Trust Region Policy Optimization & NPG

Recap on NPG:

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$

$$\boxed{\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta}$$

Intuition: maximize local adv subject to being incremental (in KL);

Recap on NPG:

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a) \right] \longrightarrow \text{First-order Taylor expansion at } \theta_t$$

$$\text{s.t., } KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta \longrightarrow \text{second-order Taylor expansion at } \theta_t$$

Intuition: maximize local adv subject to being incremental (in KL);

Recap on NPG:

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$ → First-order Taylor expansion at θ_t

s.t., $KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$ → second-order Taylor expansion at θ_t

Intuition: maximize local adv subject to being incremental (in KL);

PG

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

s.t. $(\theta - \theta_t)^\top F_{\theta_t}(\theta - \theta_t) \leq \delta$

Recap on NPG:

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a) \right]$$ → First-order Taylor expansion at θ_t

s.t., $KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$ → second-order Taylor expansion at θ_t

Intuition: maximize local adv subject to being incremental (in KL);

$$\max_\theta \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

s.t. $(\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})$$ ←

NPG

Recap on NPG:

At iteration t:

$$\max_{\pi_\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(s)} A^{\pi_{\theta_t}}(s,a)\right]$$ → First-order Taylor expansion at θ_t

s.t., $KL\left(\rho_{\pi_{\theta_t}} | \rho_{\pi_\theta}\right) \leq \delta$ → second-order Taylor expansion at θ_t

Intuition: maximize local adv subject to being incremental (in KL);

$$\max_\theta \nabla_\theta J(\pi_{\theta_t})^\top(\theta - \theta_t)$$

$$\text{s.t. } \underline{(\theta-\theta_t)^\top F_{\theta_t}(\theta-\theta_t)} \leq \delta$$

$$\theta_{t+1} = \theta_t + \eta F_{\theta_t}^{-1} \nabla_\theta J(\pi_{\theta_t})$$ ←

NPG

$$F_{\theta_t} := \mathbb{E}_{s,a\sim d_\mu^{\pi_{\theta_t}}}\left[\underbrace{\nabla_\theta \ln \pi_{\theta_t}(a|s)\left(\nabla_\theta \ln \pi_{\theta_t}(a|s)\right)^\top}_{PSD}\right] \in \mathbb{R}^{dim_\theta \times dim_\theta}$$

Outline for Today:

1. Derivation of the closed-form NPG update

2. Intuitive Explanation of Natural (Policy) Gradient

3. Review of Policy Optimization (API, CPI, PG, and NPG) & a new algorithm

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$

s.t. $(\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_{\theta} J(\pi_{\theta_t})^{\top}(\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^{\top} F_{\theta_t} (\theta - \theta_t) \leq \delta$$

Notation simplification

$\longrightarrow$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \underbrace{\nabla_\theta J(\pi_{\theta_t})}_{\nabla}^\top \underbrace{(\theta - \theta_t)}_{\Delta}$$
$$\text{s.t. } (\theta - \theta_t)^\top \underbrace{F_{\theta_t}}_{F} (\theta - \theta_t) \le \delta$$

Notation simplification $\longrightarrow$

$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \le \delta$$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$$

$\xrightarrow{\text{Notation simplification}}$

$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \leq \delta$$

$\big\downarrow$ $\widetilde{\Delta} := F^{1/2}\Delta$

$F^{\frac{1}{2}} = \sqrt{F}$

$F = U \Sigma U^T$

$F^{\frac{1}{2}} = U \sqrt{\Sigma}\, U^T$

$(F^{\frac{1}{2}})^2 = F$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \le \delta$$

Notation simplification $\longrightarrow$

$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \le \delta$$

$\rightsquigarrow \underbrace{\nabla^\top F^{\frac{1}{2}} F^{-\frac{1}{2}}}_{I} \Delta$

$\Rightarrow (F^{-\frac{1}{2}}\nabla)^\top \tilde{\Delta}$

$\Delta^\top F^{\frac{1}{2}} F^{\frac{1}{2}} \Delta$

$= (F^{\frac{1}{2}}\Delta)^\top (F^{\frac{1}{2}}\Delta)$

$= \tilde{\Delta}^\top \cdot \tilde{\Delta}$

$\tilde{\Delta} := F^{1/2}\Delta$

$$\max_{\tilde{\Delta}} \left(F^{-1/2}\nabla\right)^\top \tilde{\Delta},$$
$$\text{s.t. } \tilde{\Delta}^\top \tilde{\Delta} \le \delta$$

$\|\tilde{\Delta}\|_2 \le \delta$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
s.t. $(\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$

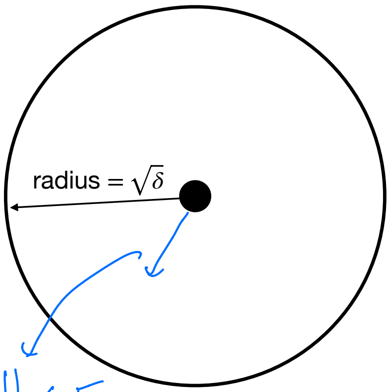
$\xrightarrow{\text{Notation simplification}}$

$$\max_{\Delta} \nabla^\top \Delta,$$
s.t. $\Delta^\top F \Delta \leq \delta$

$\big\downarrow \; \widetilde{\Delta} := F^{1/2}\Delta$

$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^\top \widetilde{\Delta},$$

s.t. $\widetilde{\Delta}^\top \widetilde{\Delta} \leq \delta$



$\|\widetilde{\Delta}\|_2 \leq \sqrt{\delta}$

At iteration t, NPG solves a convex constrained optimization problem:

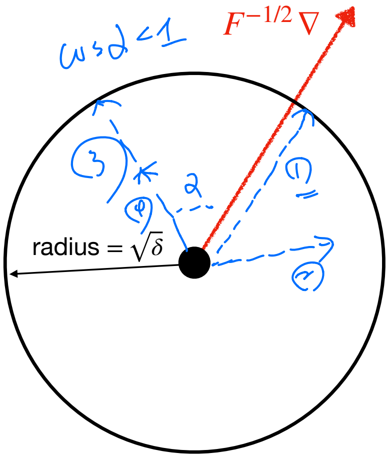
$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \le \delta$$

