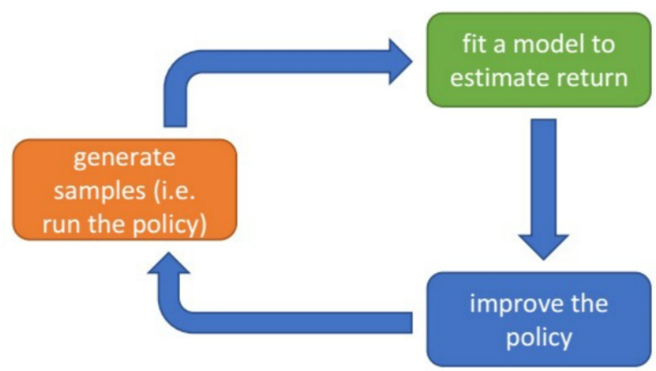
Policy Gradient as Policy Iteration

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

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'

$\hat{A}^\pi(\mathbf{x}_t, \mathbf{u}_t)$



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

Familiar to policy iteration algorithm:

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

Policy Gradient as Policy Iteration

$\theta \longrightarrow \theta'$

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

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

could be interpreted as policy improvement!

$$\theta' = \arg\max_{\theta'} J(\theta') - J(\theta)$$

$$\text{s.t. } d(\theta', \theta) \leq \epsilon$$

$$\max_{\theta'} \underbrace{J(\theta') - J(\theta)} \quad \mathbb{E}_{\tau \sim p_{\theta'}} \left[\sum_t \gamma^t \underbrace{A^{\pi_\theta}(s_t, a_t)} \right]$$

$$r(s_t, a_t) + \gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t)$$

Policy Gradient as Policy Iteration

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

initial state

proof: $J(\theta') - J(\theta) = J(\theta') - E_{\mathbf{s}_0 \sim p(\mathbf{s}_0)}\left[V^{\pi_\theta}(\mathbf{s}_0)\right]$

$$= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)}\left[V^{\pi_\theta}(\mathbf{s}_0)\right]$$

$$= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t V^{\pi_\theta}(\mathbf{s}_t) - \sum_{t=1}^{\infty} \gamma^{\overset{t+1}{t}} V^{\pi_\theta}(\mathbf{s}_{\overset{t+1}{t}})\right]$$

$t=0$

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

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

$$= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t \left(\underbrace{r(\mathbf{s}_t, \mathbf{a}_t) + \gamma V^{\pi_\theta}(\mathbf{s}_{t+1})}_{Q(s_t, a_t)} - V^{\pi_\theta}(\mathbf{s}_t)\right)\right]$$

$$= E_{\tau \sim \overset{p_\theta}{p_{\theta'}}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t)\right] = \mathbb{E}_{\tau \sim p_\theta(\tau)}\left[\frac{p_{\theta'}(\tau)}{p_\theta(\tau)} \sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(s_t, a_t)\right]$$

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

$\pi_\theta = \pi_{\theta'}$ $\Rightarrow$ $p_\theta = p_{\theta'}$

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

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

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

p_θ

importance weight

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

Policy Gradient as Policy Iteration

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

Can we ignore distribution mismatch?

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

why do we want this to be true?

initial Goal

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

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

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

is it true? and when?

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

Bounding the distribution change

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

Total Variation distance $(P_\theta, P_{\theta'})$

$= \sum_\tau |P_\theta(\tau) - P_{\theta'}(\tau)|$

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

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

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

seem familiar?

$$P_\theta(S_t) = (1-\epsilon)^t P_\theta(S_t) + (1-(1-\epsilon)^t) P_\theta(S_t)$$

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

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

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 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

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

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

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 $\mathbf{s}_t$

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

$= \sum (P_{\theta'} - P_\theta + P_\theta) f = \sum P_\theta f + \sum (P_{\theta'} - P_\theta) f$

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

x

$$\geq E_{p_\theta(\mathbf{s}_t)}[f(\mathbf{s}_t)] - 2\epsilon t \max_{\mathbf{s}_t} |f(\mathbf{s}_t)|$$

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

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

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

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

$x \geq -|x| = -\sum |P_{\theta'} - P_\theta| \, |f|$

$\geq -\left(\sum |P_{\theta'} - P_\theta|\right) \max |f|$

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:

