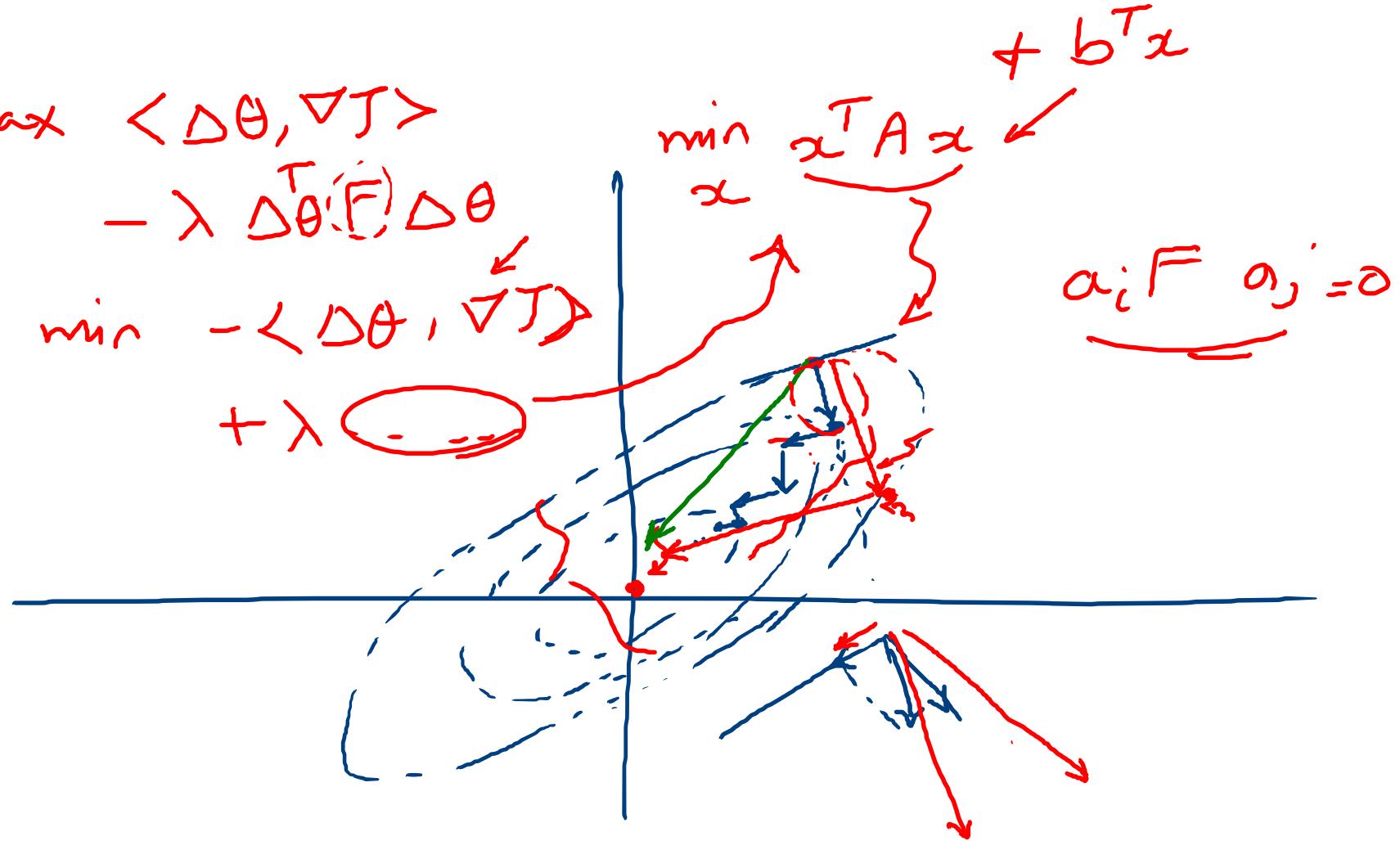
$$\Delta \theta = \alpha F \nabla J$$

∇J

$$\max \langle \Delta\theta, \nabla J \rangle$$

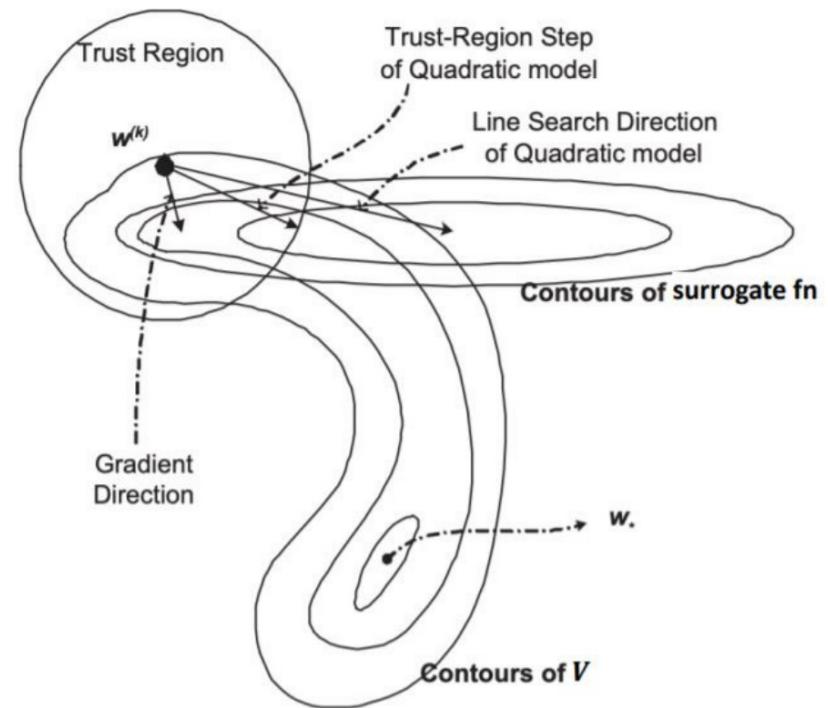
$$-\lambda \Delta\theta^T F \Delta\theta$$

$$= \min -\langle \Delta\theta, \nabla J \rangle$$



Trust Region Method

- We often optimize a surrogate objective
- Surrogate objective may be trustable only in a small region
- **Limit search to small trust region**



Cs885, waterloo 2022

Trust Region Policy Optimization

Recall from “Policy Gradient as Policy Iteration”:

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

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

surrogate loss

trust region

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

Policy Gradient with Constraints

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

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

1. Maximize $\mathcal{L}(\theta', \lambda)$ with respect to θ'
2. $\lambda \leftarrow \lambda + \alpha(D_{\text{KL}}(\pi_{\theta'}(\mathbf{a}_t | \mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t | \mathbf{s}_t)) - \epsilon)$

Intuition: raise λ if constraint violated too much, else lower it
an instance of *dual gradient descent*

Natural Gradient Based on Trust Region

Recall:

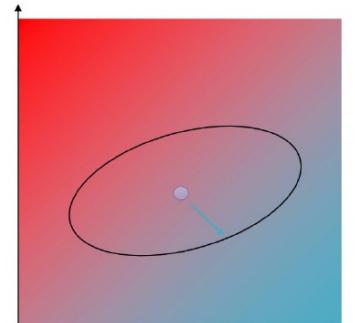
$$\theta' \leftarrow \arg \max_{\theta'} \nabla_{\theta} J(\theta)^T (\theta' - \theta)$$

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

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

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

natural gradient



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

Proximal Policy Optimization

- TRPO is conceptually and computationally challenging in large part because of the constraint in the optimization.

$$D_{KL}(\pi_{\theta'}(\cdot|s) || \pi_{\theta}(\cdot|s)) \leq \epsilon \rightarrow \text{محدود، معرض، این}$$

