Policy Gradient Methods

value-based : Q → DQN
policy-based : π → PPO

(policy optimization)

Goal: distribution ρ.

$(J(\pi))$ $V^{\pi}(\rho) = \mathbb{E}_{s_0 \sim \rho}[V^{\pi}(s_0)]$

stochastic policy (multimodality)

parametric policy class $\Pi = \{\pi_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$

$\triangleq \max_{\theta \in \Theta} V^{\pi_\theta}(\rho)$

$\theta \in \mathbb{R}^d$.

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

$f_\theta(s,a) = \theta^T \phi_{s,a}$ (log-linear)

$f_\theta(s,a)$: neural net (neural softmax)

discrete action space

(A, μ) discrete: $\mu = \#$
continuous: $\mu =$ Lebesgue.

measure-theory

continuous action space.

(canonical)

e.g. 1 softmax policy

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

$\Theta = \mathbb{R}^{|S \times A|}$

soft Q-iteration
soft policy-iteration (PSPI)

e.g. 2 Gaussian policy

$$\pi_\theta(a|s) = \frac{1}{\sqrt{2\pi}\,\sigma_\theta(s)} \cdot \exp\left(-\frac{(a-\mu_\theta(s))^2}{2\sigma_\theta(s)^2}\right)$$

e.g. $\mu_\theta(s) = \theta_\mu^T \chi_\mu(s)$

$\sigma_\theta(s) = \exp(\theta_\sigma^T \chi_\sigma(s))$

Recall: goal is to $\max_{\theta \in \Theta} V^{\pi_\theta}(\rho)$

$\theta_{t+1} = \theta_t + \eta \cdot \nabla_\theta V^{\pi_{\theta_t}}(\rho)$

d^π occupancy

τ: trajectory under policy π & starting distribution μ

$$\Pr_\mu^\pi(\tau) = \mu(s_0) \cdot \pi(a_0|s_0) \cdot P(s_1|s_0, a_0) \cdot \pi(a_1|s_1) \cdots$$

$$V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}}[R(\tau)]$$

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t)$$

Theorem (policy gradient theorem)

① $$\nabla_\theta V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}}\left[R(\tau) \cdot \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

② $$\nabla_\theta V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot Q^{\pi_\theta}(s_t, a_t) \cdot \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

REINFORCE

Proof of ①: $$\nabla V^{\pi_\theta}(\mu) = \nabla_\theta \sum_\tau R(\tau) \cdot \Pr_\mu^{\pi_\theta}(\tau)$$

$$= \sum_\tau R(\tau) \cdot \nabla_\theta \Pr_\mu^{\pi_\theta}(\tau)$$

log-trick

$\nabla p_\theta = p_\theta \cdot \nabla \log p_\theta$

$$= \sum_\tau R(\tau) \cdot \Pr_\mu^{\pi_\theta}(\tau) \cdot \nabla_\theta \log \Pr_\mu^{\pi_\theta}(\tau)$$

$$= \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}}\left[R(\tau) \cdot \nabla_\theta \log \Pr_\mu^{\pi_\theta}(\tau)\right].$$

$$\Pr_\mu^{\pi_\theta}(\tau) = \mu(s_0)\cdot \pi_\theta(a_0|s_0)\cdot P(s_1|s_0,a_0)\cdot \pi_\theta(a_1|s_1)\cdots$$

$$\nabla_\theta \log \Pr_\mu^{\pi_\theta}(\tau) = \nabla_\theta \log \pi_\theta(a_0|s_0) + \nabla_\theta \log \pi_\theta(a_1|s_1) + \cdots$$

$$\Rightarrow \nabla_\theta V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau\sim \Pr_\mu^{\pi_\theta}}\left[R(\tau)\cdot \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)$$

$$\theta_{t+1} = \theta_t + \eta\cdot \nabla_\theta V^{\pi_\theta}(\mu)$$

→ Monte Carlo version : REINFORCE.

Proof of ②: $\forall$ starting state s_0:

$$\nabla V^{\pi_\theta}(s_0) = \nabla \sum_{a_0} \pi_\theta(a_0|s_0)\cdot Q^{\pi_\theta}(s_0,a_0)$$

$$= \sum_{a_0} \left(\nabla \pi_\theta(a_0|s_0)\right) Q^{\pi_\theta}(s_0,a_0) + \sum_{a_0} \pi_\theta(a_0|s_0)\cdot \nabla Q^{\pi_\theta}(s_0,a_0)$$

$$= \sum_{a_0} \pi_\theta(a_0|s_0)\cdot Q^{\pi_\theta}(s_0,a_0)\cdot \nabla \log \pi_\theta(a_0|s_0)$$

$$+ \sum_{a_0} \pi_\theta(a_0|s_0)\cdot \nabla\left(r(s_0,a_0) + \gamma\cdot \sum_{s_1} P(s_1|s_0,a_0)\cdot V^{\pi_\theta}(s_1)\right)$$

$$= \sum_{a_0} \pi_\theta(a_0|s_0)\cdot \Big(Q^{\pi_\theta}(s_0,a_0) + \nabla \log \pi_\theta(a_0|s_0)$$

$$+ \gamma\cdot \sum_{s_1} P(s_1|s_0,a_0)\cdot \nabla V^{\pi_\theta}(s_1)\Big)$$

$$= \mathbb{E}_{\tau\sim \Pr_{s_0}^{\pi_\theta}}\left[Q^{\pi_\theta}(s_0,a_0)\cdot \nabla \log \pi_\theta(a_0|s_0)\right]$$

$$+ \gamma\cdot \mathbb{E}_{\tau\sim \Pr_{s_0}^{\pi_\theta}}\left[\nabla V^{\pi_\theta}(s_1)\right]$$

$$\nabla V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau\sim Pr_\mu^{\pi_\theta}}\left[Q^{\pi_\theta}(s_0,a_0)\cdot\nabla\log\pi_\theta(a_0|s_0)\right]$$

$$+\gamma\cdot\mathbb{E}_{\tau\sim Pr_\mu^{\pi_\theta}}\left[\nabla V^{\pi_\theta}(s_1)\right]$$

$$\Rightarrow \gamma\cdot\mathbb{E}_{\tau\sim Pr_\mu^{\pi_\theta}}\left[Q^{\pi_\theta}(s_1,a_1)\cdot\nabla\log\pi_\theta(a_1|s_1)\right]$$

$$+\gamma^2\cdot\mathbb{E}_{\tau\sim Pr_\mu^{\pi_\theta}}\left[Q^{\pi_\theta}(s_2,a_2)\cdot\nabla\log\pi_\theta(a_2|s_2)\right]$$

$$+\cdots$$

$$\Rightarrow \nabla V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau\sim Pr_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\gamma^t\cdot Q^{\pi_\theta}(s_t,a_t)\cdot\nabla\log\pi_\theta(a_t|s_t)\right]$$

$$\Leftrightarrow \nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\cdot\mathbb{E}_{(s,a)\sim d_\mu^{\pi_\theta}}\left[Q^{\pi_\theta}(s,a)\cdot\nabla\log\pi_\theta(a|s)\right]$$

occupancy $\rightarrow d_\mu^{\pi_\theta} = (1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t\cdot d_{\mu,t}^{\pi_\theta}$

$d_{\mu,t}^{\pi_\theta} \rightarrow Pr_\mu^{\pi_\theta}(s_t=\cdot,\ a_t=\cdot)$

REINFORCE:

generate a trajectory.

pick a time step t w.p. $\propto \gamma^t$.

Recall: policy gradient theorem.

$\Rightarrow$ 2 version of REINFORCE

① $\nabla V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau\sim Pr_\mu^{\pi_\theta}}\left[R(\tau)\cdot\sum_{t=0}^{\infty}\nabla\log\pi_\theta(a_t|s_t)\right]$

② $\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\cdot\mathbb{E}_{(s,a)\sim d_\mu^{\pi_\theta}}\left[Q^{\pi_\theta}(s,a)\cdot\nabla\log\pi_\theta(a|s)\right]$

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\rho) \Rightarrow \text{non-concave (in general)}$$

$\Rightarrow$ non-convex optimization

Convex optimization

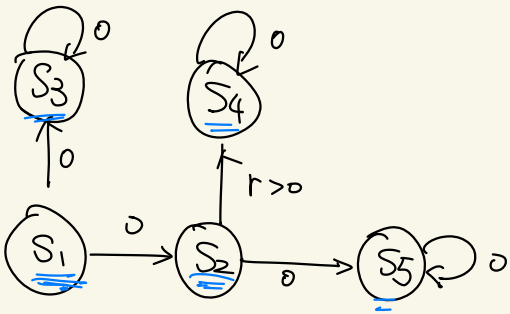
$$\min_{x \in \mathcal{X}} f(x)$$

f is convex function

$\mathcal{X}$ is convex set

Example:

M:

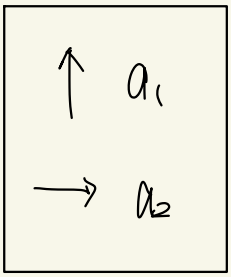


$Q^{\pi_\theta}(s, \pi_\theta)$

concave: LQR

(Linear Quadratic Regulator)

$V^{\pi_\theta}(s)$ is not concave for both direct parameterization and softmax parameterization under MDP M.



$$V^{\pi_\theta}(s_1) = \pi_\theta(a_2|s_1) \cdot \pi_\theta(a_1|s_2) \cdot r.$$

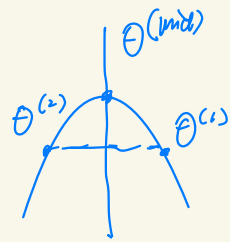
$$\theta = (\theta_{s_1,a_1}, \theta_{s_1,a_2}, \theta_{s_2,a_1}, \theta_{s_2,a_2})$$

$$\theta^{(1)} = (\log 1, \log 3, \log 3, \log 1)$$

$$\theta^{(2)} = (-\log 1, -\log 3, -\log 3, -\log 1)$$

$$\theta^{(mid)} = \frac{1}{2}(\theta^{(1)} + \theta^{(2)}) = \vec{0}$$

non-concavity: $V^{\theta^{(1)}}(s_1) + V^{\theta^{(2)}}(s_1) > 2 \cdot V^{\theta^{(mid)}}(s_1)$



concave

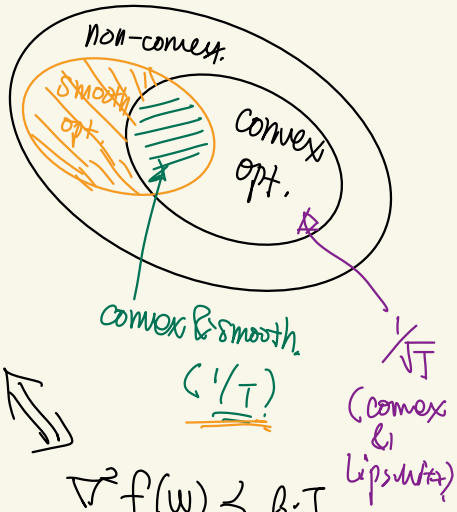
Softmax policy: $\pi_{\theta^{(1)}}(a_2|s_1) = \frac{3}{4}$. $\pi_{\theta^{(1)}}(a_1|s_2) = \frac{3}{4}$.

$\pi_{\theta^{(2)}}(a_2|s_1) = \frac{1}{4}$. $\pi_{\theta^{(2)}}(a_1|s_2) = \frac{1}{4}$

$\pi_{\theta^{(mid)}}(a_2|s_1) = \frac{1}{2}$. $\pi_{\theta^{(mid)}}(a_1|s_2) = \frac{1}{2}$

$\Rightarrow V^{\pi_{\theta^{(1)}}}(s_1) = \frac{9}{16}r$. $V^{\pi_{\theta^{(2)}}}(s_1) = \frac{1}{16}r$. $V^{\pi_{\theta^{(mid)}}}(s_1) = \frac{1}{4}r$

Smooth-optimization



def (smoothness) A function $f: \mathbb{R}^d \to \mathbb{R}$ is β-smooth if $\forall w, w'$.

$$\|\nabla f(w) - \nabla f(w')\| \le \beta \cdot \|w - w'\|$$

(gradient is β-Lipschitz)

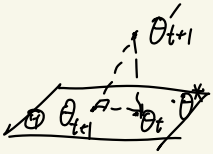
$\Leftrightarrow$ $\nabla^2 f(w) \preceq \beta \cdot I$.

convex: $\nabla^2 f(w) \succeq 0$

α-strongly-convex: $\nabla^2 f(w) \succeq \alpha \cdot I$

Gradient Ascent: $\theta_{t+1} = \theta_t + \eta \cdot \nabla V^{\pi_{\theta_t}}(\mu)$

$$\theta_{t+1} = \Pi_{\Theta}\left[\theta_t + \eta \cdot \nabla V^{\pi_{\theta_t}}(\mu)\right]$$



Lemma V^{π_θ} is β-smooth w.r.t. θ.

$V^{\pi_\theta}(\mu) \le V^{\pi_{\theta^*}}(\mu)$.

If we set the step size $\eta = 1/\beta$, $\forall T$:

$$\min_{t \le T} \|\nabla V^{(t)}(\mu)\|^2 \le \frac{2\beta \cdot (V^*(\mu) - V^{(0)}(\mu))}{T}$$

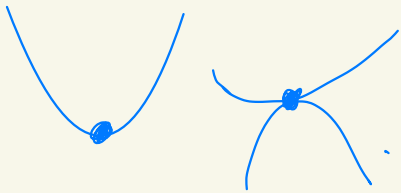
$\le V_{max}$

$V^{(t)}(\mu) = V^{\pi_{\theta^{(t)}}}(\mu)$

$V^*(\mu) = V^{\pi_{\theta^*}}(\mu)$

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla V^{(t)}(\mu)\|^2 \le \varepsilon$$

ε - approximate stationary point



(sample complexity)

$T = O(1/\varepsilon)$ rounds

OVX opt: (Descent Lemma)

$$f(x_{t+1}) \ge f(x_t) + \langle \nabla f(x_t), \underbrace{x_{t+1} - x_t}_{-\eta \cdot \nabla f(x_t)} \rangle$$

$$= f(x_t) - \eta \cdot \|\nabla f(x_t)\|^2$$

Smooth opt: (Descent Lemma)

$$f(x_{t+1}) \ge f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{\beta}{2}\|x_{t+1} - x_t\|^2$$

$$= f(x_t) - \eta \cdot \|\nabla f(x_t)\|^2 - \frac{\beta}{2}\eta^2 \cdot \|\nabla f(x_t)\|^2$$

$\eta = \frac{1}{\beta}$: $f(x_{t+1}) \ge f(x_t) - \frac{\beta}{2} \cdot \|\nabla f(x_t)\|^2$

Recall: (Policy Gradient Theorem)

① $\nabla_\theta V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}}\left[R(\tau) \cdot \sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t) \right]$.