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

Some Useful Preliminaries: Constrained Optimization

- equality constraints: method of Lagrange multipliers

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

- Inequality constraints: KKT

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

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):

$$[I(\theta)]_{i,j} = E_{x\sim p_\theta(x)}[\,(\frac{\partial}{\partial\theta_i} log p_\theta(x))(\frac{\partial}{\partial\theta_j} log p_\theta(x))]$$

$$= -E_{x\sim p_\theta(x)}[\,\frac{\partial^2}{\partial\theta_i\partial\theta_j} log p_\theta(x)]$$

variance (covariance) of score function

curvature of score function

Where are we at so far?

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

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

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

Tot. Var.

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

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: $|\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) - \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)| \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) \| \pi_\theta(\mathbf{a}_t|\mathbf{s}_t))}$

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

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

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

How do we optimize the objective?

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

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

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

How do we enforce the constraint?

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

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

$$\max_{\theta} \min_{\lambda} \mathcal{L}(\theta', \lambda)$$

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

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

$\lambda \geq 0$

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?

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

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

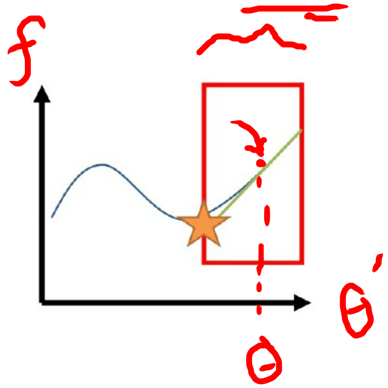
for small enough ϵ, this is guaranteed to improve $J(\theta') - J(\theta)$

$\max f(\theta)$

s.t. $g(\theta) \geq 0$

linearize

Trust Region



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

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

$$f(\theta') = f(\theta) + \nabla f \Big|_{\theta'=\theta}^{T} (\theta' - \theta) + \ldots$$

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

How (else) do we optimize the objective?

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

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

A

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

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

$\nabla_{\theta'} \log \pi_{\theta'}$

$$\frac{\cancel{\pi_{\theta'}}\, \nabla \pi_{\theta'}}{\pi_{\theta'}\, \cancel{\pi_\theta}} \gamma^t A^{\pi_\theta} \Big|_{\theta' = \theta}$$

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

(see policy gradient lecture for derivation)

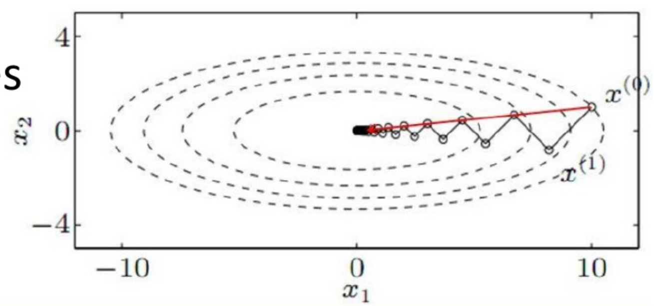
$$\nabla_\theta \bar{A}(\theta) = \sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \left[\cancel{\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}} \gamma^t \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)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

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

exactly the normal policy gradient!

Importance of step size in RL

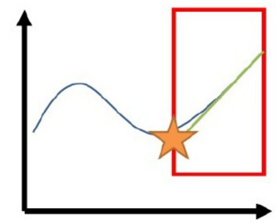
- Supervised learning: Step too far → next updates will fix it
- Reinforcement learning: Policy is determining data collection!
  - Step too far → bad policy
  - Next batch: collected under bad policy
  - May not be able to recover from a bad choice, collapse in performance!
- Learning rate tuning is 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