Notation simplification $\longrightarrow$

$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \le \delta$$

$\downarrow$ $\widetilde{\Delta} := F^{1/2}\Delta$

$$\max_{\widetilde{\Delta}} \boxed{\left(F^{-1/2}\nabla\right)}^\top \widetilde{\Delta},$$
$$\text{s.t. } \widetilde{\Delta}^\top \widetilde{\Delta} \le \delta$$



$①^\top (F^{-\frac{1}{2}} \nabla)$
$> ②^\top (F^{-\frac{1}{2}} \nabla)$

$① = \eta \cdot F^{-\frac{1}{2}} \nabla$

$\|①\|_2 = \sqrt{\delta}$

$\Rightarrow \| \eta F^{-\frac{1}{2}} \nabla \|_2 = \sqrt{\delta} \Rightarrow \eta^2 \nabla^\top F^{-\frac{1}{2}} F^{-\frac{1}{2}} \nabla = \delta$

solve for η:

$\Rightarrow \eta = \sqrt{\dfrac{\delta}{\nabla^\top F^{-1} \nabla}}$

$① = \eta \cdot F^{-\frac{1}{2}} \cdot \nabla$

At iteration t, NPG solves a convex constrained optimization problem:

$\nabla_\theta J(\theta)\big|_{\theta=\theta_t}$

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \le \delta$$

Notation simplification $\longrightarrow$

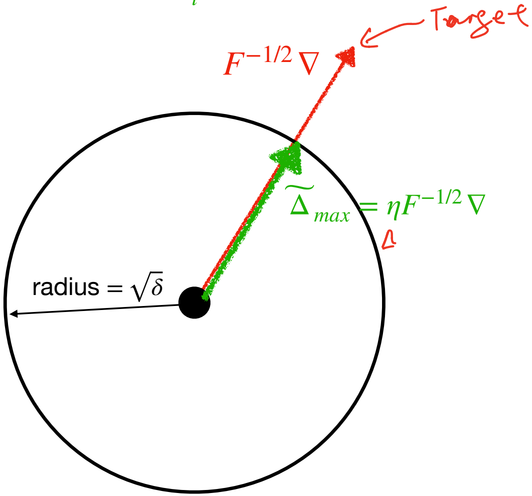
$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \le \delta$$

$$F^{-\frac{1}{2}} \nabla = \underbrace{F_{\theta_t}^{-\frac{1}{2}} \nabla_\theta J(\theta_t)}_{\text{vector defined by } \theta_t}$$

$\big\downarrow\ \widetilde{\Delta} := F^{1/2}\Delta$

$$\max_{\widetilde{\Delta}} \left(F^{-1/2} \nabla\right)^\top \widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^\top \widetilde{\Delta} \le \delta$$



At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$$

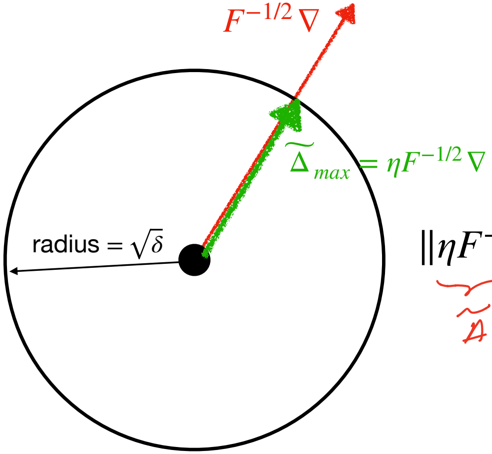
$$\xrightarrow{\text{Notation simplification}}$$

$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \leq \delta$$

$$\Big\downarrow \; \widetilde{\Delta} := F^{1/2}\Delta$$

$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^\top \widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^\top \widetilde{\Delta} \leq \delta$$



$$\underbrace{\|\eta F^{-1/2}\nabla\|}_{\widetilde{\Delta}_{max}}{}_2 = \sqrt{\delta}$$

At iteration t, NPG solves a convex constrained optimization problem:

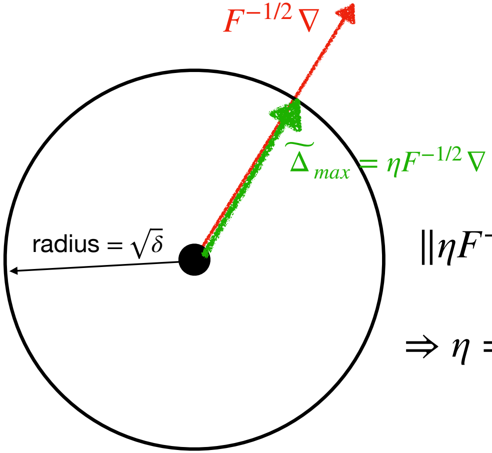
$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top(\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t}(\theta - \theta_t) \le \delta$$

$\xrightarrow{\text{Notation simplification}}$

$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \le \delta$$

$\downarrow \; \widetilde{\Delta} := F^{1/2}\Delta$

$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^\top \widetilde{\Delta},$$
$$\text{s.t. } \widetilde{\Delta}^\top \widetilde{\Delta} \le \delta$$



$$\|\eta F^{-1/2}\nabla\|_2 = \sqrt{\delta}$$
$$\Rightarrow \eta = \sqrt{\frac{\delta}{\nabla^\top F^{-1}\nabla}}$$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \le \delta$$