- What is the effect of the constraint?
- Recall KL-Divergence:

$$D_{KL}(\pi_{\theta'}(\cdot|s) || \pi_{\theta}(\cdot|s)) = \sum_a \pi_{\theta'}(a|s) \log \frac{\pi_{\theta'}(a|s)}{\pi_{\theta}(a|s)}$$

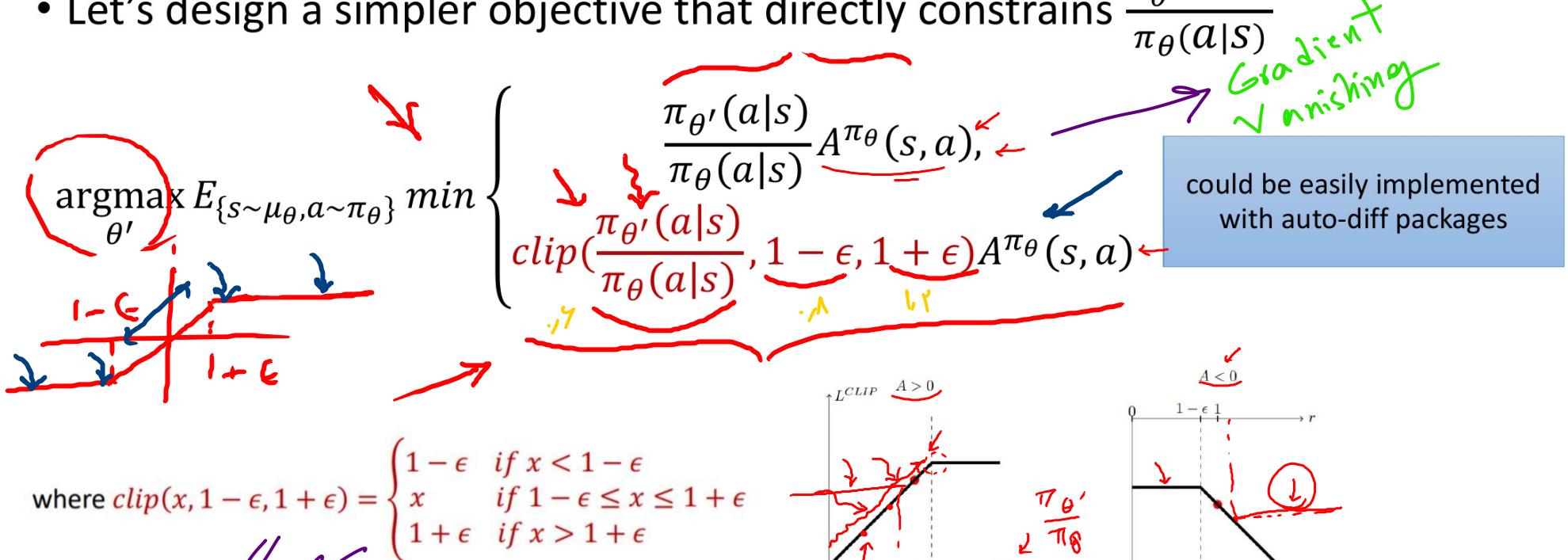
We are effectively constraining the ratio

$$\frac{\pi_{\theta'}(a|s)}{\pi_{\theta}(a|s)}$$

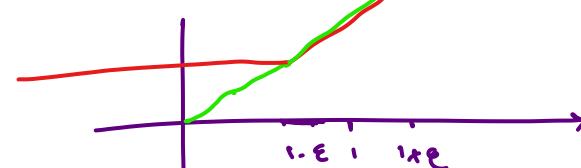
Proximal Policy Optimization

$$J(\theta') - J(\theta) = \sum_t \frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)} A^{\pi_\theta}$$

- Let's design a simpler objective that directly constrains $\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}$



gradient vanishing



$A >$