$\nabla_{\theta'} D_{KL}(\pi_{\theta'} \| \pi_\theta) = 0$

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



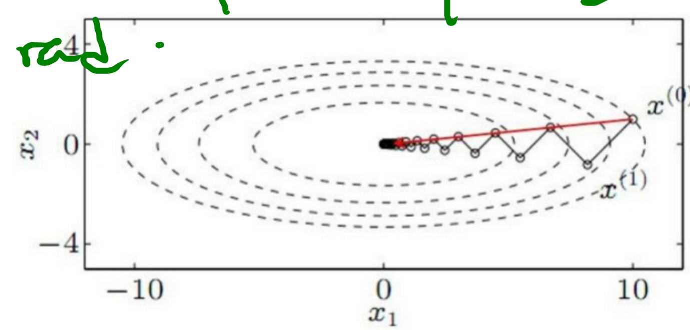
Could be shown as a constraint problem:

$$\theta' \leftarrow \arg\max_{\theta'} (\theta' - \theta)^T \nabla_\theta J(\theta) \text{ s.t. } \underbrace{\|\theta' - \theta\|^2 \leq \epsilon}_{\text{controls how far we go}}$$

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

Gradient That comes from policy grad.

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



$$\nabla_{\theta'} \overbrace{D_{KL}(\pi_{\theta'} \,\|\, \pi_\theta)}^{f(\theta')} = \nabla_{\theta'} \sum_a \pi_{\theta'}(a) \log \frac{\pi_{\theta'}(a)}{\pi_\theta(a)}$$

$$= \sum_a \overbrace{\nabla \pi_{\theta'}(a) \underbrace{\log \frac{\pi_{\theta'}}{\pi_\theta}}_{0}} + \overbrace{\cancel{\pi_{\theta'}} \frac{\nabla \pi_{\theta'}}{\cancel{\pi_{\theta'}}}} \Bigg|_{\theta' = \theta}$$

$$= \sum_a \nabla \pi_{\theta'} \Bigg|_{\theta' = \theta} = \underbrace{\nabla_{\theta'} \sum_a \pi_{\theta'}(a)}_{1} = 0$$

grad w.r.t θ'

$$= \sum_a H_{\pi_{\theta'}} \cdot \log \frac{\pi_{\theta'}}{\pi_\theta} + \nabla \pi_{\theta'}^{T} \cdot \frac{\nabla \pi_{\theta'}}{\pi_{\theta'}} + H_{\pi_{\theta'}} \Bigg|_{\theta' = \theta},$$

$$F = \sum_a \frac{\nabla \pi_{\theta'} \nabla \pi_{\theta'}^{T}}{\pi_{\theta'}}$$

$$= H \sum \pi_{\theta'}(a)$$

$$\underbrace{a a^T} = \begin{array}{c c} & j \\ i & \begin{bmatrix} & \vdots & \\ \cdots & a_i a_j & \\ & & \end{bmatrix} \end{array}$$

$$\begin{cases} \max\limits_{\theta'} & \nabla J_\theta^T (\theta' - \theta) \\ \text{s.t.} & (\theta' - \theta)^T F (\theta' - \theta) \le \epsilon \end{cases}$$

$$\underbrace{\nabla_{\theta'} J\Big|_{\theta'=\theta}}(\theta' - \theta) - \lambda\left[(\theta' - \theta)^T F (\theta' - \theta) - \epsilon\right]$$

$$\underbrace{\nabla_{\theta'} J\Big|_{\theta'=\theta}} - 2\lambda F(\theta' - \theta) = 0$$

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

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

$$\underbrace{\nabla J^T \cancel{F^{-1}}}$$

$$\alpha^2 (F^{-1} \nabla J)^T \cancel{F} (F^{-1} \nabla J) \le \epsilon$$

$$\underline{\alpha \le \sqrt{\frac{\epsilon}{\nabla J^T F^{-1} \nabla J}}}$$

$$\underline{\underline{f(\theta')}} = \underline{\underline{f(\theta)}} + \overset{0}{\nabla_{\theta'} f^T \Big|_{\underline{\underline{\theta'=\theta}}}} (\theta' - \theta)$$

$$+ \frac{1}{2}(\theta'-\theta)^T H_f (\theta'-\theta) + \cdots$$

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

Taylor expansion:

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

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

$$F = \sum_a \frac{\nabla_\theta \pi_\theta \nabla_\theta \pi_\theta^T}{\pi_\theta}$$

$\nabla^T \log \pi_\theta$

$$= \sum_a \frac{\pi_\theta \nabla \pi_\theta \nabla^T \pi_\theta}{\pi_\theta \pi_\theta}$$