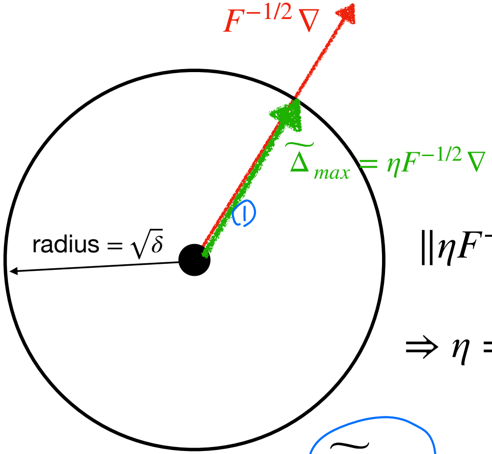
Notation simplification

$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \le \delta$$

$$\widetilde{\Delta} := F^{1/2}\Delta$$

$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^\top \widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^\top \widetilde{\Delta} \le \delta$$



$$\|\eta F^{-1/2}\nabla\|_2 = \sqrt{\delta}$$

$$\Rightarrow \eta = \sqrt{\frac{\delta}{\nabla^\top F^{-1}\nabla}}$$

1 $$\widetilde{\Delta}_{max} = \sqrt{\frac{\delta}{\nabla^\top F^{-1}\nabla}} F^{-1/2}\nabla$$

$$\tilde{\Delta} := F^{\frac{1}{2}}\Delta \Rightarrow \boxed{\Delta = F^{-\frac{1}{2}}\tilde{\Delta}}$$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \le \delta$$

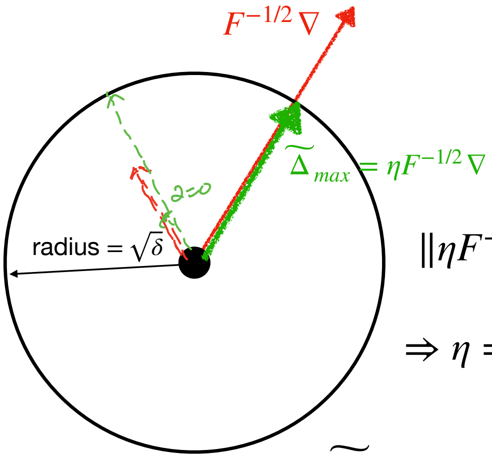
$\xrightarrow{\text{Notation simplification}}$

$$\max_{\Delta} \nabla^\top \Delta,$$
$$\text{s.t. } \Delta^\top F \Delta \le \delta$$

$$\widetilde{\Delta} := F^{1/2}\Delta$$

$$\max_{\widetilde{\Delta}} \left(F^{-1/2}\nabla\right)^\top \widetilde{\Delta},$$

$$\text{s.t. } \widetilde{\Delta}^\top \widetilde{\Delta} \le \delta$$



$$\|\eta F^{-1/2}\nabla\|_2 = \sqrt{\delta}$$

$$\Rightarrow \eta = \sqrt{\frac{\delta}{\nabla^\top F^{-1}\nabla}}$$

$$\widetilde{\Delta}_{max} := \sqrt{\frac{\delta}{\nabla^\top F^{-1}\nabla}} F^{-1/2}\nabla$$

$\theta - \theta_t$

$//$

$$\boxed{\Delta_{max} := \sqrt{\frac{\delta}{\nabla^\top F^{-1}\nabla}} F^{-1}\nabla}$$

$x^\top y$

$= \|x\|_2 \|y\|_2 \cos\langle x, y\rangle$

$F^{-\frac{1}{2}}\widetilde{\Delta}_{max} = \sqrt{\frac{\delta}{\cdots}}\, F^{-\frac{1}{2}} F^{-\frac{1}{2}} \nabla = \sqrt{\frac{\delta}{\cdots}}\, F^{-1}\nabla$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$$

A more standard and straightway is to use Lagrange multiplier $\lambda \leq 0$:

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$$

A more standard and straightway is to use Lagrange multiplier $\lambda \leq 0$:

$$\min_{\lambda \leq 0} \max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t) + \lambda \left((\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) - \delta \right)$$

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$$

A more standard and straightway is to use Lagrange multiplier $\lambda \leq 0$:

$$\min_{\lambda \leq 0} \max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t) + \lambda \left((\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) - \delta \right)$$

(This is optional: Lagrange formulation is out of scope)

At iteration t, NPG solves a convex constrained optimization problem:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t)$$
$$\text{s.t. } (\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) \leq \delta$$

A more standard and straightway is to use Lagrange multiplier $\lambda \leq 0$:

$$\min_{\lambda \leq 0} \max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t) + \lambda \left((\theta - \theta_t)^\top F_{\theta_t} (\theta - \theta_t) - \delta \right)$$

(This is optional: Lagrange formulation is out of scope)

Summary: at this stage, we complete the NPG algorithm derivation

Outline for Today:

1. Derivation of the closed-form NPG update

2. Intuitive Explanation of Natural (Policy) Gradient

3. Review of Policy Optimization (API, CPI, PG, and NPG) & a new algorithm

NPG update: $\theta_1 = \theta_0 + \eta \underbrace{F_{\theta_0}^{-1} \nabla_{\theta_0}}_{\text{Natural Gradient}}$

NPG update: $\theta_1 = \theta_0 + \eta F_{\theta_0}^{-1} \nabla_{\theta_0}$

$$KL\left(\rho_{\pi_{\theta_0}} | \rho_{\pi_\theta}\right) \leq \delta \Rightarrow \underbrace{(\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0)}_{\text{metric, } F_{\theta_0} \text{ is PD}} \leq \delta$$

NPG update: $\theta_1 = \theta_0 + \eta F_{\theta_0}^{-1} \nabla_{\theta_0}$

$$KL\left(\rho_{\pi_{\theta_0}} | \rho_{\pi_\theta}\right) \leq \delta \Rightarrow (\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$

Our goal is to make sure two distributions do not change to much,
but parameters θ could potential change a lot!

NPG update: $\theta_1 = \theta_0 + \eta F_{\theta_0}^{-1} \nabla_{\theta_0}$

$$KL\left(\rho_{\pi_{\theta_0}} | \rho_{\pi_\theta}\right) \leq \delta \Rightarrow (\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$

Our goal is to make sure two distributions do not change to much,
but parameters θ could potential change a lot!

$F_{\theta_0}^{-1} \cdot \nabla_{\theta_0}$

Consider special case where F_{θ_0} is a diagonal matrix: $F_{\theta_0} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}$ $= \begin{bmatrix} \frac{1}{\sigma_1} \cdot \nabla_\theta[1] \\ \frac{1}{\sigma_2} \cdot \nabla_\theta[2] \\ \frac{1}{\sigma_3} \cdot \nabla_\theta[3] \end{bmatrix}$

NPG update: $\theta_1 = \theta_0 + \eta F_{\theta_0}^{-1} \nabla_{\theta_0}$

$$KL\left(\rho_{\pi_{\theta_0}} | \rho_{\pi_\theta}\right) \leq \delta \Rightarrow (\theta - \theta_0)^\top F_{\theta_0} (\theta - \theta_0) \leq \delta$$

Our goal is to make sure two distributions do not change to much, but parameters θ could potential change a lot!