Monte Carlo : obtain sample trajectory $\tau \sim Pr_\mu^{\pi_\theta}$

$\tau = (s_0, a_0, s_1, a_1, \cdots, s_H, a_H)$

$$\hat{R}(\tau) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \rightarrow \text{unbiased}$$

$$\widehat{\nabla_\theta V^{\pi_\theta}(\mu)} = \hat{R}(\tau) \cdot \sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t) \quad ①$$

② $$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a)\sim d_\mu^{\pi_\theta}} \left[Q^{\pi_\theta}(s,a) \cdot \nabla \log \pi_\theta(a|s) \right]$$

$$= \mathbb{E}_{\tau \sim Pr_\mu^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t \cdot Q^{\pi_\theta}(s_t, a_t) \cdot \nabla \log \pi_\theta(a_t|s_t) \right]$$

Monte Carlo : obtain sample trajectory $\tau \sim Pr_\mu^{\pi_\theta}$

$$\widehat{Q^{\pi_\theta}}(s_t, a_t) = \sum_{t'=t}^{\infty} \gamma^{t'-t} r(s_{t'}, a_{t'}) \rightarrow \text{unbiased}$$

$$\widehat{\nabla_\theta V^{\pi_\theta}(\mu)} = \sum_{t=0}^{\infty} \hat{Q}^{\pi_\theta}(s_t, a_t) \cdot \nabla \log \pi_\theta(a_t|s_t) \quad ②$$

Algorithm : (REINFORCE)

1. Initialize θ_0
2. For $t = 0, 1, \cdots$:
   - (a) sample $\tau \sim Pr_\mu^{\pi_\theta}$ (on-policy)
   - (b) $\theta_{t+1} = \theta_t + \eta_t \widehat{\nabla_\theta V^{\pi_{\theta_t}}(\mu)}$

Lemma (unbiasedness) $\mathbb{E}_{\tau \sim Pr_\mu^{\pi_\theta}} \left[\widehat{\nabla_\theta V^{\pi_\theta}(\mu)} \right] = \nabla_\theta V^{\pi_\theta}(\mu)$

Proof of ②: $$\mathbb{E}\left[\widehat{\nabla_\theta V^{\pi_\theta}(\mu)} \right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \hat{Q}^{\pi_\theta}(s_t, a_t) \cdot \nabla \log \pi_\theta(a_t|s_t) \right]$$

$$= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{E}\left[\hat{Q}^{\pi_\theta}(s_t, a_t) \,\middle|\, s_t, a_t \right] \cdot \nabla \log \pi_\theta(a_t|s_t) \right]$$

$$= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot Q^{\pi_\theta}(s_t, a_t) \cdot \nabla \log \pi_\theta(a_t|s_t)\right] = \nabla_\theta V^{\pi_\theta}(\mu)$$

$\square$

Lemma Assume
- V^{π_θ} is β-smooth
- $\mathbb{E}\left[\| \nabla \hat{V}^{\pi_\theta}(\mu) - \nabla V^{\pi_\theta}(\mu) \|^2\right] \le \sigma^2$.

with optimally-tuned step-size: $\forall T$.

$$\min_{t \le T} \mathbb{E}\left[\| \nabla V^{(t)}(\mu) \|^2\right] \le \frac{2\beta \cdot (V^*(\mu) - V^{(0)}(\mu))}{T} + \sqrt{\frac{2\sigma^2}{T}}$$

expected convergence

Variance-term (dominate)

$$\begin{cases} \eta_t = \frac{1}{\beta}. & t \le \frac{\beta \cdot (V^*(\mu) - V^{(0)}(\mu))}{\sigma^2} \\ \eta_t = \sqrt{\frac{2}{\beta T}} & \text{o.w.} \end{cases}$$

$y = ax + \frac{b}{x}$

$$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)}\left[Q^{\pi_\theta}(s,a) \cdot \nabla_\theta \log \pi_\theta(a|s)\right]$$

Given s: $\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s) \cdot b(s)\right]$ → irrelevant w.r.t. a

$$= \sum_a \pi_\theta(a|s) \cdot \nabla_\theta \log \pi_\theta(a|s) \cdot b(s)$$

$$= \sum_a \nabla_\theta \pi_\theta(a|s) \cdot b(s)$$

$$= \Big(\underbrace{\nabla_\theta \sum_a \pi_\theta(a|s)}_{1}\Big) b(s) = 0$$

$$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\left(Q^{\pi_\theta}(s,a) - b(s)\right)\cdot\nabla_\theta \log \pi_\theta(a|s)\right]$$

baseline

In particular, $b(s) = V^{\pi_\theta}(s)$

$$Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) := A^{\pi_\theta}(s,a)$$

advantage

$$V^{\pi_\theta}(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[Q^{\pi_\theta}(s,a)\right]$$

$\Downarrow$

③ $$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim d_\mu^{\pi_\theta}}\left[A^{\pi_\theta}(s,a)\cdot\nabla_\theta \log \pi_\theta(a|s)\right]$$

$$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_\theta}(s,a)\right] = 0$$

PD Lemma

$$\widehat{\nabla_\theta V^{\pi_\theta}(\mu)} = \sum_{t=0}^{\infty} \gamma^t \cdot \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot \left(\widehat{Q}^{\pi_\theta}(s_t,a_t) - \boxed{f(s_t)}\right) +$$

Still unbiased! (with reduced variance)

policy : $\pi_\theta \to \theta$

value function : $Q^{\pi_\theta} \to \widehat{Q} \approx f_\delta \to \delta$ (δ)

actor – critic

policy based

value based

Similar to : importance sampling

$\to$ doubly robust

Ref : From IS to DR PG. (ICML-20)

Optimality

Recall: starting state distribution ρ: $\max_{\theta \in \Theta} V^{\pi_\theta}(\rho)$

given

$\downarrow$

$\rightarrow$ starting state distribution μ: $\max_{\theta \in \Theta} V^{\pi_\theta}(\mu)$

$\rightarrow$ non-convex (in general)

$$\theta_{t+1} = \theta_t + \eta_t \cdot \nabla_\theta V^{\pi_\theta}(\mu)$$

Guarantee: ε-approximate optimal point.

$$V^{(T)}(\rho) \geq V^*(\rho) - \varepsilon(T)$$

$\frac{1}{T}$, $\frac{1}{\sqrt{T}}$

β-smooth $\Rightarrow$ ε-approximate stationary point $\|\nabla V^{(t)}(\mu)\| \leq \varepsilon$

① Certain condition of μ $\Rightarrow$ lower bound

② canonical softmax policy: $\pi_\theta(a|s) = \dfrac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$

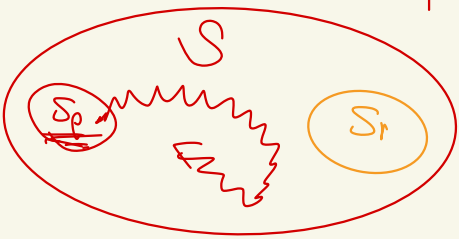
$\rightarrow$ global convergence $V^{(t)}(\rho) \rightarrow V^*(\rho)$ [Asymptotic]

③ canonical softmax policy + log-barrier regularization

$\rightarrow$ polynomial rate of convergence $\left(\frac{1}{\sqrt{T}} \text{ or } \frac{1}{\varepsilon^2}\right)$

Target: $V^{\pi_\theta}(\rho)$

Run algorithm: $V^{\pi_\theta}(\mu)$



$S_r = \{s : \exists a \text{ s.t. } r(s,a) > 0\}$

Sparse-reward MDP

$V^{\pi_\theta}(\rho) = 0$ ← sub-optimal

$\Rightarrow \nabla V^{\pi_\theta}(\rho) = 0$

Need Exploration !

(online RL)

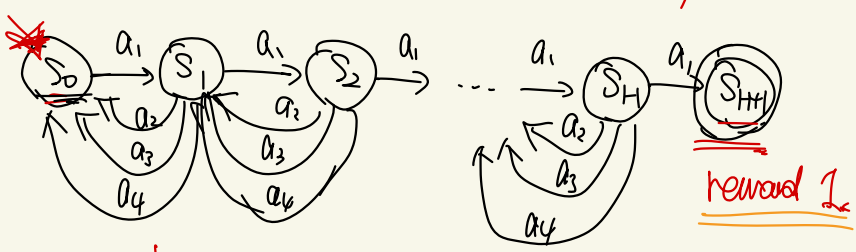
Assumption of μ:

$\forall s : \mu(s) > 0$

(coverage / support)

Chain MDP

$$r(s,a) = \mathbb{1}\{s = s_H,\ a = a_1\}$$



1) $H+1 = \frac{1}{1-\gamma} \Rightarrow \gamma = \frac{H}{H+1}$.

2) Given starting distribution $\rho = s_0$

3) Direct policy parameterization: $\pi_\theta(a|s) = \theta_{s,a}$

$0 < \theta_{s,a} < 1$. $\sum_{i=1}^{4} \theta_{s,a_i} = 1$ ($3|S|$ free parameters)

Proposition (Vanishing Gradient at suboptimal parameters)

Suppose θ is such that $\forall s$: $\theta_{s,a_1} < \frac{1}{4}$

$$\forall k \le \frac{H}{40 \cdot \log(2H)} - 1 : \begin{cases} \| \nabla_\theta^k V^{\pi_\theta}(s_0) \| \le \left(\frac{1}{3}\right)^{\frac{H}{4}} \\ V^*(s_0) - V^{\pi_\theta}(s_0) \ge \frac{H+1}{8} - \frac{(H+1)^2}{3^H} \end{cases}$$

[Agarwal et al. 2020]

k-th order tensor $J \in \mathbb{R}^{d^{\otimes k}}$.

$$\|J\| = \sup_{\substack{u_1, \dots, u_k \in \mathbb{R}^d, \\ \|u_i\|_2 = 1}} \langle J, u_1 \otimes \cdots \otimes u_d \rangle.$$

Remark:

1) Exact Gradient

→ vanishing gradients

× estimation error

2) Gradient, Hessian, ..., k-th order derivative

A ~~escaping~~ from saddle point (deep learning theory)

3) Information-theoretic lower bound

→ difficulty of the problem instance.

$\exists$ difficult MDP. $\forall$ algorithm A: Subopt-Gap $\geq \delta$.

Exploration-based algorithm: E^3. RMax

$\rightarrow$ polynomial guarantee in chain MDP

Prop. is a lower bound specific to PG methods.

4) Initial state distribution $\rho \Rightarrow \mu$

(explorative)

Recall: gradient would be small if
- π is near optimal
- π does NOT visit advantageous states. (i.e. lack of exploration)

Canonical Softmax Parameterization

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

$\theta \in \mathbb{R}^{|S \times A|}$ S_r

lazy evaluation

offline: $D = \{(s, a, r, s')\}$

omine: $\{s_0, a_0, r_0, s_1, a_1, r_1, s_2, \dots\}$

Lemma For softmax policy class,

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s',a'}} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s') \cdot \pi_\theta(a'|s') \cdot A^{\pi_\theta}(s',a')$$

$(= Q^{\pi_\theta}(s',a') - V^{\pi_\theta}(s'))$

Proof: By policy gradient theorem,

$$\boxed{\nabla_\theta V^{\pi_\theta}(\mu)} = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d_\mu^{\pi_\theta},\, a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_\theta}(s,a) \cdot \boxed{\nabla_\theta \log \pi_\theta(a|s)} \right]$$

$/ Q^{\pi_\theta}(s,a)$ $\quad \in \mathbb{R}^{|S \times A|}$.

$$\log \pi_\theta(a|s) = \theta_{s,a} - \log\left(\sum_{a''} \exp(\theta_{s,a''})\right)$$

$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_{s',a'}} = \mathbb{1}\{s'=s\}\cdot\left(\mathbb{1}\{a'=a\} - \frac{\exp(\theta_{s,a'})}{\sum_{a'}\exp(\theta_{s,a'})}\right)$$

$$= \mathbb{1}\{s'=s\}\cdot\left(\mathbb{1}\{a'=a\} - \pi_\theta(a'|s)\right)$$

$$\Rightarrow \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s',a'}} = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_\theta}(s,a)\cdot\mathbb{1}\{s'=s\}\cdot\left(\mathbb{1}\{a'=a\} - \pi_\theta(a'|s)\right)\right]$$

$$= \frac{1}{1-\gamma}\, d_\mu^{\pi_\theta}(s')\cdot\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A^{\pi_\theta}(s',a)\cdot\left(\mathbb{1}\{a'=a\} - \pi_\theta(a'|s)\right)\right]$$

① $\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A^{\pi_\theta}(s',a)\cdot\mathbb{1}\{a'=a\}\right] = \sum_a \pi_\theta(a|s')\cdot A^{\pi_\theta}(s',a)\cdot\mathbb{1}\{a'=a\}$

$$= \pi_\theta(a'|s')\cdot A^{\pi_\theta}(s',a')$$

② $\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A^{\pi_\theta}(s',a)\cdot\pi_\theta(a'|s')\right] = \pi_\theta(a'|s')\cdot\underbrace{\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[A^{\pi_\theta}(s',a)\right]}_{\hookrightarrow\, 0} = 0$

$$A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s',a) - V^{\pi_\theta}(s') \qquad \square$$

$$= Q^{\pi_\theta}(s',a) - \mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[Q^{\pi_\theta}(s',a)\right]$$

$$\theta_{t+1} = \theta_t + \eta\cdot\nabla_\theta V^{(t)}(\mu)$$

$$\Leftrightarrow \pi_{t+1}(a|s) \propto \pi_t(a|s)\cdot\exp\left(\eta\cdot\pi_t(a|s)\cdot A^{(t)}(s,a)\right)$$

$$\propto \pi_t(a|s)\cdot\exp\left(\eta\cdot\pi_t(a|s)\cdot Q^{(t)}(s,a)\right)$$ → per-state → lazy evaluation.

Note: this is different to $\pi_{t+1}(a|s) \propto \pi_t(a|s)\exp(\eta\cdot A^{(t)}(s,a))$

(by NPG : Hedge-style)

Theorem (Global convergence) [Agarwal et al. 2020]

Assume distribution μ is such that $\mu(s) > 0 \quad \forall s\in S$

For $\eta \le (1-\gamma)^3/8$: $V^{(t)}(s) \to V^*(s)$ as $t\to\infty \quad \forall s\in S$

Remarks: 1) $\mu(s) > 0, \forall s \in S$ coverage (support)

Guess: $d_\mu^{\pi_\theta}(s)$ very small for states $s \in S_{eff}$

$\rightarrow$ may NOT globally converge if NOT holds.

2) asymptotic

Guess: worst case $\rightarrow$ exponentially slow. $(|S|, |A|)$

~~poly~~

Proof (sketch): monotone improvement ☆

$\pi_{t+1} \leftarrow \Pi_Q \pi_t$ (policy iteration)

$\forall s: \; V^{(t+1)}(s) \geq V^{(t)}(s)$

$\rightarrow$ limited value function: $V^{(\infty)}(s)$, $Q^{(\infty)}(s,a)$

It suffices to show that $V^{(\infty)}(s) = V^*(s)$

$$\boxed{I_s^0 = \{a : Q^{(\infty)}(s,a) = V^{(\infty)}(s)\}}$$

convergence.

$$I_s^+ = \{a : Q^{(\infty)}(s,a) > V^{(\infty)}(s)\} \rightarrow |I_s^+| = 0 \quad \triangle$$

$$I_s^- = \{a : Q^{(\infty)}(s,a) < V^{(\infty)}(s)\}$$

(no positive advantage)

$|I_s^-| = 0$: otherwise contradicts to monotone improvement.

$\mu(s) > 0 \Rightarrow d_\mu^{\pi_\theta}(s) > 0$

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} \cdot \underbrace{d_\mu^{\pi_\theta}(s)}_{>0} \cdot \underbrace{\pi_\theta(a|s)}_{>0} \cdot \underbrace{A^{\pi_\theta}(s,a)}_{\rightarrow 0} = \pi_\theta(a|s) \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \pi_\theta(a|s)} \quad \square$$

$\rightarrow 0$

Polynomial Guarantee with Log-Barrier Regularization

Softmax: $\pi_\theta(a|s) = \dfrac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$

def (Relative entropy) $KL(p \| q) = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right]$.

Unif_A

Log-barrier regularized objective:

$$\max_{\theta\in\Theta} L_\lambda(\theta) := V^{\pi_\theta}(\mu) - \lambda \cdot \mathbb{E}_{s\sim \mathrm{Unif}_S}\left[KL(\mathrm{Unif}_A \| \pi_\theta(\cdot|s))\right]$$

stochasticity

exploration

$$-\sum_s \frac{\lambda}{|S|} \cdot \sum_a \frac{1}{|A|} \cdot \log \frac{|A|}{\pi_\theta(a|s)}$$

$$= \frac{\lambda}{|S|\cdot|A|} \cdot \sum_{s,a} \log \pi_\theta(a|s) + \lambda \cdot \log|A|$$

$$L_\lambda(\theta) = V^{\pi_\theta}(\mu) + \underbrace{\frac{\lambda}{|S|\cdot|A|} \cdot \sum_{s,a} \log \pi_\theta(a|s)}_{<0} + \cancel{\lambda \cdot \log|A|}$$

Note: Entropy $H(p) = -\mathbb{E}_{x\sim p}[\log p(x)]$.

$$\frac{1}{|S|} \sum_s H(\pi_\theta(\cdot|s)) = \frac{1}{|S|} \sum_{s,a} -\pi_\theta(a|s) \log \pi_\theta(a|s)$$

far less aggressive in penalizing small probabilities!