$\nabla \log \pi_\theta$

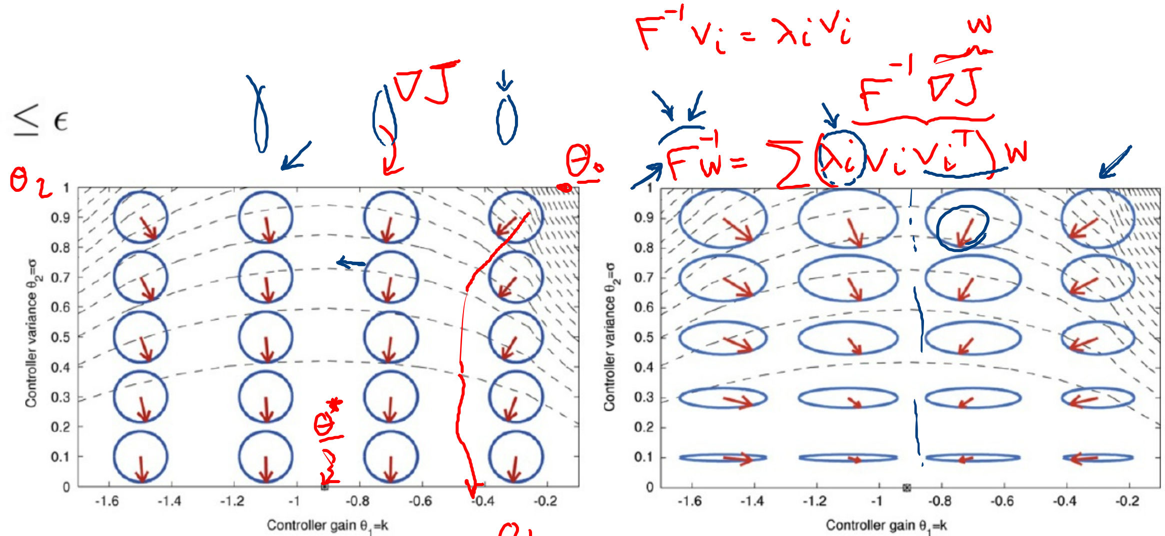
Natural Policy Gradient

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

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

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

natural gradient: pick α



(a) Vanilla policy gradient. (b) Natural policy gradient.

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)

$x^T A x$

$A \succ 0$

$\Delta\theta \propto \nabla_\theta J$

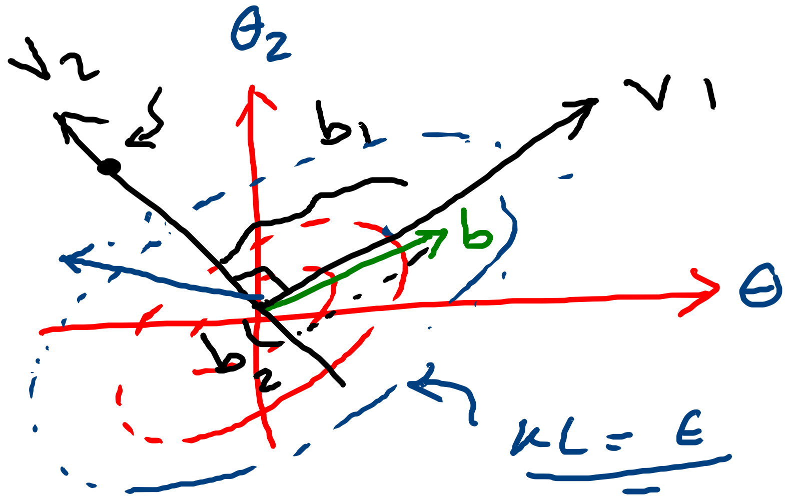
$D_{KL}(\pi_{\theta'} \| \pi_\theta)$

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

$$\begin{cases} \max\limits_{\theta'} & (\theta' - \theta)^T \nabla_\theta J \\ \text{s.t.} & (\theta' - \theta)^T F (\theta' - \theta) \leq \epsilon \end{cases}$$