Consider special case where F_{θ_0} is a diagonal matrix: $F_{\theta_0} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}$

$$\forall i: \ \theta_1[i] = \theta_0[i] + \boxed{\left(\eta \sigma_i^{-1}\right)} \nabla_{\theta_0}[i]$$

NPG update: $\theta_1 = \theta_0 + \eta F_{\theta_0}^{-1} \nabla_{\theta_0}$

$$KL\left(\rho_{\pi_{\theta_0}} | \rho_{\pi_\theta}\right) \leq \delta \Rightarrow (\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta$$

Our goal is to make sure two distributions do not change to much,
but parameters θ could potential change a lot!

Consider special case where F_{θ_0} is a diagonal matrix: $F_{\theta_0} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}$

$$\forall i : \ \theta_1[i] = \theta_0[i] + \left(\eta \sigma_i^{-1}\right) \nabla_{\theta_0}[i]$$

For tiny σ_i, we indeed have a **huge** learning rate, i.e., $\eta \sigma_i^{-1}$, at coordinate i !

NPG update: $\theta_1 = \theta_0 + \eta F_{\theta_0}^{-1} \nabla_{\theta_0}$

$(\theta - \theta_0)^\top \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{bmatrix} (\theta - \theta_0)$

$$KL\left(\rho_{\pi_{\theta_0}} | \rho_{\pi_\theta}\right) \leq \delta \Rightarrow (\theta - \theta_0)^\top F_{\theta_0} (\theta - \theta_0) \leq \delta$$

Our goal is to make sure two distributions do not change to much,
but parameters θ could potential change a lot!

Consider special case where F_{θ_0} is a diagonal matrix: $F_{\theta_0} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}$

η / σ_i

$$\forall i: \ \theta_1[i] = \theta_0[i] + \left(\eta \sigma_i^{-1}\right) \nabla_{\theta_0}[i]$$

For tiny σ_i, we indeed have a **huge** learning rate, i.e., $\eta \sigma_i^{-1}$, at coordinate i !

In other words, NPG **allows a big jump** on some coordinates which do not affect KL-div too much

Example of Natural Gradient on 1-d problem:

$\theta \in \mathbb{R}$

$p[1]$ $p[2]$ $p[1] + p[2] = 1$

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

$$\lim_{\theta \to \infty} \frac{\exp(\theta)}{1+\exp(\theta)} = \frac{\exp(\theta)}{\exp(\theta)} = 1$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

$\infty^+ = \underset{\theta}{\operatorname{argmax}}\, g(\theta)$??

$p_{\theta^*}[1] = 1$

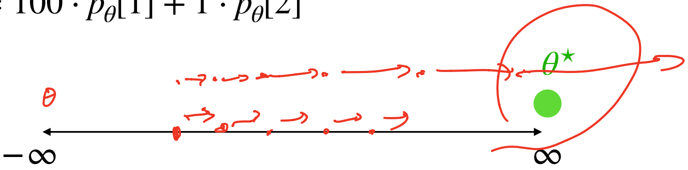
$p_{\theta^*}[2] = 0$

Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

$$\theta^\star = \arg\max_\theta g(\theta)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

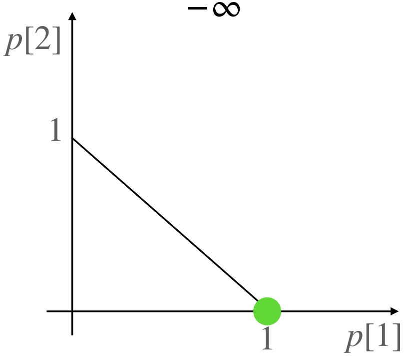


$$\theta' = \theta + g'(\theta)$$

Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1 + \exp(\theta)}, \frac{1}{1 + \exp(\theta)} \right)$$

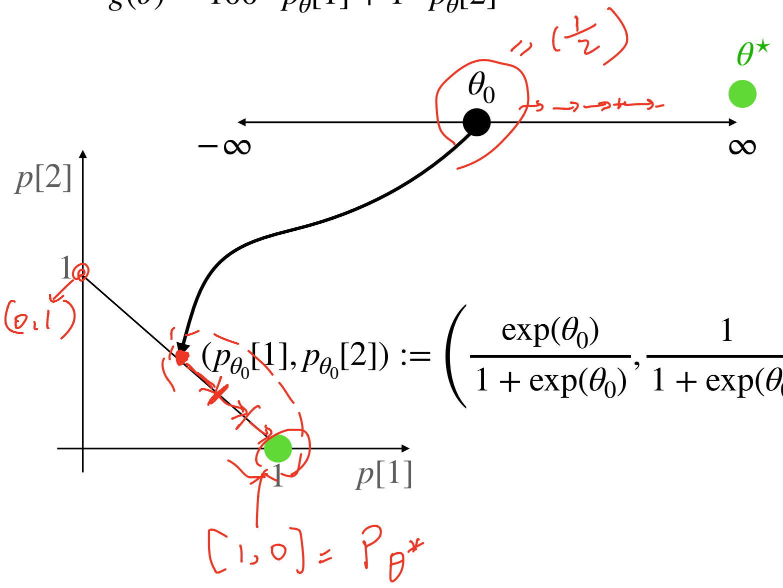
$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$



Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)} \right) = \left(\frac{1}{2}, \frac{1}{2}\right)$$