Algorithm: $\theta_{t+1} = \theta_t + \eta \cdot \nabla_\theta L_\lambda(\theta_t)$ $(*)$

Theorem Suppose θ is such that $\boxed{\|\nabla_\theta L_\lambda(\theta)\|_2 \le \varepsilon_{opt}}$.

and $\varepsilon_{opt} \le \dfrac{\lambda}{2|S|\cdot|A|}$

approximate stationary point

$\Downarrow$

approximate optimal solution

$\Rightarrow$ for all starting state distributions ρ:

$$V^{\pi_\theta}(\rho) \ge V^*(\rho) - \frac{2\lambda}{1-\gamma} \cdot \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty$$

distribution mismatch coefficient (coverage param).

Corollary (Sample Complexity)

$$\beta_\lambda = \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|S|}$$

$L_\lambda(\theta)$ smoothness

Need to be proved

$(*)$ with $\lambda = \dfrac{\varepsilon\cdot(1-\gamma)}{2\cdot\|d_\rho^{\pi^*}/\mu\|_\infty}$, $\eta = \dfrac{1}{\beta_\lambda}$. then:

$$\forall \rho: \underbrace{\min_{t\le T} \{V^*(\rho) - V^{(t)}(\rho)\} \le \varepsilon}_{\varepsilon\text{-optimal value function}} \text{ when } T \ge \frac{320|S|^2|A|^2}{(1-\gamma)^6\varepsilon^2}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty^2 = O(1/\varepsilon^2)$$

Proof of Corollary: By smooth-optimization:

$$\min_{t\le T} \|\nabla_\theta L_\lambda(\theta_t)\|_2^2 \le \frac{2\beta_\lambda\cdot(L_\lambda(\theta^*) - L_\lambda(\theta_0))}{T}$$

$$\Rightarrow \min_{t \le T} \|\nabla_\theta L_\lambda(\theta_t)\|_2 \le \boxed{\frac{2\beta_\lambda}{T(1-\gamma)}}$$

Assume $r \in [0,1]$: $L_\lambda(\theta) \in [0, \frac{1}{1-\gamma}]$

$$\varepsilon_{opt} \le \frac{\lambda}{2\cdot|S|\cdot|A|} \Rightarrow \boxed{T \ge \frac{8\beta_\lambda |S|^2|A|^2}{(1-\gamma)\lambda^2}}$$

upper bound

$$\Rightarrow \frac{8\beta_\lambda |S|^2|A|^2}{(1-\gamma)\cdot\lambda^2} \le \frac{64|S|^2|A|^2}{(1-\gamma)^4\lambda^2} + \frac{16|S|\cdot|A|^2}{(1-\gamma)\cdot\lambda}$$

$$\le \frac{80|S|^2\cdot|A|^2}{(1-\gamma)^4\cdot\lambda^2} = \frac{320|S|^2|A|^2}{(1-\gamma)^6\cdot\varepsilon^2}\cdot\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty^2 \quad \square$$

If $T \ge \frac{320|S|^2|A|^2}{(1-\gamma)^6\cdot\varepsilon^2}\cdot\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty^2 \Rightarrow$ ε-approximate optimal solution.

$= O(1/\varepsilon^2)$ → convergence rate: $O(1/\sqrt{T})$

Recall (PD Lemma) $\forall \pi, \pi'$. $\forall s \to \rho$.

$$\boxed{V^{\pi'}(\rho) - V^{\pi}(\rho) = \frac{1}{1-\gamma}\cdot \mathbb{E}_{s'\sim d_\rho^{\pi'}, a\sim\pi'(\cdot|s)}[A^\pi(s,a)]}$$

(Proof see MDP Planning)

We want to show: $\|\nabla_\theta L_\lambda(\theta)\|_2 \le \varepsilon_{opt} \le \frac{\lambda}{2|S|\cdot|A|}$.

$$\Rightarrow \forall \rho: V^{\pi_\theta}(\rho) \ge V^*(\rho) - \frac{2\lambda}{1-\gamma}\cdot\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty.$$

Claim: $\forall s$: $\max_a A^{\pi_\theta}(s,a) \le \frac{2\lambda}{\mu(s)\cdot|S|}$.

Proof of Theorem: $\forall \rho$: By PD Lemma

$$V^*(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\sum_{s,a} d_\rho^{\pi^*}(s)\cdot \pi^*(a|s)\cdot A^{\pi_\theta}(s,a)$$

$$\le \frac{1}{1-\gamma}\cdot\sum_s d_\rho^{\pi^*}(s)\cdot \max_a A^{\pi_\theta}(s,a)$$

$$\le \frac{1}{1-\gamma}\cdot\sum_s 2\, d_\rho^{\pi^*}(s)\cdot\frac{\lambda}{\mu(s)\cdot|S|}$$

$$\le \frac{2\lambda}{1-\gamma}\cdot\max_s\cdot\frac{d_\rho^{\pi^*}(s)}{\mu(s)} = \frac{2\lambda}{1-\gamma}\cdot\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty \quad \square$$

Proof of claim: It suffices to show $\forall (s,a)$ s.t. $A^{\pi_\theta}(s,a) > 0$

$$\Rightarrow A^{\pi_\theta}(s,a) \le \frac{2\lambda}{\mu(s)\cdot|S|}$$

$$L_\lambda(\theta) = V^{\pi_\theta}(\mu) + \frac{\lambda}{|S|\cdot|A|}\cdot\sum_{s',a'}\log\pi_\theta(a'|s')$$

$$\frac{\partial L_\lambda(\theta)}{\partial\theta_{s,a}} = \frac{1}{1-\gamma}\, d_\mu^{\pi_\theta}(s)\cdot\pi_\theta(a|s)\cdot A^{\pi_\theta}(s,a) + \frac{\lambda}{|S|}\cdot\left(\frac{1}{|A|} - \pi_\theta(a|s)\right)$$

$$\frac{\partial\log\pi_\theta(a|s)}{\partial\theta_{s',a'}} = \mathbb{1}\{s'=s\}\cdot\left(\mathbb{1}\{a'=a\} - \pi_\theta(a|s)\right)$$

By gradient norm assumption $\|\nabla_\theta L_\lambda(\theta)\|_2 \le \varepsilon_{opt}$;

$$\frac{\lambda}{2|S|\cdot|A|} \geq \varepsilon_{opt} \geq \frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma} d_\mu^{\pi_\theta}(s)\cdot \pi_\theta(a|s)\cdot \underbrace{A^{\pi_\theta}(s,a)}_{>0} + \frac{\lambda}{|S|}\left(\frac{1}{|A|} - \pi_\theta(a|s)\right)$$

$$\geq \frac{\lambda}{|S|}\cdot\left(\frac{1}{|A|} - \pi_\theta(a|s)\right)$$

$$\Rightarrow \pi_\theta(a|s) \geq \frac{1}{|A|} - \frac{\varepsilon_{opt}|S|}{\lambda} \geq \frac{1}{2|A|}.$$

$$A^{\pi_\theta}(s,a) = \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)\cdot\pi_\theta(a|s)}\cdot\left(\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} - \frac{\lambda}{|S|}\cdot\left(\frac{1}{|A|} - \pi_\theta(a|s)\right)\right)$$

$$= \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\cdot\left(\frac{1}{\pi_\theta(a|s)}\cdot\frac{\partial L_\lambda(\theta)}{\partial \theta_{s,a}} + \frac{\lambda}{|S|}\cdot\left(1 - \frac{1}{\pi_\theta(a|s)\cdot|A|}\right)\right)$$

$$\leq \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\cdot\left(2|A|\cdot\varepsilon_{opt} + \frac{\lambda}{|S|}\right) \leq \frac{\lambda}{2|S|\cdot|A|}$$

$$\leq 2\cdot\frac{1-\gamma}{d_\mu^{\pi_\theta}(s)}\cdot\frac{\lambda}{|S|} \leq \frac{2\lambda}{|S|}\cdot\frac{1}{\mu(s)}$$

$$\left(\begin{aligned} d_\mu^{\pi_\theta}(s) &= (1-\gamma)\cdot\left(\mu(s) + \sum_{t=1}^{\infty} d_{\mu,t}^{\pi_\theta}(s)\right) \\ &\geq (1-\gamma)\cdot\mu(s) \quad \Rightarrow \frac{1-\gamma}{d_\mu^{\pi_\theta}(s)} \leq \frac{1}{\mu(s)}. \end{aligned}\right) \quad \square$$

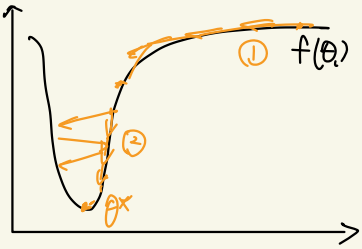
Natural Policy Gradient

Recall: $\min_{\theta\in\Theta} f(\theta)$. $\rightarrow$ GD: $\theta_{t+1} = \theta_t - \eta\cdot\nabla_\theta f(\theta_t)$

vanilla

Pre-conditioned GD : $\theta_{t+1} = \theta_t - \eta \cdot G(\theta_t) \nabla f(\theta_t)$

pre-conditioner



1) $G(\theta_t) = \left(\nabla_\theta^2 H(\theta_t)\right)^{-1}$ → Newton's Method

curvature

local geometry

2) θ parameterize some family P_θ (density)

$G(\theta_t) = \left(F(\theta_t)\right)^{-1}$

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

$\mathbb{E}_{x\sim P_\theta}[\nabla_\theta \log P_\theta(x)] = 0$

Fisher information: $F(\theta) = \mathrm{Var}_{x\sim P_\theta}[\nabla \log P_\theta(x)]$ (FIM)

$= \mathbb{E}_{x\sim P_\theta}[\nabla \log P_\theta(x) \cdot (\nabla \log P_\theta(x))^T]$.

geometry over P-space

$\Rightarrow \boxed{\theta_{t+1} = \theta_t - \eta \cdot F(\theta_t)^{-1} \nabla_\theta f(\theta_t)}$

Natural Gradient Descent

① Netwon's method $\xrightarrow[\infty]{\text{MLE}}$ NGD

$$P_{t+1} = \underset{P \in \mathcal{P}}{\arg\min} \left\{ \langle g_t, P \rangle + \frac{1}{\eta} KL(P_t \| P) \right\}$$

$$g_t = \frac{\delta f}{\delta P}\Big|_{P = P_{\theta_t}}$$

② NGD in parameter-space

$\Leftrightarrow$ Mirror Descent in distribution-space.

$$x_{t+1} = \underset{x \in \mathcal{X}}{\arg\min} \left\{ \langle \nabla f(x_t), x \rangle + \frac{1}{\eta} D_f(x_t, x) \right\}$$

MLE (Maximum Likelihood Estimation)

$$l(\theta; x) = -\log P_\theta(x)$$

$$\nabla^2 l(\theta; x) = -\nabla^2 \log P_\theta(x)$$

$$= -\nabla \left(\frac{\nabla P_\theta(x)}{P_\theta(x)} \right)$$

$$= -\frac{\nabla^2 P_\theta(x) \cdot P_\theta(x) - \nabla P_\theta(x) \cdot \nabla P_\theta(x)^T}{P_\theta(x)^2}$$

$$= \frac{\nabla P_\theta(x)}{P_\theta(x)} \cdot \left(\frac{\nabla P_\theta(x)}{P_\theta(x)} \right)^T - \frac{\nabla^2 P_\theta(x)}{P_\theta(x)}$$

$$= (\nabla \log P_\theta(x))(\nabla \log P_\theta(x))^T - \frac{\nabla^2 P_\theta(x)}{P_\theta(x)}$$

③ $\theta_{t+1} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \underbrace{\langle g_t, P_\theta \rangle}_{\phi_t(\theta)} + \frac{1}{\eta} \underbrace{KL(P_\theta \| P_{\theta_t})}_{K(\theta)} \right\}$

$$\phi_t(\theta) \approx \phi_t(\theta_t) + \langle \nabla_\theta \phi_t(\theta_t), \theta - \theta_t \rangle$$

$$= \langle g_t, P_{\theta_t} \rangle + \langle \nabla_\theta f(\theta_t), \theta - \theta_t \rangle$$

$$K(\theta) \approx \underbrace{K(\theta_t)}_{KL(P_\theta \| P_{\theta_t})} + \underbrace{\langle \nabla K(\theta_t), \theta - \theta_t \rangle}_{0} + \frac{1}{2} (\theta - \theta_t)^T \underbrace{\nabla^2 K(\theta_t)}_{F(\theta_t)\ (\text{FIM})} (\theta - \theta_t)$$

$$= \sum_x P_\theta(x) \cdot \log \frac{P_\theta(x)}{P_{\theta_t}(x)}$$

$$\Rightarrow \theta_{t+1} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \langle \nabla_\theta f(\theta_t), \theta - \theta_t \rangle + \frac{1}{2\eta} \|\theta - \theta_t\|^2_{F(\theta_t)} \right\}$$

$$\Rightarrow \nabla f(\theta_t) + \frac{1}{\eta} F(\theta_t) \cdot (\theta_{t+1} - \theta_t) = 0$$

$$\Rightarrow \text{NGD: } \theta_{t+1} = \theta_t - \eta \cdot F(\theta_t)^{-1} \nabla f(\theta_t)$$

$x \sim P_\theta$. $\Rightarrow$ expected Hessian:

$$\mathbb{E}_{x \sim P_\theta} [\nabla^2 l(\theta; x)]$$

$$= F(\theta) - \underbrace{\sum_x \cancel{P_\theta(x)} \cdot \frac{\nabla^2 P_\theta(x)}{\cancel{P_\theta(x)}}}_{\nabla^2 \left(\underbrace{\sum_x P_\theta(x)}_{=1} \right) = 0}$$

$$= F(\theta)$$

Summary:

① Vanilla GD + local geometry $\rightarrow$ Newton's Method.

② Netwon's Method + MLE $\rightarrow$ NGD

③ OMD with KL regularizer + approximation $\rightarrow$ NGD

with softmax parameterization.

With softmax distribution

$P_\theta \in \Delta_m$ s.t. $P_\theta(i) = \dfrac{\exp(\theta_i)}{\sum_{j=1}^m \exp(\theta_j)}$

$$\log P_\theta(i) = \theta_i - \log \sum_{j=1}^m \exp(\theta_j)$$

score function: $\nabla \log p_\theta(i) = \vec{e}_i - p_\theta$

FIM: $F(\theta) = \mathbb{E}_{i\sim p_\theta}\left[\nabla \log p_\theta(i)\,(\nabla \log p_\theta(i))^T\right]$

$$= \mathbb{E}_{i\sim p_\theta}\left[e_i e_i^T - e_i p_\theta^T - p_\theta e_i^T + p_\theta p_\theta^T\right].$$

$$\mathbb{E}_{i\sim p_\theta}[e_i e_i^T] = \sum_{i=1}^{m} p_\theta(i)\cdot \begin{array}{c} \begin{pmatrix} 0 & \vdots & 0 \\ \cdots & 1 & \cdots \\ 0 & \vdots & 0 \end{pmatrix} \\ \text{i-th} \end{array} \text{i-th} = \begin{pmatrix} p_\theta(1) & & & \\ & p_\theta(2) & & \\ & & \ddots & \\ & & & p_\theta(m) \end{pmatrix}$$

$= \mathrm{diag}(p_\theta)$

$$\mathbb{E}_{i\sim p_\theta}[e_i p_\theta^T] = \sum_{i=1}^{m} p_\theta(i)\cdot e_i \cdot p_\theta^T = \left(\sum_{i=1}^{m} p_\theta(i)\cdot e_i\right)\cdot p_\theta^T = p_\theta \cdot p_\theta^T$$

$\Rightarrow$ FIM: $F(\theta) = \mathrm{diag}(p_\theta) - p_\theta p_\theta^T$.