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

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



$$\langle \overset{b}{\overbrace{\nabla J}}, \Delta\theta \rangle \ \text{s.t.}\ D_{KL} \le \epsilon$$

$$\frac{b_1}{\lambda_1} + \frac{b_2}{\lambda_2}$$

$$\Delta\theta = \alpha \underbrace{F^{-1} \nabla J}$$

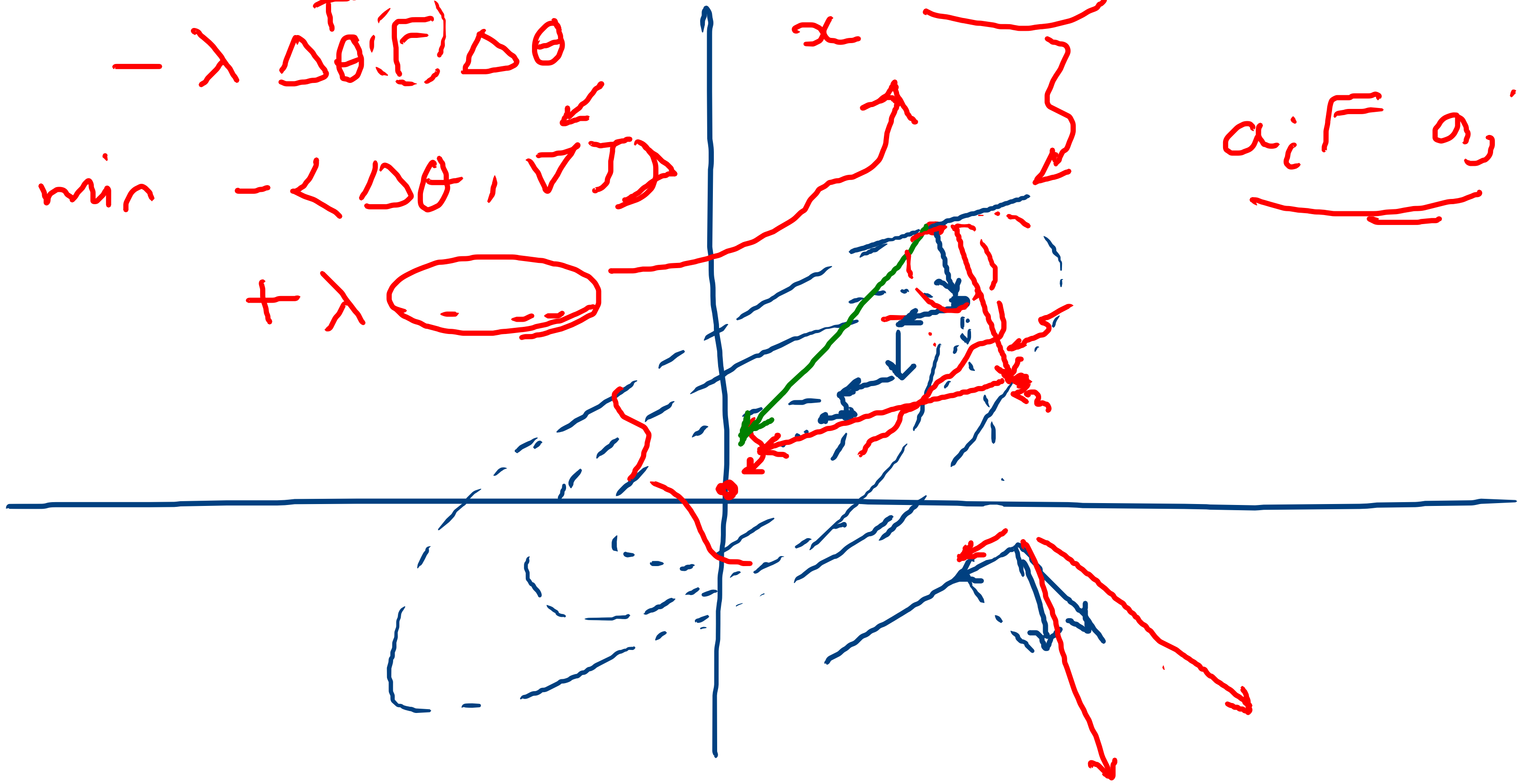
$$\nabla J$$

$$\max \langle \Delta\theta, \nabla J \rangle - \lambda \Delta\theta^T F \Delta\theta$$