$F_\theta \in R^{d_\theta \times d_\theta}$

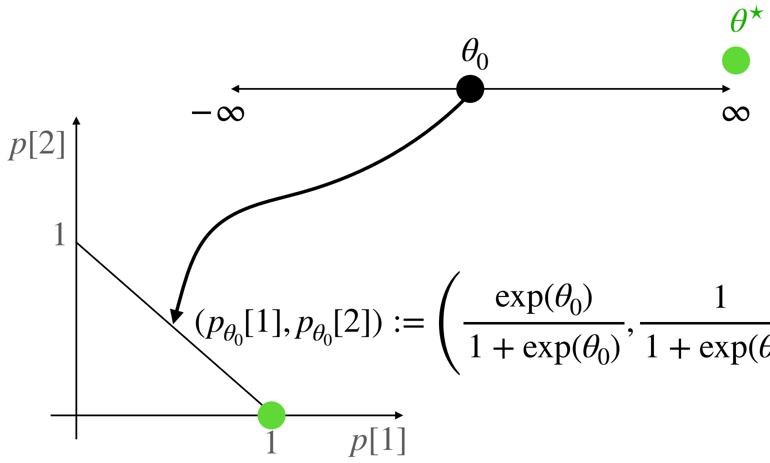
Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)}\right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

$= KL(p_{\theta_0} \| p_\theta)\big|_{\theta=\theta_0}$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)}\right)$$

Example of Natural Gradient on 1-d problem:

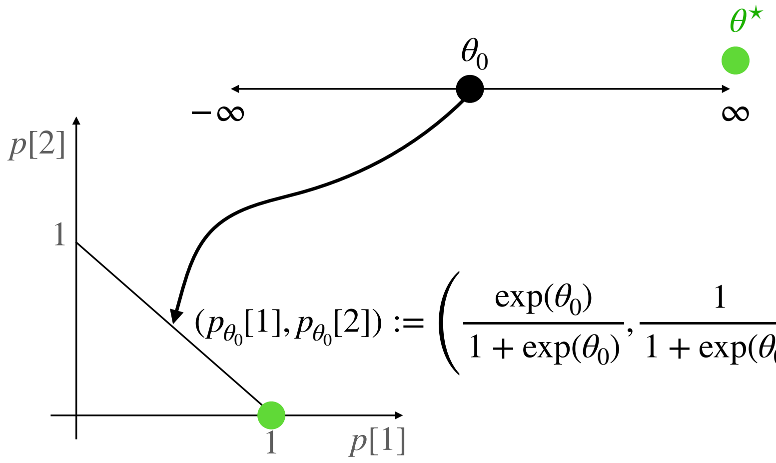
$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)}\right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

Hence: $f_{\theta_0} \to 0^+$, as $\theta_0 \to \infty$

$$\lim_{\theta_0 \to \infty} \frac{\exp(\theta_0)}{(1+\exp(\theta_0))^2} = 0^+$$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)}\right)$$

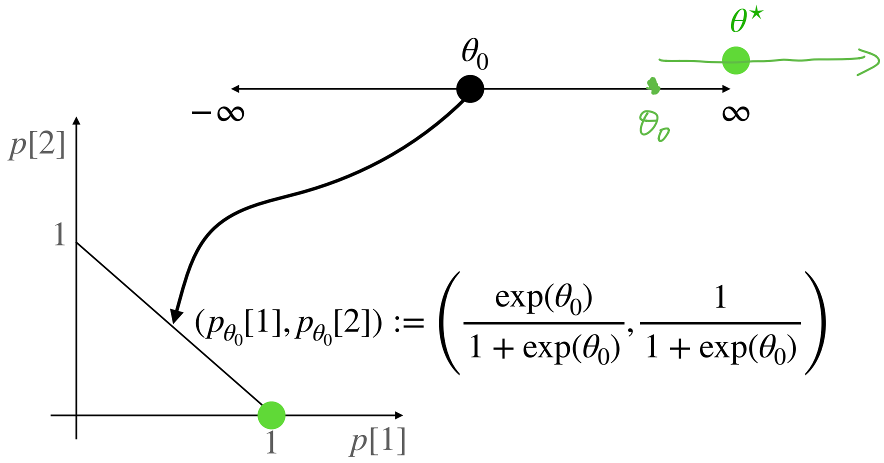
Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)}\right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

Hence: $f_{\theta_0} \to 0^+$, as $\theta_0 \to \infty$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)}\right)$$

NPG: $\theta_1 = \theta_0 + \eta \dfrac{g'(\theta_0)}{f_{\theta_0}}$

$= \theta_0 + \dfrac{\eta}{f_{\theta_0}} g'(\theta_0)$

$\dfrac{\eta}{f_{\theta_0}} \to \infty^+, \ \theta_0 \to \infty^+$

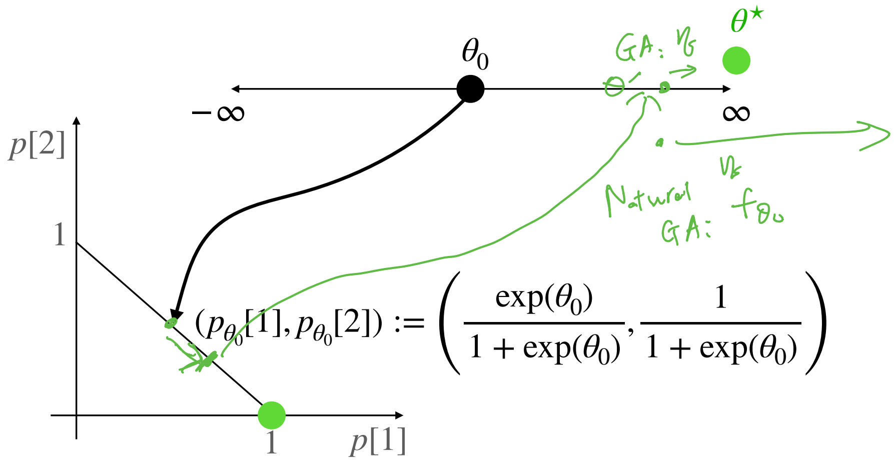
Example of Natural Gradient on 1-d problem:

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)}\right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

Hence: $f_{\theta_0} \to 0^+$, as $\theta_0 \to \infty$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)}\right)$$

NPG: $\theta_1 = \theta_0 + \eta \dfrac{g'(\theta_0)}{f_{\theta_0}}$

GA: $\theta_1 = \theta_0 + \eta g'(\theta_0)$

Example of Natural Gradient on 1-d problem:

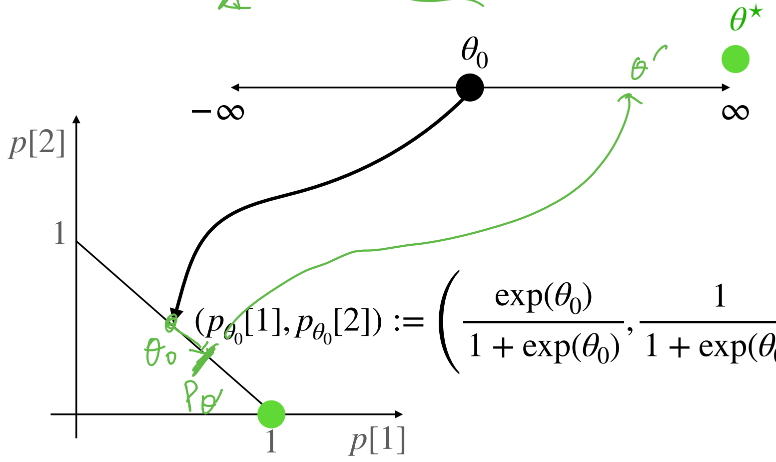
$\mathbb{E}_{s a} \nabla \ln \pi_{sa} \nabla \ln \pi_{sa}^T = U \Sigma Q^T$

$$p_\theta = \left(\frac{\exp(\theta)}{1+\exp(\theta)}, \frac{1}{1+\exp(\theta)} \right)$$

$$g(\theta) = 100 \cdot p_\theta[1] + 1 \cdot p_\theta[2]$$

Fisher information scalar: $f_{\theta_0} = \dfrac{\exp(\theta_0)}{(1+\exp(\theta_0))^2}$

Hence: $f_{\theta_0} \to 0^+$, as $\theta_0 \to \infty$



$$(p_{\theta_0}[1], p_{\theta_0}[2]) := \left(\frac{\exp(\theta_0)}{1+\exp(\theta_0)}, \frac{1}{1+\exp(\theta_0)} \right)$$

NPG: $\theta_1 = \theta_0 + \eta \dfrac{g'(\theta_0)}{f_{\theta_0}}$

GA: $\theta_1 = \theta_0 + \eta g'(\theta_0)$

i.e., Plain GA in θ will move to $\theta = \infty$ at a constant speed,
while Natural GA can traverse faster and faster when θ gets bigger
(subject to the same learning rate)

Outline for Today:

✓ 1. Derivation of the closed-form NPG update

✓ 2. Intuitive Explanation of Natural (Policy) Gradient

(In HW2, try to compare PG and NPG, see how they perform differently in practice!)

3. Review of Policy Optimization (API, CPI, PG, and NPG) & a new algorithm

Review on Policy Optimization:

We have huge ~~space~~ state space, i.e., $|S|$ might be $255^{3\times512\times512}$

We can only reset from initial state distribution $s_0 \sim \mu$

Review on Policy Optimization:

We have huge space space, i.e., $|S|$ might be $255^{3\times512\times512}$

We can only reset from initial state distribution $s_0 \sim \mu$

Numeration over state (e.g., a for loop) is not possible!

Review on Policy Optimization:

We have huge space space, i.e., $|S|$ might be $255^{3\times512\times512}$

We can only reset from initial state distribution $s_0 \sim \mu$

Numeration over state (e.g., a for loop) is not possible!

Goal: learn w/ function approximation

Review on Policy Optimization:

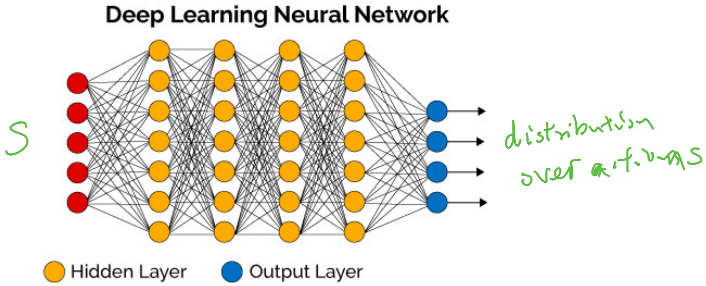
We have huge space space, i.e., $|S|$ might be $255^{3\times512\times512}$

We can only reset from initial state distribution $s_0 \sim \mu$

Numeration over state (e.g., a for loop) is not possible!

Goal: learn w/ function approximation

A Policy is a classifier w/ A many classes



Review on Policy Optimization:

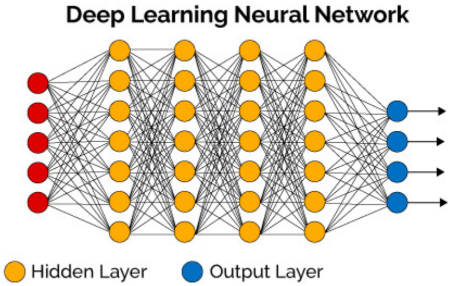
We have huge space space, i.e., $|S|$ might be $255^{3\times 512\times 512}$

We can only reset from initial state distribution $s_0 \sim \mu$

Numeration over state (e.g., a for loop) is not possible!

Goal: learn w/ function approximation

A Policy is a classifier w/ A many classes



What about continuous actions $a \in \mathbb{R}^d$?

Review on Policy Optimization:

We have huge space space, i.e., $|S|$ might be $255^{3\times512\times512}$

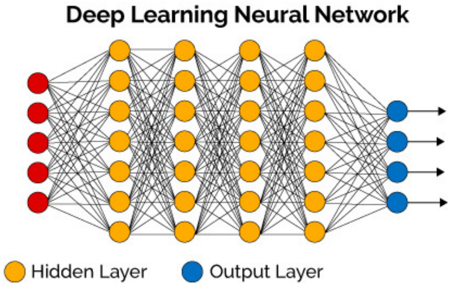
We can only reset from initial state distribution $s_0 \sim \mu$

Numeration over state (e.g., a for loop) is not possible!

Goal: learn w/ function approximation

$s \to \mu_\beta(s) \in \mathbb{R}^d$

A Policy is a classifier w/ A many classes



What about continuous actions $a \in \mathbb{R}^d$?