NGD: $\theta_{t+1} = \theta_t - \eta\cdot (F(\theta_t))^{-1}\nabla f_t(\theta_t)$

$l_t = \dfrac{\partial f_t}{\partial p}$.

$\boxed{f_t(\theta) \triangleq \langle p_\theta, l_t\rangle}$

$= \sum_{i=1}^{m} p_\theta(i)\cdot l_t(i)$

$$\begin{aligned}
\nabla f(\theta) &= \nabla\langle p_\theta, l_t\rangle \\
&= \nabla \sum_{i=1}^{m} p_\theta(i)\cdot l_t(i) \\
&= \sum_{i=1}^{m} \nabla p_\theta(i)\cdot l_t(i) \\
&= \sum_{i=1}^{m} p_\theta(i)\cdot \nabla \log p_\theta(i)\cdot l_t(i) \\
&= \sum_{i=1}^{m} p_\theta(i)\cdot l_t(i)\cdot(\vec{e}_i - \vec{p}_\theta) \\
&= \sum_{i=1}^{m} p_\theta(i)\cdot l_t(i)\cdot \vec{e}_i - \sum_{i=1}^{m} p_\theta(i)\cdot l_t(i)\cdot p_\theta
\end{aligned}$$

score

$$\begin{pmatrix} p_\theta(1)\, l_t(1) \\ p_\theta(2)\, l_t(2) \\ \vdots \\ p_\theta(m)\, l_t(m) \end{pmatrix} = \mathrm{diag}(p_\theta)\cdot l_t$$

$\langle p_\theta, l_t\rangle\, p_\theta = (p_\theta^T l_t)\, p_\theta$

$= p_\theta\cdot(p_\theta^T l_t) = (p_\theta p_\theta^T)\, l_t$

Problem Setting: expert

$l_t = (l_t(1), l_t(2), \cdots, l_t(m))$

$\theta_t \to p_{\theta_t} \in \Delta_m$

expected loss: $f(\theta_t)$

$= \langle p_{\theta_t}, l_t\rangle$.

$$\Rightarrow \nabla f(\theta) = \left(\mathrm{diag}(p_\theta) - p_\theta p_\theta^T\right) \cdot \ell_t$$

$$= F(\theta)\, \ell_t \quad ☆$$

$$☆ \quad \boxed{\nabla_\theta \langle \ell_t, p_\theta \rangle = F(\theta) \cdot \ell_t}$$

Softmax: ℓ_t serves as natural gradient

NGD: $\theta_{t+1} = \theta_t - \eta \cdot F(\theta)^{-1} \nabla f(\theta) = \theta_t - \eta \cdot \ell_t$ ☆

$\Rightarrow$ distribution-space: $p_{t+1}(i) \propto \exp(\theta_{t+1,i})$

$$= \exp(\theta_{t,i} - \eta \cdot \ell_t(i))$$

$$= \exp(\theta_{t,i}) \cdot \exp(-\eta \cdot \ell_t(i))$$

△ Hedge Algorithm for expert problem $\boxed{p_{t+1}(i) \propto p_t(i) \cdot \exp(-\eta \ell_t(i))}$

(a.k.a. Multiplicative weight Update)

OMD: (E: exact)

distribution space: $p_{t+1} = \arg\min_p \langle \ell_t, p \rangle + \frac{1}{\eta} KL(p \| p_t)$

parameter space: $\theta_{t+1} = \arg\min_\theta \langle \ell_t, p_\theta \rangle + \frac{1}{\eta} KL(p_\theta \| p_t)$

1-st order approx + 2-nd order approx

$$\theta_{t+1} = \theta_t - \eta \cdot F(\theta_t)^{-1} \nabla_\theta \langle \ell_t, p_{\theta_t} \rangle \quad \text{NGD}$$

(A: approx) OMD: $\theta_{t+1} = \arg\min_\theta \langle \nabla_\theta \langle \ell_t, p_{\theta_t} \rangle, \theta \rangle + \frac{1}{2\eta} \cdot \|\theta - \theta_t\|^{-1}_{F(\theta_t)}$

With softmax distribution:

(E): $\nabla_\theta \langle \ell_t, p_\theta \rangle + \frac{1}{\eta} F(\theta)(\theta - \theta_t) = 0$ $\quad F(\theta) \cdot \ell_t$

(A): $\nabla_\theta \langle \ell_t, p_{\theta_t} \rangle + \frac{1}{\eta} \cdot F(\theta_t)(\theta - \theta_t) = 0$ $\quad F(\theta_t) \cdot \ell_t$

$\Rightarrow \theta_{t+1} = \theta_t - \eta \cdot \ell_t$ (identical)

Natural Policy Gradient

We do NOT replace ρ with μ.

$$\Pi_\theta = \{ \pi_\theta(\cdot|s) : s \in S \}$$

FIM: $$F_\rho(\theta) = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)} \left[(\nabla \log \pi_\theta(a|s)) \cdot (\nabla \log \pi_\theta(a|s))^T \right]$$

NPG: $$\theta_{t+1} = \theta_t + \eta \cdot (F_\rho(\theta_t))^+ \cdot \nabla_\theta V^{(t)}(\rho)$$

(*)

in offline RL: PSPI

in online RL: ONPG

Lemma (Softmax NPG as Soft Policy Iteration)

parameter space

$$\theta_{t+1} = \theta_t + \frac{\eta}{1-\gamma} \cdot A_t + \eta \cdot v$$

W^*

state-dependent offset.

$Q_t(s,a)$

policy space

$$\pi_{t+1}(a|s) = \pi_t(a|s) \cdot \frac{\exp(\eta A_t(s,a)/(1-\gamma))}{Z_t(s)}$$

where $$Z_t(s) = \sum_a \pi_t(a|s) \cdot \exp(\eta A_t(s,a)/(1-\gamma))$$

Proof: $$(F_\rho(\theta))^+ \nabla_\theta V^{\pi_\theta}(\rho) = W^*$$

natural gradient if and only if

W^* is the minimum-norm solution of the following least-square regression:

$$\min_w \| \nabla V^{\pi_\theta}(\rho) - F_\rho(\theta) \cdot w \|^2$$

Compatible Function Approximation.

$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_{s',a'}} = \mathbb{1}\{s'=s\} \cdot (\mathbb{1}\{a'=a\} - \pi_\theta(a'|s))$$

$$\Rightarrow W^T \nabla_\theta \log \pi_\theta(a|s) = W_{s,a} - \sum_{a'} \pi_\theta(a'|s) \cdot W_{s,a'} \rightarrow \bar{W}_s$$

$$\Rightarrow F_\rho(\theta)\cdot w = \mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\nabla\log\pi_\theta(a|s)\,\nabla\log\pi_\theta(a|s)^T\cdot w\right]$$

$$= \mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\nabla\log\pi_\theta(a|s)\cdot(w_{s,a}-\bar{w}_s)\right]$$

indep to a

$$= \mathbb{E}_{s\sim d_\rho^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\nabla\log\pi_\theta(a|s)\cdot w_{s,a}\right]$$

$$\left(\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla\log\pi_\theta(a|s)\right] = \nabla\sum_a\pi_\theta(a|s) = \nabla 1 = 0\right)$$

$$\left[F_\rho(\theta)\cdot w\right]_{s',a'} = d_\rho^{\pi_\theta}(s')\cdot\pi_\theta(a'|s')\cdot\left(w_{s',a'} - \sum_{a''}\pi_\theta(a''|s')\cdot w_{s',a''}\right)$$

$$= d_\rho^{\pi_\theta}(s')\cdot\pi_\theta(a'|s')\cdot\left(w_{s',a'} - \bar{w}_{s'}\right)$$

$$\left[\nabla V^{\pi_\theta}(\rho)\right]_{s',a'} = d_\rho^{\pi_\theta}(s')\cdot\pi_\theta(a'|s')\cdot\frac{1}{1-\gamma}A^{\pi_\theta}(s,a)$$

$$\Rightarrow \left\|\nabla V^{\pi_\theta}(\rho) - F_\rho(\theta)\cdot w\right\|^2 = \sum_{s,a}\left(d_\rho^{\pi_\theta}(s)\,\pi_\theta(a|s)\underbrace{\left(\frac{A^{\pi_\theta}(s,a)}{1-\gamma} - w_{s,a} + \bar{w}_s\right)}_{=0}\right)^2$$

$$w_{s,a} = \frac{A^{\pi_\theta}(s,a)}{1-\gamma} \rightarrow 0$$

$\Rightarrow$ all 0-error solution: $w = \dfrac{A^{\pi_\theta}}{1-\gamma} + \underline{\underline{v}}$ (state-dependent offset) $\square$

Remark: $\theta_{t+1} = \theta_t + \eta\cdot\left(\overset{s\sim d_\rho^{\pi_\theta}}{F_\rho(\theta_t)}\right)^{-1}\nabla\overset{s\sim d_\rho^{\pi_\theta}}{V^{(t)}(\rho)}$

+ softmax $\rightarrow$ $\boxed{\theta_{t+1} = \theta_t + \frac{\eta}{1-\gamma}\cdot A_t + \underline{v}}$ $\rightarrow$ per-state (distribution-independent)

Actually, NPG update (in softmax) has nothing to do with state distribution!

Theorem (Global convergence of NPG) $\rho \in \Delta_S$. $\theta_0 = 0$.

$$V^{(T)}(\rho) \geq V^*(\rho) - \underbrace{\frac{\log|A|}{\eta T} - \frac{1}{(1-\gamma)^2 T}}_{\varepsilon}$$

Remarks: 1) holds for arbitrary distribution ρ.

2). $\eta \geq (1-\gamma)^2 \log|A|$ $\Rightarrow$ sample-complexity guarantee:

NPG finds an ε-approximate optimal solution within $T \geq \frac{2}{(1-\gamma)^2 \varepsilon}$ (independent to $|A|, |S|$)

$\Rightarrow 1/T \leftarrow$

---

* By Hedge algorithm guarantee: $\mathrm{Reg}_K \leq \sqrt{T \log|A|} \leq T\varepsilon$.

$\Leftrightarrow T \geq \frac{\log|A|}{\varepsilon^2}$.

$\Rightarrow 1/\sqrt{T} \leftarrow$

Lemma (Monotone Improvement) $\forall \rho \in \Delta_S$:

$$V^{(t+1)}(\rho) - V^{(t)}(\rho) \geq \frac{1-\gamma}{\eta} \cdot \mathbb{E}_{s\sim\rho}[\underbrace{\log Z_t(s)}_{\text{log-partition}}] \geq 0$$

Proof of Lemma: first show $\log Z_t(s) \geq 0$.

$$\log Z_t(s) = \log \sum_a \pi_t(a|s) \cdot \exp\left(\frac{\eta A_t(s,a)}{1-\gamma}\right)$$

$$\overset{\text{Jenson}}{\geq} \sum_a \pi_t(a|s) \cdot \log \exp\left(\frac{\eta A_t(s,a)}{1-\gamma}\right)$$

$$= \frac{\eta}{1-\gamma} \cdot \sum_a \pi_t(a|s) \cdot A_t(s,a) = 0$$

advantage.

By PD Lemma : $V^{(t+1)}(\rho) - V^{(t)}(\rho)$ $\qquad \pi_{t+1}(a|s) = \pi_t(a|s) \cdot \frac{\exp(\frac{\eta}{1-\gamma} A_t(s,a))}{Z_t(s)}$

$$= \frac{1}{1-\gamma} \cdot \mathbb{E}_{s\sim d^{(t+1)}}\left[\sum_a \pi_{t+1}(a|s) \cdot A_t(s,a) \right]$$

$$= \frac{1}{\eta} \cdot \mathbb{E}_{s\sim d_\rho^{(t+1)}}\left[\underbrace{\sum_a \pi_{t+1}(a|s)}_{\to 1} \cdot \log \frac{\pi_{t+1}(a|s) \cdot Z_t(s)}{\pi_t(a|s)} \right]$$

$$= \frac{1}{\eta} \cdot \mathbb{E}_{s\sim d_\rho^{(t+1)}}\left[\underbrace{KL(\pi_{t+1}(\cdot|s) \,\|\, \pi_t(\cdot|s))}_{\geq 0} \right]$$

$$+ \frac{1}{\eta} \mathbb{E}_{s\sim d_\rho^{(t+1)}}\left[\log Z_t(s) \right]$$

$$\geq \frac{1}{\eta} \mathbb{E}_{s\sim d_\rho^{(t+1)}}\left[\log Z_t(s) \right] \geq \frac{1-\gamma}{\eta} \cdot \mathbb{E}_{s\sim\rho}\left[\log Z_t(s) \right] \quad \square$$

Proof of Theorem : $V^{(T)}(\rho) \geq V^*(\rho) - \frac{\log|A|}{\eta T} - \frac{1}{(1-\gamma)^2 T}$.

PD Lemma : $V^*(\rho) - V^{(t)}(\rho) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s\sim d_\rho^*}\left[\sum_a \pi^*(a|s) \cdot A_t(s,a) \right]$.

$$= \frac{1}{\eta} \cdot \mathbb{E}_{s\sim d_\rho^*}\left[\sum_a \pi^*(a|s) \cdot \log \frac{\pi_{t+1}(a|s) Z_t(s) \cdot \pi^*(a|s)}{\pi_t(a|s) \cdot \pi^*(a|s)} \right]$$

$$= \frac{1}{\eta} \mathbb{E}_{s\sim d_\rho^*}\left[KL(\pi^*(\cdot|s) \,\|\, \pi_t(\cdot|s)) - KL(\pi^*(\cdot|s) \,\|\, \pi_{t+1}(\cdot|s)) + \log Z_t(s) \right]$$

A $\boxed{\frac{1}{\eta} \mathbb{E}_{s\sim d_\rho^*}\left[\log Z_t(s) \right] \leq \frac{1}{1-\gamma} \cdot \left(V^{(t+1)}(d_\rho^*) - V^{(t)}(d_\rho^*) \right)}$

$$\Rightarrow V^*(\rho) - V^{(T)}(\rho) \leq V^*(\rho) - V^{(T-1)}(\rho)$$

$$\leq \frac{1}{T}\sum_{t=0}^{T-1}\left(V^*(\rho)-V^{(t)}(\rho)\right) \quad \text{(by monotone improvement)}$$

$$= \frac{1}{\eta T}\cdot\sum_{t=0}^{T-1}\mathbb{E}_{s\sim d_\rho^*}\left[KL(\pi^*(\cdot|s)\,\|\,\pi_t(\cdot|s)) - KL(\pi^*(\cdot|s)\,\|\,\pi_{t+1}(\cdot|s))\right]$$

$$+ \frac{1}{(1-\gamma)T}\cdot\sum_{t=0}^{T-1}\left(V^{(t+1)}(d_\rho^*)-V^{(t)}(d_\rho^*)\right)$$

$$\leq \frac{1}{\eta T}\cdot\mathbb{E}_{s\sim d_\rho^*}\left[\underline{KL(\pi^*(\cdot|s)\,\|\,\pi_0(\cdot|s))}\right] \leq \log|A|$$

uniform

$$+ \frac{1}{(1-\gamma)T}\cdot\left(\underline{V^{(T)}(d_\rho^*)-V^{(0)}(d_\rho^*)}\right) \leq \frac{1}{1-\gamma}$$

$\gamma\in[0,1]$

$$\leq \frac{\log|A|}{\eta T} + \frac{1}{(1-\gamma)^2 T} \quad \square.$$

Function Approximation and Distribution Shift

parametric policy class : $\Pi = \{\pi_\theta : \theta\in\mathbb{R}^d\}$

restricted

→ expressivity $\quad d \ll |S\times A|$

- Agnostic setting : $\max_{\theta\in\mathbb{R}^d} V^{\pi_\theta}(\rho)$ < learned policy / best policy in policy class

- NPG : $\theta_{t+1} = \theta_t + \eta\cdot\left(F_\rho(\theta)\right)^{+}\nabla_\theta V^{\pi_\theta}(\rho)$

① NPG update $\Leftrightarrow$ CFA (Compatible Function Approximation)

② Analysis of Q-NPG under log-linear policy

③ Analysis of NPG (bias-variance trade-off)

[Agarwal et al. 2020]

Lemma (Gradient and CFA) Let w^* be the following minimizer:

$$\boxed{w^*} \in \underset{w \in \mathbb{R}^d}{\arg\min} \; \mathbb{E}_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\left(A^{\pi_\theta}(s,a) - w^T \nabla_\theta \log \pi_\theta(a|s)\right)^2\right]$$

LS error of compatible function approximation.