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

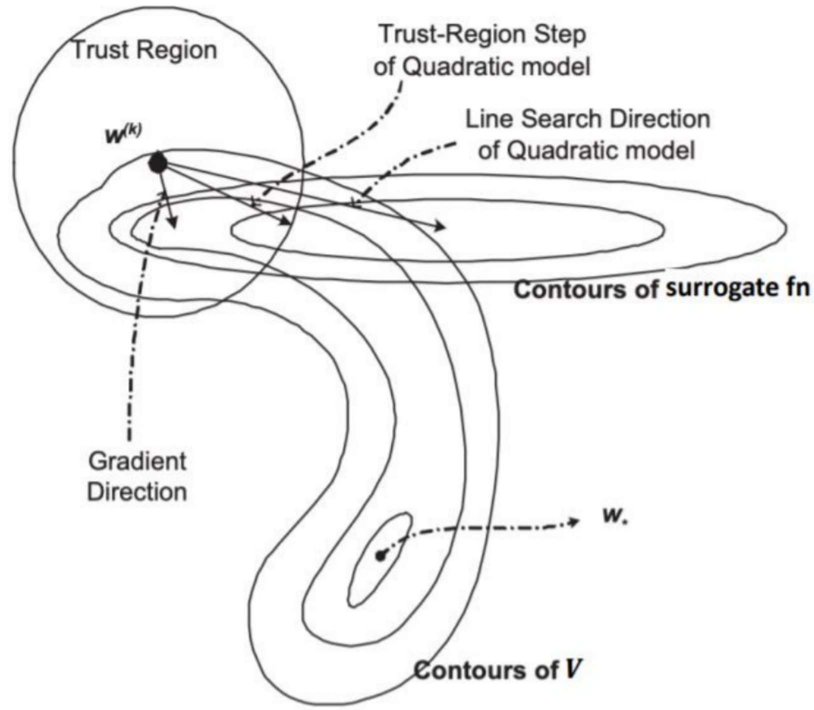
$$\min_x \underbrace{x^T A x}_{} + b^T x$$

$$\underbrace{a_i F a_j = 0}_{}$$



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":

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

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

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?

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

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

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

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

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

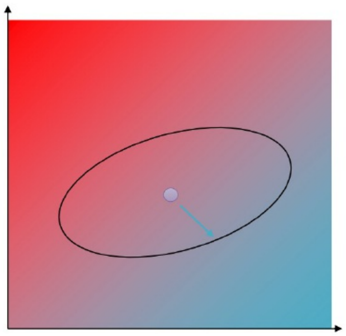
an instance of *dual gradient descent*

Natural Gradient Based on Trust Region

Recall:

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

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



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

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

natural gradient

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

Proximal Policy Optimization

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

$$D_{KL}(\pi_{\theta'}(.|s)\,||\pi_\theta(.|s)) \leq \epsilon$$

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

$$D_{KL}(\pi_{\theta'}(.|s)\,||\pi_\theta(.|s)) = \Sigma_a \pi_{\theta'}(a|s) log \frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}$$

We are effectively constraining the ratio $\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}$

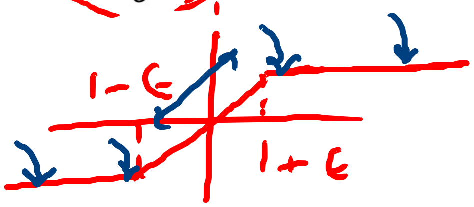
Proximal Policy Optimization

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

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

$$\underset{\theta'}{\operatorname{argmax}}\, E_{\{s\sim\mu_\theta, a\sim\pi_\theta\}}\, min \begin{cases} \frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)} A^{\pi_\theta}(s,a), \\ clip(\frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}, 1-\epsilon, 1+\epsilon) A^{\pi_\theta}(s,a) \end{cases}$$

could be easily implemented with auto-diff packages



where $clip(x, 1-\epsilon, 1+\epsilon) = \begin{cases} 1-\epsilon & if\ x < 1-\epsilon \\ x & if\ 1-\epsilon \le x \le 1+\epsilon \\ 1+\epsilon & if\ x > 1+\epsilon \end{cases}$