$\nabla_\theta \ln \pi_\theta(a|s)$ ✓

$$\pi_{\beta,\alpha}(\cdot\,|\,s) = \mathcal{N}\left(\mu_\beta(s), \exp(\alpha) I_{d\times d}\right)$$

$$\theta := [\beta, \alpha]$$

Review on Policy Optimization:

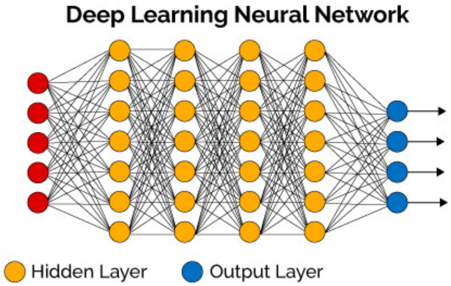
We have huge space space, i.e., $|S|$ might be $255^{3\times512\times512}$

We can only reset from initial state distribution $s_0 \sim \mu$

Numeration over state (e.g., a for loop) is not possible!

Goal: learn w/ function approximation

A Policy is a classifier w/ A many classes

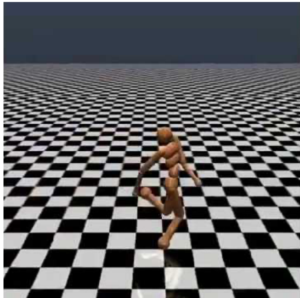


What about continuous actions $a \in \mathbb{R}^d$?

$$\pi_{\beta,\alpha}(\cdot \mid s) = \mathcal{N}\left(\mu_\beta(s), \exp(\alpha) I_{d\times d}\right)$$

$a \sim \pi_{\beta,\alpha}(\cdot|s)$

$$\theta := [\beta, \alpha]$$



Review on Policy Optimization: API

Given an current policy π^t, we perform policy update to π^{t+1}

First attempt: **Approximate Policy Iteration**

$$\pi^{t+1} = \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

Review on Policy Optimization: API

Given an current policy π^t, we perform policy update to π^{t+1}

First attempt: **Approximate Policy Iteration**

$$\pi^{t+1} = \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

i.e., find the greedy policy that maximizes the local advantage (e.g., via regression)

Review on Policy Optimization: API

Given an current policy π^t, we perform policy update to π^{t+1}

First attempt: **Approximate Policy Iteration**

$$\pi^{t+1} = \arg\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi_t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

i.e., find the greedy policy that maximizes the local advantage (e.g., via regression)

Unfortunately, π^{t+1} might be very different from π^t,
and API could fail to make any progress

Review on Policy Optimization: CPI

Given an current policy π^t, we perform policy update to π^{t+1}

second attempt: **Conservative Policy Iteration**

$$\pi_{grd} = \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

Review on Policy Optimization: CPI

Given an current policy π^t, we perform policy update to π^{t+1}

second attempt: **Conservative Policy Iteration**

$$\pi_{grd} = \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

$$\forall s : \pi^{t+1}(\cdot \mid s) = (1 - \alpha)\pi^t(\cdot \mid s) + \alpha \pi_{grd}(\cdot \mid s)$$

π^t π^{t+1} π_{grd}

Review on Policy Optimization: CPI

Given an current policy π^t, we perform policy update to π^{t+1}

second attempt: **Conservative Policy Iteration**

$$\pi_{grd} = \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

$$\forall s : \pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^t(\cdot \mid s) + \alpha \pi_{grd}(\cdot \mid s)$$

i.e., CPI find the greedy policy, and move towards it a little bit!

Review on Policy Optimization: CPI

Given an current policy π^t, we perform policy update to π^{t+1}

second attempt: **Conservative Policy Iteration**

$$\pi_{grd} = \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

$$\forall s : \pi^{t+1}(\cdot \mid s) = (1 - \alpha)\pi^t(\cdot \mid s) + \alpha \pi_{grd}(\cdot \mid s)$$

i.e., CPI find the greedy policy, and move towards it a little bit!

Two nice properties:

Review on Policy Optimization: CPI

Given an current policy π^t, we perform policy update to π^{t+1}

second attempt: **Conservative Policy Iteration**

$$\pi_{grd} = \arg\max_{\pi \in \Pi} \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left[A^{\pi^t}(s, \pi(s)) \right]$$

$$\forall s : \pi^{t+1}(\cdot \mid s) = (1-\alpha)\pi^t(\cdot \mid s) + \alpha \pi_{grd}(\cdot \mid s)$$

i.e., CPI find the greedy policy, and move towards it a little bit!

$\sum_s \left| d^{\pi^t}(s) - d^{\pi^{t+1}}(s) \right|$

Two nice properties:

$$\left\| d^{\pi^t}(\cdot) - d^{\pi^{t+1}}(\cdot) \right\|_1 \leq O\left(\frac{\alpha}{1-\gamma}\right), \quad V^{\pi^{t+1}} > V^{\pi^t} \text{ (if not terminate yet)}$$

Review on Policy Optimization: PG

Given an current policy π^t, we perform policy update to π^{t+1}

Third attempt: **PG on parameterized policy**

$$\max_{\theta} \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s, a) \right]$$

Review on Policy Optimization: PG

Given an current policy π^t, we perform policy update to π^{t+1}

Third attempt: **PG on parameterized policy**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \right]$$

Locally Improve the local-adv a little bit via one-step gradient ascent:

$$\nabla_\theta \left[\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \right] \right] = \nabla_\theta J(\theta_t)$$

$$= \nabla_\theta \left[\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a) \right] \right]$$

$$= \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \nabla_\theta \ln \pi_{\theta_t}(a|s) A^{\pi_{\theta_t}}(s,a)$$

Review on Policy Optimization: PG

Given an current policy π^t, we perform policy update to π^{t+1}

Third attempt: **PG on parameterized policy**

$$\max_{\theta} \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot | s)} A^{\pi_{\theta_t}}(s, a) \right]$$

Locally Improve the local-adv a little bit via one-step gradient ascent:

$$\theta_{t+1} = \theta_t + \eta \cdot \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_{\theta_t}(s)} \nabla \ln \pi_{\theta_t}(a \mid s) \cdot A^{\pi_{\theta_t}}(s, a) \right]$$