$\Downarrow$

Best linear predictor of the advantage function:

$$\hat{A}^{\pi_\theta}(s,a) = (w^*)^T \nabla \log \pi_\theta(a|s).$$

Then: $$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\left[\hat{A}^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a|s)\right]$$

Proof: By optimality of w^*:

$$\mathbb{E}_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\Big[\Big(A^{\pi_\theta}(s,a) - \overbrace{(w^*)^T \nabla \log \pi_\theta(a|s)}^{\hat{A}^{\pi_\theta}(s,a)}\Big) \cdot \nabla \log \pi_\theta(a|s)\Big] = 0.$$

$$\Rightarrow \nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s\sim d_\mu^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\Big[\underbrace{A^{\pi_\theta}(s,a) \cdot \nabla \log \pi_\theta(a|s)}_{\hat{A}^{\pi_\theta}(s,a) \cdot \nabla \log \pi_\theta(a|s)}\Big] \quad \square$$

NPG ascent direction!

Compatible $\langle$ value (advantage) $\leftarrow A^{\pi_\theta}(s,a)$; policy $\rightarrow \nabla \log \pi_\theta(a|s)$

Actor - Critic

Lemma $\quad w^* \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; \mathbb{E}_{s \sim d_\rho^{\pi_\theta},\, a \sim \pi_\theta(\cdot|s)} \left[\left(A^{\pi_\theta}(s,a) - w^T \nabla \log \pi_\theta(a|s) \right)^2 \right]$

$\Rightarrow$ Natural gradient: $\boxed{(F_\rho(\theta))^+ \nabla_\theta V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} w^*}$

Proof: $\nabla V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi_\theta},\, a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_\theta}(s,a) \cdot \nabla \log \pi_\theta(a|s) \right]$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi_\theta},\, a \sim \pi_\theta(\cdot|s)} \left[\nabla \log \pi_\theta(a|s) \cdot \nabla \log \pi_\theta(a|s)^T w^* \right]$$

$$= \frac{1}{1-\gamma} \cdot F_\rho(\theta) \cdot w^*. \qquad \square$$

$\Downarrow$

NPG update: $\theta \leftarrow \theta + \frac{\eta}{1-\gamma} \cdot w^*$

minimizer of CFA error.

In general: $\theta_{t+1} = \theta_t + \eta \cdot w_t$

Examples: NPG & Q-NPG.

$$\Pi = \left\{ \pi_\theta(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_{a'} \exp(f_\theta(s,a'))} : \theta \in \mathbb{R}^d \right\}$$

$f_\theta(s,a)$: function approximator (e.g. linear, neural nets)

Log-linear policy

$\forall (s,a)$, $\exists$ feature $\phi_{s,a} \in \mathbb{R}^d$.

$$\Rightarrow \pi_\theta(a|s) = \frac{\exp(\theta^T \phi_{s,a})}{\sum_{a'} \exp(\theta^T \phi_{s,a'})}$$

$$\nabla_\theta \log \pi_\theta(a|s) = \phi_{s,a} - \mathbb{E}_{a' \sim \pi_\theta(\cdot|s)}[\phi_{s,a'}] = \bar{\phi}^{\theta}_{s,a}$$

NPG: $\theta \leftarrow \theta + \eta \cdot w_*$,

$$w_* \in \arg\min_{w \in \mathbb{R}^d} \mathbb{E}_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\left(A^{\pi_\theta}(s,a) - w^T \bar{\phi}^{\theta}_{s,a} \right)^2 \right]$$

centered feature

Q-NPG: $\theta \leftarrow \theta + \eta \cdot w_*$

$$w_* \in \arg\min_{w \in \mathbb{R}^d} \mathbb{E}_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\left(Q^{\pi_\theta}(s,a) - w^T \phi_{s,a} \right)^2 \right]$$

original feature

$\downarrow$

NPG / Q-NPG as soft policy iteration:

$\bar{\phi}^{\pi}_{s,a}$

NPG: $\pi(a|s) \leftarrow \pi(a|s) \cdot \exp(\eta \cdot w_*^T \phi_{s,a}) / Z_s$

Q-NPG: $\pi(a|s) \leftarrow \pi(a|s) \cdot \exp(\eta \cdot w_*^T \phi_{s,a}) / Z_s$

$f_\theta(s,a)$ is a neural net, $\nabla_\theta \log \pi_\theta(a|s) = g_\theta(s,a) = \nabla_\theta f_\theta(s,a) - \mathbb{E}_{a' \sim \pi_\theta(\cdot|s)}[\nabla_\theta f_\theta(s,a')]$

NPG: $\theta \leftarrow \theta + \eta \cdot w^*$, $w^* \in \arg\min \mathbb{E}_{s,a}\left[\left(A^{\pi_\theta}(s,a) - w^T g_\theta(s,a) \right)^2 \right]$

Q-NPG: $\theta \leftarrow \theta + \eta \cdot w^*$, $w^* \in \arg\min \mathbb{E}_{s,a}\left[\left(Q^{\pi_\theta}(s,a) - w^T \nabla f_\theta(s,a) \right)^2 \right]$

NPG "Regret Lemma"

Consider NPG more abstractly: $\theta_{t+1} = \theta_t + \eta \, w_t$ "natural gradient" proxy at round t.

For now: arbitrary (bounded) sequence $\{w_t\}_{t=1}^{T}$ s.t. $\|w_t\|_2 \le W$.

Recall: $f: \mathbb{R}^d \to \mathbb{R}$ is β-smooth if $\forall x, x' \in \mathbb{R}^d$

- (Zeroth-order) $|f(x') - f(x) - \nabla f(x) \cdot (x' - x)| \le \frac{\beta}{2} \|x' - x\|_2^2$
- (first-order) $\|\nabla f(x) - \nabla f(x')\|_2 \le \beta \cdot \|x - x'\|_2$
- (second-order) $\nabla^2 f(x) \preceq \beta I$.

Assumption $\log \pi_\theta(a|s)$ is β-smooth w.r.t. θ, $\forall (s,a)$

Lemma (NPG regret lemma) With $\theta_{t+1} = \theta_t + \eta \cdot w_t$, $\theta_0 = \vec{0}$

def $$\mathrm{err}_t = \mathbb{E}_{(s,a)\sim d_\rho^{\tilde{\pi}}} \left[A_t(s,a) - w_t^T \nabla \log \pi_t(a|s) \right]$$

Using step-size $\eta = \sqrt{2\log|A| / \beta W^2 T}$: $\forall$ comparator policy $\tilde{\pi}$

$$\min_{t<T} \left\{ V^{\tilde{\pi}}(\rho) - V^{(t)}(\rho) \right\} \le \frac{1}{1-\gamma} \left(W \cdot \sqrt{\frac{2\beta \log|A|}{T}} + \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{err}_t \right)$$

Remark: 1) OMD with KL regularizer $\to$ Hedge. $W \cdot \sqrt{T \cdot \log|A|}$

online convex optimization (OCO) ×

2) Tabular: $O(1/T)$. $\Leftrightarrow$ FA: $O(1/\sqrt{T})$

Proof: By smoothness of $\log \pi_\theta(a|s)$

$\theta_t \Leftrightarrow \pi_t$

$\theta_{t+1} \Leftrightarrow \pi_{t+1}$

$$\log \pi_{t+1}(a|s) \geq \log \pi_t(a|s) + \nabla \log \pi_t(a|s)^T (\theta_{t+1} - \theta_t) - \frac{\beta}{2} \cdot \|\theta_{t+1} - \theta_t\|^2$$

$$\Rightarrow \log \frac{\pi_{t+1}(a|s)}{\pi_t(a|s)} \geq \nabla \log \pi_t(a|s)^T (\theta_{t+1} - \theta_t) - \frac{\beta}{2} \cdot \|\theta_{t+1} - \theta_t\|^2$$

$$= \eta \cdot \nabla \log \pi_t(a|s)^T w_t - \frac{\beta}{2} \cdot \eta^2 \cdot \|w_t\|^2 \quad (*)$$

By PD Lemma,

$$V^{\tilde{\pi}}(\rho) - V^{(t)}(\rho) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s \sim d_\rho^{\tilde{\pi}}, a \sim \tilde{\pi}(\cdot|s)} [A_t(s,a)]$$

$$= \frac{1}{1-\gamma} \cdot \mathbb{E}_{(s,a) \sim d_\rho^{\tilde{\pi}}} [A_t(s,a) - \nabla \log \pi_t(a|s)^T w_t + \nabla \log \pi_t(a|s)^T w_t]$$

$$= \frac{1}{1-\gamma} \mathrm{err}_t + \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim d_\rho^{\tilde{\pi}}} [\nabla \log \pi_t(a|s)^T w_t]$$

$$\overset{(*)}{\leq} \frac{1}{1-\gamma} \mathrm{err}_t + \frac{1}{\eta \cdot (1-\gamma)} \cdot \mathbb{E}_{(s,a) \sim d_\rho^{\tilde{\pi}}} \left[\log \frac{\pi_{t+1}(a|s)}{\pi_t(a|s)} + \frac{\beta}{2} \eta^2 \|w_t\|^2 \right]$$

$$= \frac{1}{1-\gamma} \left(\mathrm{err}_t + \frac{\beta}{2} \cdot \eta \cdot \|w_t\|^2 + \frac{1}{\eta} \mathbb{E}_{s \sim d_\rho^{\tilde{\pi}}, a \sim \tilde{\pi}(\cdot|s)} \left[\log \frac{\pi_{t+1}(a|s)}{\pi_t(a|s)} \right] \right)$$

$$= \frac{1}{1-\gamma} \left(\mathrm{err}_t + \frac{\beta}{2} \cdot \eta \|w_t\|^2 + \frac{1}{\eta} \mathbb{E}_{s \sim d_\rho^{\tilde{\pi}}} \left[KL(\tilde{\pi}^s \| \pi_t^s) - KL(\tilde{\pi}^s \| \pi_{t+1}^s) \right] \right)$$

Telescoping, $\frac{1}{T} \sum_{t=0}^{T-1} \left(V^{\tilde{\pi}}(\rho) - V^{(t)}(\rho) \right) = \frac{1}{1-\gamma} \cdot \frac{1}{T} \sum_{t=1}^{T} \mathrm{err}_t$

$$+\frac{1}{1-\gamma}\left(\frac{\beta}{2}\cdot\eta\cdot\underbrace{\sum_{t=0}^{T-1}\|w_t\|^2}_{\le W^2T}+\frac{1}{\eta}\underbrace{\mathbb{E}_{s\sim d_\rho^{\tilde\pi}}\left[\tilde\pi(\cdot|s)\,\|\,\underbrace{\pi_0(\cdot|s)}_{\text{Uniform}}\right]}_{\le\log|A|}\right)$$

With $\eta=\sqrt{\frac{2\beta\cdot\log|A|}{W^2T}}$ $\Rightarrow \frac{1}{1-\gamma}\cdot W\cdot\sqrt{\frac{2\beta\log|A|}{T}}$.

$$\Rightarrow \min_{t=T}\left\{V^{\tilde\pi}(\rho)-V^{(t)}(\rho)\right\}\le\frac{1}{T}\sum_{t=0}^{T-1}\left(V^{\tilde\pi}(\rho)-V^{(t)}(\rho)\right)$$

$$\le\frac{1}{1-\gamma}\cdot\left(W\cdot\sqrt{\frac{2\beta\cdot\log|A|}{T}}+\frac{1}{T}\sum_{t=1}^{T}\mathrm{err}_t\right)\quad\square$$

Q-NPG : Performance Bound for Log-Linear Policies.

Algorithm: $\theta_{t+1}=\theta_t+\eta\cdot w_t^*$ (*) $\qquad \pi_\theta(a|s)\propto\exp(\theta^T\phi_{s,a})$

$$w_t^*\in\arg\min_w \mathbb{E}_{d_\mu^{\pi_t},a\sim\pi_t(\cdot|s)}\left[\left(Q^{\pi_t}(s,a)-w^T\phi_{s,a}\right)^2\right]$$

def $L(w;\theta,\mu)=\mathbb{E}_{(s,a)\sim\mu}\left[\left(Q^{\pi_\theta}(s,a)-w^T\phi_{s,a}\right)^2\right]$.

$\Rightarrow w_t^*\in\arg\min_w L(w;\theta_t,d_t)$ $\qquad d_\mu^{\pi_t}$

Approximate Version: $\boxed{w_t\approx\arg\min_w L(w;\theta_t,d_t)}$

Assume $\|w_t\|_2\le W$ (l_2-regularized). $\qquad \theta_{t+1}=\theta_t+\eta\cdot w_t$.

Assumption (Estimation / Transfer Errors)

Suppose N samples, $O(1/\sqrt{N}), O(1/N)$

1. (Excess risk / Estimation error)

$$\mathbb{E}_{w_0, w_1, \cdots, w_{t-1}}\left[L(w_t ; \theta_t, d_t) - L(w_t^* ; \theta_t, d_t) \right] \le \varepsilon_{stat}$$

best linear predictor

2. (Transfer error / Bias)

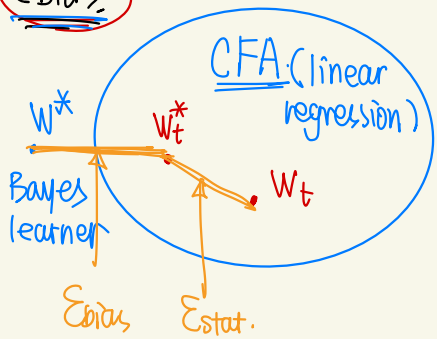
fixed measure

$$\mathbb{E}_{w_0, w_1, \cdots, w_{t-1}}\left[L(w_t^* ; \theta_t, d^*) \right] \le \varepsilon_{bias}$$

Any comparator policy

$d^*(s,a) = d_\rho^{\pi^*}(s) \circ \mathrm{Unif}(A)$

(irrelevant to w_t, $t \in T$)



Assumption (Relative condition number)

$$\Sigma_\nu = \mathbb{E}_{(s,a)\sim\nu}\left[\phi_{s,a} \phi_{s,a}^T \right], \quad \text{then}$$

$$\sup_{w \in \mathbb{R}^d} \frac{w^T \Sigma_{d^*} w}{w^T \Sigma_\nu w} = \kappa \quad \text{and} \quad \kappa \text{ is finite}$$

"coverage" of ν w.r.t. d^*

$\sigma_{max}\left(\Sigma_\nu^{-1/2} \Sigma_{d^*} \Sigma_\nu^{-1/2}\right)$

Remark: 1) log-linear $\Rightarrow$ a set of features $\Phi = \{\phi_{s,a}\}$

$\Rightarrow$ Generalized Rayleigh Quotient.

2) $K \le d$ is always possible.

Löwner-John ellipsoid

Claim: Suppose $\Phi = \{\phi_{s,a}\}$ is compact. $\exists$ some $\nu \in \Delta_{S\times A}$ which is supported on at most d^2 state-action pairs and can be constructed only with knowledge of Φ such that: $K \le d$

(without knowledge of d^* or MDP)

Theorem (Agnostic Learning of Q-NPG)

Fix distribution $\rho \in \Delta_S$. $\nu \in \Delta_{S\times A}$. comparator policy π^*.

Suppose $\|w_t\|_2 \le W$, $\|\phi_{s,a}\|_2 \le B$. $\forall s, a, t$.

With $\theta_{t+1} = \theta_t + \eta \cdot w_t$. $\theta_0 = 0$. $\eta = \sqrt{2\log|A| / (B^2 W^2 T)}$:

$$\mathbb{E}\left[\min_{t<T}\left\{V^{\pi^*}(\rho) - V^{(t)}(\rho)\right\}\right] \le \underbrace{\frac{BW}{1-\gamma}\sqrt{\frac{2\log|A|}{T}}}_{\text{standard OMD results}} + \underbrace{\sqrt{\frac{4|A|K\cdot\varepsilon_{stat}}{(1-\gamma)^3}}}_{\text{estimation error}} + \frac{\sqrt{4|A|\cdot\varepsilon_{bias}}}{1-\gamma}$$

$\sim 1/\sqrt{N}$

Cases that $\varepsilon_{bias} = 0$:

1) Canonical Softmax Policy: $\pi_\theta(a|s) \propto \exp(\theta_{s,a})$

2) "linear" MDP setting (transition dynamics are low-rank) if we know the set of features $\Phi = \{\phi_{s,a}\}$.

Approximation - Estimation error decomposition :

$$L(\underline{w_t}; \theta_t, d_t) = \underbrace{\left(L(w_t; \theta_t, d_t) - L(w_t^*; \theta_t, d_t)\right)}_{\substack{\text{excess risk} \\ \varepsilon_{stat} \text{ (estimation error)}}} + \underbrace{L(w_t^*; \theta_t, d_t)}_{\substack{\cancel{\varepsilon_{bias}} \\ \text{approximation error}}}$$

Corollary (Estimation error / Approximation error bound of Q-NPG)

Approximation error: $\mathbb{E}[\underline{L(w_t^*; \theta_t, d_t)}] \leq \varepsilon_{approx}$.

$$\mathbb{E}\left[\min_{t<T}\{V^{\pi^*}(\rho) - V^{(t)}(\rho)\}\right] \leq \frac{BW}{1-\gamma}\sqrt{\frac{2\log|A|}{T}} + \sqrt{\frac{4|A|}{(1-\gamma)^3}\cdot\left(\kappa\cdot\varepsilon_{stat} + \underline{\left\|\frac{d^*}{\mu}\right\|_\infty}\cdot\varepsilon_{approx}\right)}$$

Proof: $\underline{L(w_t^*; \theta_t, d^*)} \leq \underline{\left\|\frac{d^*}{d_t}\right\|_\infty}\cdot L(w_t^*; \theta_t, d_t)$ $\quad d_\mu^{\pi_t}$

$$\leq \frac{1}{1-\gamma}\cdot\left\|\frac{d^*}{\mu}\right\|_\infty\cdot\underbrace{L(w_t^*; \theta_t, d_t)}_{\varepsilon_{approx}} \qquad \square.$$

Now: Prove theorem

$$\mathbb{E}\left[\min_{t<T}\{V^{\pi^*}(\rho) - V^{(t)}(\rho)\}\right] \leq \frac{BW}{1-\gamma}\sqrt{\frac{2\log|A|}{T}} + \sqrt{\frac{4|A|\kappa\cdot\varepsilon_{stat}}{(1-\gamma)^3}} + \frac{\sqrt{4|A|\cdot\varepsilon_{bias}}}{1-\gamma}$$

Proof: By NPG regret Lemma:

$$\mathbb{E}\left[\min_{t<T}\{V^{\pi^*}(\rho) - V^{(t)}(\rho)\}\right] \leq \frac{W}{1-\gamma}\sqrt{\frac{2\beta\cdot\log|A|}{T}} + \frac{1}{1-\gamma}\cdot\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}err_t\right]$$

CFA error: $err_t = \mathbb{E}_{s\sim d_\rho^{\pi^*}, a\sim\pi^*(\cdot|s)}\left[\underline{A_t(s,a) - w_t^T\nabla\log\pi_t(a|s)}\right]$

① Smoothness of log-linear class:

$$\pi_\theta(a|s) = \frac{\exp(\theta^T \phi_{s,a})}{\sum_{a'} \exp(\theta^T \phi_{s,a'})} \Rightarrow \nabla \log \pi_\theta(a|s) = \phi_{s,a} - \bar{\phi}_s^\theta .$$

$$\|\nabla \log \pi_\theta(a|s)\|_2^2 = \|\phi_{s,a} - \bar{\phi}_s^\theta\|_2^2 \le B^2 .$$

② Bounding the CFA error term ☆

$$\mathrm{err}_t = \mathbb{E}_{s\sim d_\rho^{\pi^*}, a\sim\pi^*(\cdot|s)}\left[A_t(s,a) - (w_t^*)^T \nabla \log \pi_t(a|s)\right] \to (A)$$

$$+ \mathbb{E}_{s\sim d_\rho^{\pi^*}, a\sim\pi^*(\cdot|s)}\left[(w_t^* - w_t)^T \nabla \log \pi_t(a|s)\right] \to (B)$$

$$(A): \mathbb{E}_{s\sim d_\rho^{\pi^*}, a\sim\pi^*(\cdot|s)}\left[Q_t(s,a) - \bar{Q}_t(s) - (w_t^*)^T(\phi_{s,a} - \bar{\phi}_s)\right]$$

$\mathbb{E}_{a\sim\pi_t(\cdot|s)}[Q_t(s,a)]$

$\mathbb{E}_{a\sim\pi_t(\cdot|s)}[\phi_{s,a}]$

$$= \mathbb{E}_{s\sim d_\rho^{\pi^*}, a\sim\pi^*(\cdot|s)}\left[Q_t(s,a) - (w_t^*)^T\phi_{s,a}\right]$$

$$- \mathbb{E}_{s\sim d_\rho^{\pi^*}, a\sim\pi_t(\cdot|s)}\left[Q_t(s,a) - (w_t^*)^T\phi_{s,a}\right]$$

$$\overset{C\text{-}S}{\le} \sqrt{\mathbb{E}_{s\sim d_\rho^{\pi^*}, a\sim\pi^*(\cdot|s)}\left[\left(Q_t(s,a) - (w_t^*)^T\phi_{s,a}\right)^2\right]}$$

$$+ \sqrt{\mathbb{E}_{s\sim d_\rho^{\pi^*}, a\sim\pi_t(\cdot|s)}\left[\left(Q_t(s,a) - (w_t^*)^T\phi_{s,a}\right)^2\right]},$$

π^* & π_t $\downarrow$ Unif_A

$$\le 2\sqrt{|A| \cdot \mathbb{E}_{\underbrace{s\sim d_\rho^{\pi^*}}_{d^*}, a\sim \mathrm{Unif}_A}\left[\left(Q_t(s,a) - (w_t^*)^T\phi_{s,a}\right)^2\right]}$$

$$= \sqrt{4|A|^2 \, L(w_t^*; \theta_t, d^*)} \to \varepsilon_{bias}$$

$$(B) = \mathbb{E}_{s\sim d_\rho^{\pi^*},\, a\sim\pi^*(\cdot|s)}\left[(w_t^*-w_t)^T \cdot \nabla \log \pi_t(a|s)\right]$$

$$\nabla \log \pi_t(a|s) \to \phi_{s,a} - \bar{\phi}_s$$

$$= \mathbb{E}_{s\sim d_\rho^{\pi^*},\, a\sim\pi^*(\cdot|s)}\left[(w_t^*-w_t)^T \phi_{s,a}\right]$$

$$- \mathbb{E}_{s\sim d_\rho^{\pi^*},\, a\sim\pi_t(\cdot|s)}\left[(w_t^*-w_t)^T \phi_{s,a}\right]$$

$$\le 2\sqrt{|A| \cdot \mathbb{E}_{s\sim d_\rho^{\pi^*},\, a\sim \mathrm{Unif}_A}\left[\left((w_t^*-w_t)^T\phi_{s,a}\right)^2\right]}$$

$$= 2\sqrt{|A| \cdot \mathbb{E}_{(s,a)\sim d^*}\left[(w_t^*-w_t)^T\phi_{s,a}\,\phi_{s,a}^T\,(w_t^*-w_t)\right]}$$

$$= (w_t^*-w_t)^T\, \mathbb{E}_{(s,a)\sim d^*}\left[\phi_{s,a}\phi_{s,a}^T\right] (w_t^*-w_t)$$

$$= (w_t^*-w_t)^T \Sigma_{d^*} (w_t^*-w_t)$$

$$= \|w_t^*-w_t\|^2_{\Sigma_{d^*}}$$

$$\kappa = \sup_w \frac{w^T\Sigma_{d^*}w}{w^T\Sigma_\nu w}$$

$$v^TAv = \|v\|_A^2$$

$$\Sigma_{d^*} \preceq \kappa \cdot \Sigma_\nu \preceq \frac{\kappa}{1-\gamma}\Sigma_{d_t} \qquad \left(d_t = d_\nu^{\pi_t} \ge (1-\gamma)\cdot\nu\right)$$

$$\Rightarrow (B) \le \sqrt{4|A| \cdot \|w_t^*-w_t\|^2_{\Sigma_{d^*}}} \le \sqrt{4|A| \cdot \frac{\kappa}{1-\gamma} \cdot \|w_t^*-w_t\|^2_{\Sigma_{d_t}}}$$

$$w_t^* \in \operatorname{argmin} L(w;\theta_t,d_t)$$

$\Rightarrow$ first-order optimality: $\forall w$: $(w-w_t^*)^T \nabla_w L(w_t^*;\theta_t,d_t) \ge 0$.

$$L(w;\theta_t,d_t) - L(w_t^*;\theta_t,d_t)$$

$$= \mathbb{E}_{(s,a)\sim d_t}\left[\left(w^T\phi_{s,a} - (w_t^*)^T\phi_{s,a} + (w_t^*)^T\phi_{s,a} - Q_t(s,a)\right)^2\right] - L(w_t^*;\theta_t,d_t)$$

$$= \mathbb{E}_{(s,a)\sim d_t}\left[\left(w^T\phi_{s,a} - (w_t^*)^T\phi_{s,a}\right)^2\right] + 2(w - w_t^*)^T \mathbb{E}_{(s,a)\sim d_t}\left[\phi_{s,a}\left((w_t^*)^T\phi_{s,a} - Q_t(s,a)\right)\right]$$

$$+ \cancel{\mathbb{E}_{(s,a)\sim d_t}\left[\left((w_t^*)^T\phi_{s,a} - Q_t(s,a)\right)^2\right]} - \cancel{L(w_t^*;\theta_t,d_t)}$$

$$= \|w - w_t^*\|^2_{\Sigma_{d_t}} + 2(w - w_t^*)^T \cdot \underbrace{\nabla_w L(w_t^*;\theta_t,d_t)}_{\geq 0}$$

$$\geq \|w - w_t^*\|^2_{\Sigma_{d_t}} \qquad \forall w \in \mathbb{R}^d$$

$$\Rightarrow (B) \leq \sqrt{4|A| \cdot \frac{K}{1-\gamma} \cdot \left(L(w_t;\theta_t,d_t) - L(w_t^*;\theta_t,d_t)\right)} \rightarrow \varepsilon_{stat.}$$

$\square$

Q-NPG Sample Complexity

Recall: $\theta_{t+1} = \theta_t + \eta \cdot w_t$, $w_t \approx \arg\min_w \mathbb{E}_{(s,a)\sim d_t}\left[\left(Q_t(s,a) - w^T \nabla \log \pi_\theta(a|s)\right)^2\right]$

Assumption (Episodic Sampling Oracle) For some $\nu \in \Delta_{S\times A}$:

$\exists$ oracle: 1) start $(s_0, a_0) \sim \nu$

2) continue to act thereafter in the MDP according to any π.

3) terminate this "rollout" when desire.

Algorithm: Sampler for $(s,a) \sim d_\nu^\pi$. unbiased estimate of $Q^\pi(s,a)$

1. sample $(s_0, a_0) \sim \nu$

2. sample d_ν^π:

$$d_\nu^\pi = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \cdot \Pr(s_t = \cdot, a_t = \cdot \mid \nu)$$

at time step h with probability γ, act according to π;

else, accept (s_h, a_h) as the sample.

3. get $\hat{Q}^{\pi}(s_h, a_h)$: from (s_h, a_h), continue to execute π.

and use a termination probability of $1-\gamma$.

Upon termination, set $\hat{Q}^{\pi}(s,a)$ as the undiscounted

sum of rewards from time h onwards.

- $$W_t \approx \arg\min_{w} \mathbb{E}_{(s,a)\sim d_t}\left[\left(Q_t(s,a) - w^T \nabla \log \pi_\theta(a|s)\right)^2\right] \quad A$$

① Linear Regression: $$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{Q}_t(s^{(i)}, a^{(i)}) - w^T \nabla \log \pi_\theta(a^{(i)}|s^{(i)})\right)^2.$$

$\rightarrow$ Concentration $\quad \varepsilon_{stat} \approx O(d/N)$

c.f. [Hsu et al. 2012] Random Design Analysis of Ridge Regression.

② SGD: $(s^{(i)}, a^{(i)}) \overset{i.i.d}{\sim} d_\mu^{\pi_t}$

for $n = 0, 1, \cdots, N-1$:

$$w^{(n+1)} = \mathrm{Proj}_{W}\left[w^{(n)} - 2\alpha \cdot \underbrace{\left(\phi^{(n)T} w^{(n)} - \hat{Q}^{(n)}\right) \cdot \phi^{(n)}}_{\nabla L(w^{(n)})}\right]$$

on $(s^{(n)}, a^{(n)})$

$$W_t = \frac{1}{N}\sum_{n=1}^{N} w^{(n)}$$

Lemma (SGD convergence) Assume $C = \{x : \|x\| \le B\}$, $B > 0$

Let f be a convex function. $x^* \in \arg\min_{x \in C} f(x)$.

Assume $\forall t$: $\|v_t\| \le \rho$, where v_t is an unbiased SG.

By $\eta = \sqrt{B^2 / \rho^2 N}$: $\mathbb{E}\left[f\left(\frac{1}{N} \sum_{n=1}^{N} x_t \right) \right] - f(x^*) \le \frac{B\rho}{\sqrt{N}}$

Corollary (Sample Complexity of Q-NPG)

$$\mathbb{E}\left[\min_{t<T} \{ V^*(\rho) - V^{(t)}(\rho) \} \right] \le \frac{BW}{1-\gamma} \sqrt{\frac{2 \log |A|}{T}} + \frac{\sqrt{4|A| \cdot \varepsilon_{bias}}}{1-\gamma}$$

$$+ \sqrt{\frac{8 \kappa |A| \cdot BW(BW+1)}{(1-\gamma)^4}} \frac{1}{N^{1/4}}.$$

each episode → expected length $\frac{2}{1-\gamma}$

each round

rounds

$\Rightarrow$ expected # of total samples : $\frac{2NT}{1-\gamma}$.

Proof: $\|\nabla L(w^{(n)})\| \le B \cdot \left(BW + \frac{1}{1-\gamma} \right)$.

By $\alpha = \sqrt{\frac{W^2}{G^2 N}}$ $\Rightarrow$ $\varepsilon_{stat} \le \frac{2BW\left(BW + \frac{1}{1-\gamma}\right)}{\sqrt{N}}$ $\square$

NPG : Performance Bounds for Smooth Policy Classes

$$L_A(w;\theta,\nu) := \mathbb{E}_{(s,a)\sim\nu}\left[\left(A^{\pi_\theta}(s,a) - w^T \nabla \log \pi_\theta(a|s)\right)^2\right]$$

(Approx) NPG : $\theta_{t+1} = \theta_t + \eta \cdot w_t$

where $w_t \approx \arg\min_w L_A(w;\theta_t, d_t)$

$w_t^* = \arg\min_w L_A(w;\theta_t, d_t)$

Assumptions

1. (smoothness). $\log \pi_\theta(a|s)$ is β-smooth wrt. θ.

   (log-linear: $\beta = B^2$)

2. (bias-variance decomposition)

   1) (excess risk): $\mathbb{E}\left[L_A(w_t;\theta_t,d_t) - L_A(w_t^*;\theta_t,d_t) \,\middle|\, \theta_t\right] \le \varepsilon_{stat}$

   Advantage

   $d_\nu^{\pi_t}$

   ε_{stat}: $1/\sqrt{N}$.