Review on Policy Optimization: PG

Given an current policy π^t, we perform policy update to π^{t+1}

Third attempt: **PG on parameterized policy**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right]$$

Locally Improve the local-adv a little bit via one-step gradient ascent:

$$\theta_{t+1} = \theta_t + \eta \cdot \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(s)} \nabla \ln \pi_{\theta_t}(a\,|\,s) \cdot A^{\pi_{\theta_t}}(s,a)\right]$$

When $\eta \to 0^+$, gradient ascent ensures
we improve the objective function

Review on Policy Optimization: NPG

Given an current policy π^t, we perform policy update to π^{t+1}

Fourth attempt: **Natural Policy Gradient**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right]$$

Review on Policy Optimization: NPG

Given an current policy π^t, we perform policy update to π^{t+1}

Fourth attempt: **Natural Policy Gradient**

$$\max_{\theta} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s, a) \right]$$

$$s.t., \mathrm{KL}(\rho_{\theta_t} | \rho_\theta) \leq \delta$$

Review on Policy Optimization: NPG

Given an current policy π^t, we perform policy update to π^{t+1}

Fourth attempt: **Natural Policy Gradient**

$$\max_{\theta} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathbb{E}_{a \sim \pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \right]$$

$$s.t., \mathrm{KL}(\rho_{\theta_t} | \rho_\theta) \leq \delta$$

Hessian

Define fisher info-matrix $F_{\theta_t} = \nabla_\theta^2 \mathrm{KL}(\rho_{\theta_t} | \rho_\theta)|_{\theta=\theta_t}$,

a convex approximation, e.g., linearize obj and quadratize constraint,
gives us the following NPG update:

Review on Policy Optimization: NPG

Given an current policy π^t, we perform policy update to π^{t+1}

Fourth attempt: **Natural Policy Gradient**

$$\max_{\theta} \mathbb{E}_{s\sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \right]$$

$$s.t., \mathrm{KL}(\rho_{\theta_t} | \rho_\theta) \leq \delta$$

Define fisher info-matrix $F_{\theta_t} = \nabla_\theta^2 \mathrm{KL}(\rho_{\theta_t} | \rho_\theta) |_{\theta=\theta_t}$,
a convex approximation, e.g., linearize obj and quadratize constraint,
gives us the following NPG update:

$$\max_{\theta} \nabla_\theta J(\pi_{\theta_t})^\top (\theta - \theta_t), \text{ s.t., } (\theta-\theta_t)^\top F_{\theta_t}(\theta-\theta_t) \leq \delta$$

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right]$$

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right] - \underbrace{\lambda \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\mathrm{KL}\left(\pi_{\theta_t}(a\,|\,s)\,|\,\pi_\theta(a\,|\,s)\right)\right]}_{\text{regularization}}$$

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \right] \underbrace{- \lambda \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[\mathrm{KL}\left(\pi_{\theta_t}(a|s) | \pi_\theta(a|s) \right) \right]}_{\text{regularization}}$$

Use importance weighting & expand KL divergence:

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_{\theta} \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a) \right] \underbrace{- \lambda \mathbb{E}_{s\sim d_\mu^{\pi^t}} \left[\mathrm{KL}\left(\pi_{\theta_t}(a|s) | \pi_\theta(a|s) \right) \right]}_{\text{regularization}}$$

Use importance weighting & expand KL divergence:

$$\ell(\theta) := \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \left[\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a) \right] - \lambda \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}} \underbrace{\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)} \left[-\ln \pi_\theta(a|s) \right]}_{\text{expand KL}}$$

IW

$F_\theta \in R^{d_\theta \times d_\theta}$

$F_\theta^{-1}: \; d_\theta^3$

An extension of NPG (even faster in practice):

Given an current policy π^t, we perform policy update to π^{t+1}

fifth attempt (new): **Proximal Policy Optimization (PPO)**

$$\max_\theta \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s)} A^{\pi_{\theta_t}}(s,a)\right] - \underbrace{\lambda \mathbb{E}_{s\sim d_\mu^{\pi^t}}\left[\mathrm{KL}\left(\pi_{\theta_t}(a\,|\,s)\,|\,\pi_\theta(a\,|\,s)\right)\right]}_{\text{regularization}}$$

SGA to approximately optimize $l(\theta)$

Use importance weighting & expand KL divergence:

$$\ell(\theta) := \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)}\frac{\pi_\theta(a\,|\,s)}{\pi_{\theta_t}(a\,|\,s)} A^{\pi_{\theta_t}}(s,a)\right] - \lambda \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\mathbb{E}_{a\sim\pi_{\theta_t}(\cdot|s)}\left[-\ln \pi_\theta(a\,|\,s)\right]$$

PPO: Perform a few steps of mini-batch SGA on $\ell(\theta)$ to approximate $\arg\max_\theta \ell(\theta)$

$$\pi_\theta(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_a \exp(f_\theta(s,a))}$$

Next a few lectures:

Imitation Learning
(Learning from Demonstrations)

Can we learn a good policy purely from expert demonstrations?

$$F_{\theta_0} = \mathop{E}_{s\,a}\left[\nabla \ln \pi_\theta(a|s)\, \nabla \ln \pi_\theta(a|s)^T\right] \in \mathbb{R}^{d\times d} \qquad d:\theta$$

Hessian

$KL|_{\theta=\theta_0}$

$$F_{\theta_0} = U \Sigma U^T$$

$$F_{\theta_0}^{-1} = U \Sigma^{-1} U^T \qquad U^T U = I$$

$$F_{\theta_0}^{-1} \cdot \nabla_\theta = \underbrace{U}_{\text{Rotate}}\ \underbrace{\Sigma^{-1}}_{\text{Scale}} \big(\underbrace{U^T}_{\text{Rotate}} \nabla_{\theta_0}\big)$$

$$= \sum_{i=1}^{d} u_i \frac{1}{\sigma_i}\left(u_i^T \nabla_{\theta_0}\right)$$

$$KL(P_{\theta_0} \| P_\theta) \approx (\theta-\theta_0)^T \nabla^2 KL|_{\theta=\theta_0} (\theta-\theta_0) + \cancel{\text{first-ord}} + \cancel{\text{zero}} \quad =0$$