   2) (transfer error): $\mathbb{E}\left[L_A(w_t^*;\theta_t,d^*)\right] \le \varepsilon_{bias}$

   $d^*(s,a) = d_\rho^{\pi^*}(s) \times \pi^*(a|s)$ ~~Unif~~

3. (relative condition number) $\forall t \in [T]$

$$\mathbb{E}\left[\sup_w \frac{w^T \Sigma_{d^*}^{(t)} w}{w^T \Sigma_\nu^{(t)} w}\right] \le K < \infty$$

$$\Sigma_\nu^\theta = \mathbb{E}_{(s,a)\sim\nu}\left[\nabla \log \pi_\theta(a|s)\, \nabla \log \pi_\theta(a|s)^T\right]$$

Theorem (Agnostic Learning of NPG) With $\eta = \sqrt{2\log|A| / \beta W^2 T}$

$$\mathbb{E}\left[\min_{t<T}\{V^{\pi^*}(\rho)-V^{(t)}(\rho)\}\right] \le \frac{W}{1-\gamma}\cdot\sqrt{\frac{2\beta\cdot\log|A|}{T}} + \sqrt{\frac{\kappa\,\varepsilon_{stat}}{(1-\gamma)^3}} + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma}$$

Proof: (sketch)

$$\mathrm{err}_t = \mathbb{E}_{s\sim d_\rho^{\pi^*},\, a\sim\pi^*(\cdot|s)}\left[A_t(s,a) - W_t^T \nabla\log\pi_t(a|s)\right]$$

$$= \mathbb{E}_{(s,a)\sim d^*}\left[A_t(s,a) - (W_t^*)^T\nabla\log\pi_t(a|s)\right] \quad (A)$$

$$+\mathbb{E}_{(s,a)\sim d^*}\left[(W_t - W_t^*)^T\nabla\log\pi_t(a|s)\right] \quad (B)$$

$$(A) \overset{CS}{\le} \sqrt{L_A(W_t^*;\theta_t,d^*)} \le \sqrt{\varepsilon_{bias}}$$

$$(B)^2 \overset{CS}{\le} \|W_t - W_t^*\|^2_{\Sigma^t_{d^*}} \le \kappa\cdot\|W_t - W_t^*\|^2_{\Sigma^t_\nu}$$

$$\le \frac{\kappa}{1-\gamma}\|W_t - W_t^*\|^2_{\Sigma_{d_t}} \le \frac{\kappa}{1-\gamma}\cdot\varepsilon_{stat}. \quad \square$$

NPG Sample Complexity $\rightarrow$ Algorithm Design

$\rightarrow$ oracle $\begin{cases} (s_0,a_0)\sim\nu \\ \text{continue to act in MDP by } \pi \\ \text{terminate whenever you want.} \end{cases}$ $\rightarrow$ get sample $(s,a)\sim d_\nu^\pi$

Key challenge: how to estimate $A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$

$\hat{A}^{\pi_\theta}(s,a)$

$A^{\pi_\theta}(s,a)$ sampling : set Sample Q = True with probability $\frac{1}{2}$.

Sample Q = True : execute a_h at state s_h and then continue executing π with a termination probability $1-\gamma$.

Upon termination $\rightarrow$ $\hat{Q}^\pi$ as undiscounted sum of rewards from h.

Sample Q = False : sample $a_h' \sim \pi_\theta(\cdot|s_h)$. execute a_h' at state s_h and then continue executing π with a termination probability $1-\gamma$.

Upon termination $\rightarrow$ $\hat{V}^\pi$ as undiscounted sum of reward from h

$$\hat{A}^{\pi_\theta}(s_h, a_h) = 2(\hat{Q}^\pi - \hat{V}^\pi)$$

Run SGD on $L_A(w; \theta_t, d_t) = \mathbb{E}_{(s,a)\sim d_t}\left[\left(A_t(s,a) - w^T \nabla \log \pi_\theta(a|s)\right)^2\right]$.

$$w^{(n+1)} = \mathrm{Proj}_W\left[w^{(n)} - 2\alpha \cdot \left((w^{(n)})^T \nabla \log \pi_t(a^{(n)}|s^{(n)}) - \hat{A}_t(s^{(n)}, a^{(n)})\right) \cdot \nabla \log \pi_t(a^{(n)}|s^{(n)})\right]$$

Corollary (Sample Complexity of NPG) $\|\nabla_\theta \log \pi_t(a|s)\|_2 \le B$.

$\Rightarrow$ By optimally tuning η & α: $\mathbb{E}\left[\min_{t\le T}\{V^{\pi^*}(\rho) - V^{(t)}(\rho)\}\right]$

$$\le \frac{W}{1-\gamma}\sqrt{\frac{2\beta \log|A|}{T}} + \frac{\sqrt{\varepsilon_{bias}}}{1-\gamma} + \sqrt{\frac{8\kappa BW(BW+1)}{(1-\gamma)^4}} \cdot \frac{1}{N^{1/4}}$$

NPG $\Leftrightarrow$ CFA

$$\frac{L_A(w; \theta_t, d_t)}{L(w; \theta_t, d_t)}$$

Conservative Policy Iteration (CPI)

Kakade & Langford. Approximately Optimal Approximate Reinforcement Learning

Policy Optimization

↓

Policy Gradient : Actor Critic

π $Q/V(A)$

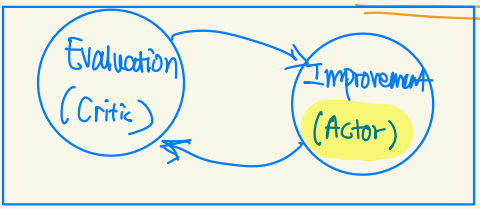
Policy Iteration :

① Policy Evaluation : Q^{π_k}

② Policy Improvement : $\pi_{k+1} = \Pi_Q \pi_k$

Key idea : policy optimization

→ supervised learning



Approximate Greedy Policy Selector

$$\hat{\pi} = G_\varepsilon(\pi, \Pi, \mu)$$

→ oracle

restart distribution

$(S, A, P, R, \gamma, \rho)$

s.t. $$\mathbb{E}_{s\sim d_\mu^\pi}\left[A^\pi(s, \hat{\pi})\right] \geq \max_{\tilde{\pi}\in\Pi} \mathbb{E}_{d_\mu^\pi}\left[A^\pi(s, \tilde{\pi})\right] - \varepsilon$$

"current" policy

"improvement"

Classification → Cost-sensitive.

$C(s,a) = -A^\pi(s,a)$

$$h: \min_{h\in\Pi} \mathbb{E}_{s\sim d_\mu^\pi, a\sim h(\cdot|s)}\left[C(s,a)\right]$$

Hypothesis class

Regression

loss function

$f_\pi \in F$: $\mathcal{L}(f_\pi) = \mathbb{E}_{s\sim d_\mu^\pi}\left[\max_{a\in A} |A^\pi(s,a) - f_\pi(s,a)|\right] \le \frac{\varepsilon}{2}$

$$\max_{\hat{\pi}\in\Pi} \mathbb{E}_{d_\mu^\pi}[A^\pi(s,\hat{\pi})] - \mathbb{E}_{s\sim d_\mu^\pi}[A^\pi(s,\pi)] \le \varepsilon$$

CPI Algorithm:

accuracy parameter ε

stepsize α

for $t = 1, 2, \cdots, T$:

$\pi' = G_\varepsilon(\pi_t, \Pi, \mu)$ → ε-approximate greedy policy

if $\mathbb{E}_{s\sim d_\mu^\pi}[A^{\pi_t}(s,\pi')] \le \varepsilon$

then return π_t.

☆ $\pi_{t+1} = (1-\alpha)\cdot\pi_t + \alpha\cdot\pi'$

Conservative

(let π_{t+1} be close to π_t)

Intuition: (improvement)

TRPO: KL

PPO: Clipping

PD Lemma

$$V^{\pi_{t+1}}(\mu) - V^{\pi_t}(\mu) = \frac{1}{1-\gamma}\cdot\mathbb{E}_{s\sim d_\mu^{\pi_{t+1}}}[A^{\pi_t}(s,\pi_{t+1})]$$

$\mathbb{E}_{a\sim\pi_{t+1}(\cdot|s)}[A^{\pi_t}(s,a)]$

$$= \frac{1}{1-\gamma}\left((1-\alpha)\cdot\mathbb{E}_{s\sim d_\mu^{\pi_{t+1}}}[\underbrace{A^{\pi_t}(s,\pi_t)}_{0}] + \alpha\cdot\mathbb{E}_{s\sim d_\mu^{\pi_{t+1}}}[A^{\pi_t}(s,\pi')]\right)$$

$$= \frac{\alpha}{1-\gamma}\cdot\mathbb{E}_{s\sim d_\mu^{\pi_{t+1}}}[A^{\pi_t}(s,\pi')]$$

Goal: find π' to maximize $\mathbb{E}_{s\sim d_\mu^{\pi_{t+1}}}[A^{\pi_t}(s,\pi')]$ >0

at round t

conservative

$\Rightarrow G_\varepsilon(\pi_t, \Pi, \mu)$

$\approx \mathbb{E}_{s\sim d_\mu^{\pi_t}}[A^{\pi_t}(s,\pi')]$

Lemma (Similar Policies imply similar state occupancy) $\forall t$:

1) $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 \le 2\alpha$. $\forall s$

2) $\|d_\mu^{\pi_{t+1}} - d_\mu^{\pi_t}\|_1 \le \frac{2\alpha\cdot\gamma}{1-\gamma}$.

TV distance

$TV(p\|q) = \frac{1}{2}\|p-q\|_1$

Proof: $\|\pi_{t+1}(\cdot|s) - \pi_t(\cdot|s)\|_1 = \alpha\cdot\|\pi_t(\cdot|s) - \pi'(\cdot|s)\|_1 \le 2\alpha$

$d_\mu^\pi(s) = (1-\gamma)\cdot\sum_{h=0}^{\infty}\gamma^h\, d_{\mu,h}^\pi(s)$

$\Pr(s_h = \cdot \mid s_0\sim\mu)$

$\|\pi(\cdot|s)\|_1 \le 1$

$\forall s'$: $d_{\mu,h}^{\pi_{t+1}}(s') - d_{\mu,h}^{\pi_t}(s')$

$$= \sum_{s,a}\Big(d_{\mu,h-1}^{\pi_{t+1}}(s)\cdot\pi_{t+1}(a|s) - d_{\mu,h-1}^{\pi_{t+1}}(s)\cdot\pi_t(a|s) + d_{\mu,h-1}^{\pi_{t+1}}(s)\cdot\pi_t(a|s) - d_{\mu,h-1}^{\pi_t}(s)\cdot\pi_t(a|s)\Big)\cdot P(s'|s,a)$$

$$= \sum_s d_{\mu,h-1}^{\pi_{t+1}}(s)\cdot\sum_a\big(\pi_{t+1}(a|s) - \pi_t(a|s)\big)\,P(s'|s,a)$$

$$+ \sum_s \left(d_{\mu,h-1}^{\pi_{t+1}}(s) - d_{\mu,h-1}^{\pi_t}(s) \right) \cdot \sum_a \pi_t(a|s) \cdot P(s'|s,a)$$

$$\Rightarrow \sum_{s'} \left| d_{\mu,h}^{\pi_{t+1}}(s') - d_{\mu,h}^{\pi_t}(s') \right|$$

$$\leq \underbrace{\sum_s d_{\mu,h-1}^{\pi_{t+1}}(s)}_{\leq 1} \cdot \underbrace{\sum_a \left| \pi_{t+1}(a|s) - \pi_t(a|s) \right|}_{\leq 2\alpha} \cdot \underbrace{\sum_{s'} P(s'|s,a)}_{=1}$$

$$+ \sum_{s'} \left| d_{\mu,h-1}^{\pi_{t+1}}(s) - d_{\mu,h-1}^{\pi_t}(s) \right| \cdot \underbrace{\sum_a \pi_t(a|s)}_{\leq 1} \cdot \underbrace{\sum_{s'} P(s'|s,a)}_{\leq 1}$$

$$\Rightarrow \| d_{\mu,h}^{\pi_{t+1}} - d_{\mu,h}^{\pi_t} \|_1 \leq 2\alpha + \| d_{\mu,h-1}^{\pi_{t+1}} - d_{\mu,h-1}^{\pi_t} \|_1,$$

$$\leq 4\alpha + \| d_{\mu,h-2}^{\pi_{t+1}} - d_{\mu,h-2}^{\pi_t} \|_1 \leq \cdots \leq 2\alpha \cdot h.$$

$$\Rightarrow \| d_{\mu}^{\pi_{t+1}} - d_{\mu}^{\pi_t} \|_1 \leq (1-\gamma) \cdot \sum_{h=0}^{\infty} \gamma^h \cdot \| d_{\mu,h}^{\pi_{t+1}} - d_{\mu,h}^{\pi_t} \|_1$$

$$\leq (1-\gamma) \cdot \sum_{h=0}^{\infty} \gamma^h \cdot 2\alpha \cdot h \leq 2\alpha \cdot \frac{\gamma}{1-\gamma}. \qquad \square$$

Theorem (Monotone Improvement)

$$\sum_{h=0}^{\infty} \gamma^h \cdot h = \frac{\gamma}{(1-\gamma)^2}.$$

Let $\boxed{A = \mathbb{E}_{s \sim d_\mu^{\pi_t}} \left[A^{\pi_t}(s, \pi') \right]}$ where $\pi' = G_\varepsilon(\pi_t, \pi, \mu)$

$$V^{\pi_{t+1}}(\mu) - V^{\pi_t}(\mu) \geq \frac{\alpha}{1-\gamma} \cdot \left(A - \frac{2\alpha\gamma}{(1-\gamma)^2} \right) \qquad \Delta$$

Set $\alpha = \dfrac{A(1-\gamma)^2}{4\gamma}$. $\Rightarrow$ $V^{\pi_{t+1}}(\mu) - V^{\pi_t}(\mu) \geq \dfrac{A^2(1-\gamma)}{8\gamma}$. $\star$

Proof: PD Lemma: $V^{\pi_{t+1}}(\mu) - V^{\pi_t}(\mu) = \frac{\alpha}{1-\gamma} \cdot \mathbb{E}_{s\sim d_\mu^{\pi_{t+1}}}[A^{\pi_t}(s,\pi')]$.

$$= \frac{\alpha}{1-\gamma} \cdot A - \frac{\alpha}{1-\gamma} \cdot \left(\mathbb{E}_{s\sim d_\mu^{\pi_t}}[A^{\pi_t}(s,\pi')] - \mathbb{E}_{s\sim d_\mu^{\pi_{t+1}}}[A^{\pi_t}(s,\pi')] \right)$$

$$\left| \mathbb{E}_{x\sim p_1}[f(x)] - \mathbb{E}_{x\sim p_2}[f(x)] \right|$$

$$= \left| \sum_x (p_1(x) - p_2(x)) f(x) \right|$$

$$\overset{\text{Hölder}}{\leq} \|p_1 - p_2\|_1 \cdot \|f\|_\infty$$

$\hookrightarrow \sup_x |f(x)|$

$$\leq \|d_\mu^{\pi_{t+1}} - d_\mu^{\pi_t}\|_1 \cdot \|A^{\pi_t}(s,\pi')\|_\infty$$

$$\leq \frac{2\alpha\cdot\gamma}{1-\gamma} \cdot \frac{1}{1-\gamma}$$

$$= \frac{2\alpha\gamma}{(1-\gamma)^2}$$

$\left(R \in [0,1] \quad V \in [0, \frac{1}{1-\gamma}] \quad A \in [0, \frac{1}{1-\gamma}] \right)$

$\square$

Convergence of CPI

CPI will terminate

Theorem (Local Convergence of CPI) Within $T \leq \frac{8\gamma}{\varepsilon^2}$ rounds, the output policy π_T satisfies: $\max_{\pi\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi_T}}[A^{\pi_T}(s,\pi)] \leq 2\varepsilon$

Proof: If CPI does NOT terminate at round t,

$$A = \mathbb{E}_{s\sim d_\mu^{\pi_t}}[A^{\pi_t}(s,\pi')] \geq \varepsilon$$

$$\Rightarrow V^{\pi_{t+1}}(\mu) - V^{\pi_t}(\mu) \geq \frac{\varepsilon^2(1-\gamma)}{8\gamma}$$

$V \in [0, \frac{1}{1-\gamma}] \Rightarrow$ CPI will terminate within $\frac{8\gamma}{\varepsilon^2} = T$ rounds.

$$\mathbb{E}_{s\sim d_\mu^{\pi_t}}\left[A^{\pi_t}(s,\pi')\right] \geq \max_{\tilde{\pi}\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi_t}}\left[A^{\pi_t}(s,\tilde{\pi})\right] - \varepsilon.$$

$$\Rightarrow \max_{\tilde{\pi}\in\Pi} \mathbb{E}_{s\sim d_\mu^{\pi_T}}\left[A^{\pi_T}(s,\tilde{\pi})\right] \leq \varepsilon + \varepsilon = 2\varepsilon. \qquad \square$$

local optimum $\Rightarrow$ global optimum. ($V^\pi \to V^*$)

initial distribution ρ

Theorem (Global Convergence of CPI) Upon termination,

we have a policy π s.t.

bias

$$\mathbb{E}_{s\sim d_\mu^\pi}\left[\max_{a\in A} A^\pi(s,a)\right] - \max_{\tilde{\pi}\in\Pi} \mathbb{E}_{s\sim d_\mu^\pi}\left[A^\pi(s,\tilde{\pi})\right]$$

$$V^*(\rho) - V^\pi(\rho) \leq \frac{2\varepsilon + \varepsilon_\pi}{(1-\gamma)^2} \cdot \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty$$

initial state distribution

distribution mismatch

Proof: By PD Lemma, $V^*(s) - V^\pi(s)$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi^*}}\left[A^\pi(s,\pi^*)\right]$$

$$\leq \frac{1}{1-\gamma} \cdot \mathbb{E}_{s\sim d_\rho^{\pi^*}}\left[\max_{a\in A} A^\pi(s,a)\right] \qquad \geq 0$$

$\forall \pi$: $d_\mu^\pi \geq (1-\gamma)\cdot\mu$

$$\leq \frac{1}{1-\gamma} \cdot \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \cdot \mathbb{E}_{s\sim\mu}\left[\max_{a\in A} A^\pi(s,a)\right]$$

$$\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \cdot \mathbb{E}_{s\sim d_\mu^\pi}\left[\max_{a\in A} A^\pi(s,a)\right]$$

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

$$\leq \varepsilon_\pi + 2\varepsilon \qquad \square$$

CPI encourages to have policies that can achieve optimal action.

Trust Region Methods

[Schulman et al. 2015] Trust Region Policy Optimization — TRPO

[Schulman et al. 2017] Proximal Policy Optimization Algorithms — PPO

CPI: conservative

$\pi_{t+1} = (1-\alpha)\cdot\pi_t + \alpha\cdot\pi'$

step-size

$$\boxed{\begin{aligned} &\max_{\theta} \ \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[A^{\pi_{\theta_t}}(s, \pi_\theta) \right] \rightarrow 1 \\ &\text{s.t. } KL\left(Pr_\mu^{\pi_{\theta_t}} \,\|\, Pr_\mu^{\pi_\theta} \right) \le \delta \rightarrow 2. \end{aligned}}$$

Local Approximation (in neighborhood of θ_t)

$Pr_\mu^{\pi}(\tau)$: trajectory distribution.

$$\max_{\theta} \ \left\langle \underbrace{\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}},\, a\sim\pi_{\theta_t}(\cdot|s)}\left[\nabla \log \pi_{\theta_t}(a|s)\cdot A^{\pi_{\theta_t}}(s,a) \right]}_{\nabla_\theta V^{\pi_{\theta_t}}(\mu)} , \theta \right\rangle$$

$$\text{s.t. } \left\langle \nabla_\theta KL\left(Pr_\mu^{\pi_{\theta_t}} \,\|\, Pr_\mu^{\pi_\theta} \right)\Big|_{\theta=\theta_t} , \theta-\theta_t \right\rangle + \frac{1}{2}(\theta-\theta_t)^T \left(\nabla_\theta^2 KL\left(Pr_\mu^{\pi_{\theta_t}} \,\|\, Pr_\mu^{\pi_\theta} \right)\Big|_{\theta=\theta_t} \right)\cdot(\theta-\theta_t)$$

Claim:

① $\nabla_\theta KL\left(Pr_\mu^{\pi_{\theta_t}} \,\|\, Pr_\mu^{\pi_\theta} \right)\Big|_{\theta=\theta_t} = 0$

② $\nabla_\theta^2 KL\left(Pr_\mu^{\pi_{\theta_t}} \,\|\, Pr_\mu^{\pi_\theta} \right)\Big|_{\theta=\theta_t} = H\cdot F(\theta_t)$

horizon (finite-horizon episodic MDP)

Fisher Information Matrix

Proof: $\tau = (s_0, a_0, s_1, a_1, \cdots, s_{H-1}, a_{H-1}, s_H)$

$$Pr_\mu^{\pi}(\tau) = \mu(s_0)\cdot\prod_{h=0}^{H-1} \pi(a_h|s_h)\cdot P(s_{h+1}|s_h, a_h)$$

$$F(\theta_t) = \mathbb{E}_{(s,a)\sim d^{\pi_{\theta_t}}}\left[\nabla\log\pi_\theta(a|s)\cdot\nabla\log\pi_\theta(a|s)^T \right]$$

$$KL\left(Pr_\mu^{\pi_{\theta_t}} \| Pr_\mu^{\pi_\theta}\right) = \sum_\tau Pr_\mu^{\pi_{\theta_t}}(\tau) \cdot \log \frac{Pr_\mu^{\pi_{\theta_t}}(\tau)}{Pr_\mu^{\pi_\theta}(\tau)}$$

$$= \sum_\tau Pr_\mu^{\pi_{\theta_t}}(\tau) \cdot \left(\sum_{h=0}^{H-1} \log \frac{\pi_{\theta_t}(a_h|s_h)}{\pi_\theta(a_h|s_h)} \right)$$

$$= \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)\sim d_{\mu,h}^{\pi_{\theta_t}}} \left[\log \frac{\pi_{\theta_t}(a_h|s_h)}{\pi_\theta(a_h|s_h)} \right],$$

① $$\nabla_\theta KL\left(Pr_\mu^{\pi_{\theta_t}} \| Pr_\mu^{\pi_\theta}\right)\Big|_{\theta=\theta_t} = \sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)} \left[-\nabla \log \pi_\theta(a_h|s_h) \Big|_{\theta=\theta_t} \right].$$

$$= \sum_{h=0}^{H-1} \mathbb{E}_{s_h,\, a_h\sim\pi_{\theta_t}(\cdot|s_h)} \left[-\nabla \log \pi_\theta(a_h|s_h) \right]\Big|_{\theta=\theta_t}$$

$$= -\sum_{h=0}^{H-1} \mathbb{E}_{s_h} \left[\sum_{a_h} \pi_{\theta_t}(a_h|s_h) \cdot \nabla \log \pi_{\theta_t}(a_h|s_h) \right] = 0$$

② $$\nabla_\theta^2 KL\left(Pr_\mu^{\pi_{\theta_t}} \| Pr_\mu^{\pi_\theta}\right)\Big|_{\theta=\theta_t} = -\sum_{h=0}^{H-1} \mathbb{E}_{(s_h,a_h)} \left[\nabla^2 \log \pi_\theta(a_h|s_h) \right]\Big|_{\theta=\theta_t}$$

$$= -\sum_{h=0}^{H-1} \mathbb{E}_{s_h,\, a_h} \left[\underbrace{\frac{\nabla^2 \pi_\theta(a_h|s_h)}{\pi_\theta(a_h|s_h)^2}}_{0} - \frac{\nabla \pi_\theta(a_h|s_h) \cdot \nabla \pi_\theta(a_h|s_h)^T}{\pi_\theta(a_h|s_h)^2} \right]\Bigg|_{\theta=\theta_t}$$

$$= \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h} \left[\nabla \log \pi_\theta(a_h|s_h) \left(\nabla \log \pi_\theta(a_h|s_h)\right)^T \right]\Big|_{\theta=\theta_t}$$

$$= H \cdot F(\theta_t) \qquad \square.$$

Let $\Delta = \theta - \theta_t$. TRPO update:

$$\boxed{\begin{aligned} &\max_\Delta \ \langle \Delta, \nabla V^{\pi_{\theta_t}}(\mu) \rangle \\ &\text{s.t. } \Delta^T \cdot F_t \Delta \le \delta \end{aligned}}$$

where $F_t = H \cdot F(\theta_t)$

QP (Quadratic Program)

$\Rightarrow$ TRPO update :

self-normalized step-size

$$\theta_{t+1} = \theta_t + \sqrt{\frac{\delta}{\left(\nabla V^{\pi_{\theta_t}}(\mu)\right)^T F_t^{-1} \left(\nabla V^{\pi_{\theta_t}}(\mu)\right)}} \cdot F_t^{-1} \nabla V^{\pi_{\theta_t}}(\mu)$$

NPG : $\theta_{t+1} = \theta_t + \eta \cdot F(\theta_t)^{-1} \nabla V^{\pi_{\theta_t}}(\mu)$

CPI + KL- trust region $\rightarrow$ TRPO.

CPI + ℓ_∞- trust region $\rightarrow$ PPO

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

$$\text{s.t.} \ \sup_s \| \pi^\theta(\cdot|s) - \pi^{\theta_t}(\cdot|s) \|_{TV} \le \delta.$$

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

importance weight

Clipping

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

Conservative

lower bound of the objective

$$\text{Clip}(x, \varepsilon) = \begin{cases} 1-\varepsilon, & \text{if } x < 1-\varepsilon \\ 1+\varepsilon, & \text{if } x > 1-\varepsilon \\ x, & \text{otherwise} \end{cases}$$

$$\downarrow$$

If $\dfrac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)} \notin [1-\varepsilon, 1+\varepsilon]$.

$$\Rightarrow \nabla_\theta \, \mathrm{Clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}, \varepsilon\right) \cdot A^{\pi_{\theta_t}}(s,a) = 0.$$

PPO at round t:

1) collect a dataset $(s,a) \sim d_\mu^{\pi_{\theta_t}}$.

2) run (mini-batch) SGA on $f_t(\theta)$

Note that PPO directly optimize (lower bound of)

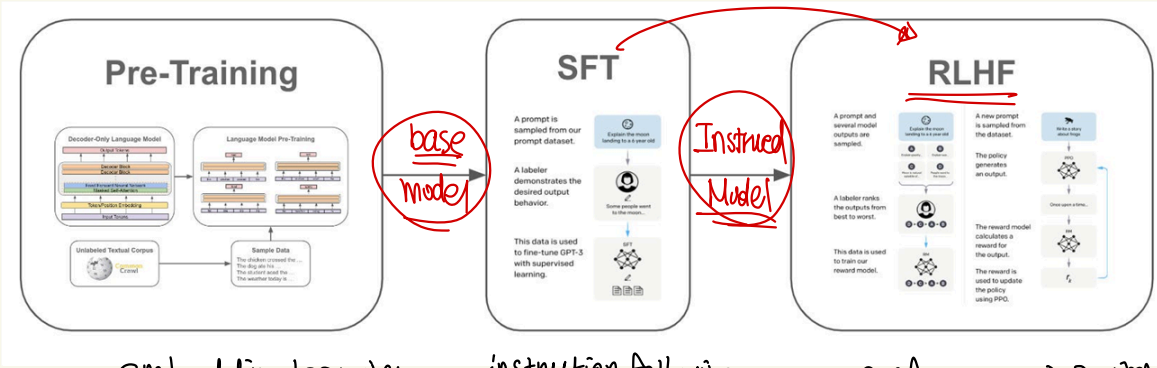
$$\mathbb{E}_{s \sim d^{\pi_{\theta_t}},\, a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_{\theta_t}}(s,a) \right]$$

approximate version

TRPO/NPG/CPI: $\max_\theta \left\langle \theta - \theta_t, \mathbb{E}_{s,a \sim d_\mu^{\pi_{\theta_t}}} \left[\nabla_\theta \log \pi_{\theta_t}(a|s) \cdot A^{\pi_{\theta_t}}(s,a) \right] \right\rangle$.

From PPO to GRPO

Recap: 3 training stages of LLM.



embedding learning instruction following preference optimization

RLHF (Preference Optimization) in LLM :

Input : a preference pair (x, a, a', y) → label $y \in \{0, 1\}$

prompt, 1st response, 2nd response

$$\begin{cases} y=1 : a' \succ a \\ y=0 : a \succ a' \end{cases}$$

↓

a fine-tuned LLM

Goal : align LLM with human's preference

Method :
1) ☆ RL-based (need a reward model, using RL algorithm to optimize) ☆
2) Direct-Update (e.g. DPO)

① Get a reward model
② Use RL to fine-tune LLM

① Get a reward model

def (Bradley-Terry Model) prompt x. two actions $a, a' \in A$.

$$\Pr(y=1 \mid x, a, a') = \frac{\exp(r(x,a))}{\exp(r(x,a)) + \exp(r(x,a'))}.$$

r_ϕ is a latent reward function

$$\sigma(w) = \frac{1}{1+e^{-w}}.$$

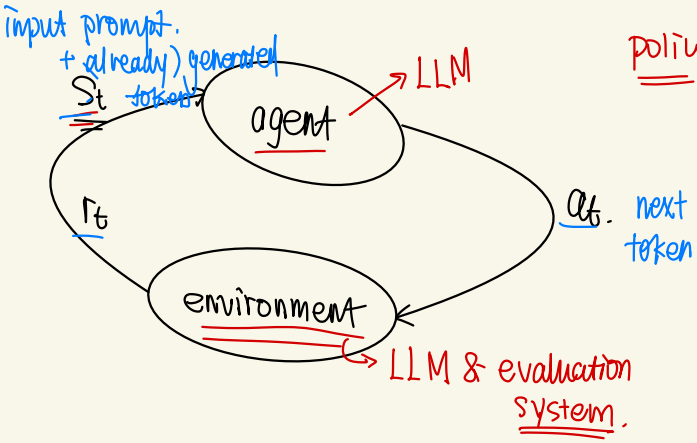
MLE :

$$\underset{r_\phi}{\operatorname{argmin}} \mathcal{L}_R(r_\phi, D) = -\mathbb{E}_{(x, a_w, a_l) \sim D} \left[\log \sigma(r_\phi(x, a_w) - r_\phi(x, a_l)) \right],$$

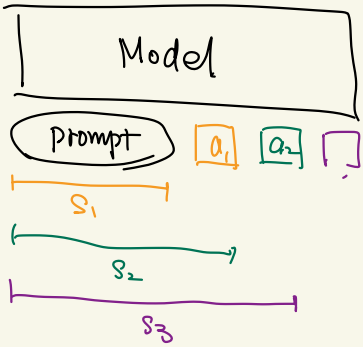
r_ϕ ← DNN / LLM

winning losing

② Use RL to fine-tune LLM.



Policy: the probability of generating next token at current state.



Recall PPO (Proximal Policy optimization):

$$J(\theta) = \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}},\, a\sim\pi_{\theta_t}(\cdot|s)}\left[\min\left\{\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}\cdot A^{\pi_{\theta_t}}(s,a),\ \mathrm{Clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)};1-\varepsilon,1+\varepsilon\right)\cdot A^{\pi_{\theta_t}}(s,a)\right\}\right]$$

GAE (Generalized Advantage Estimation)

$$A_t^{GAE} = \delta_t + (\gamma\cdot\lambda)\cdot\delta_{t+1} + \dots + (\gamma\cdot\lambda)^{T-t+1}\cdot\delta_{T-1}$$

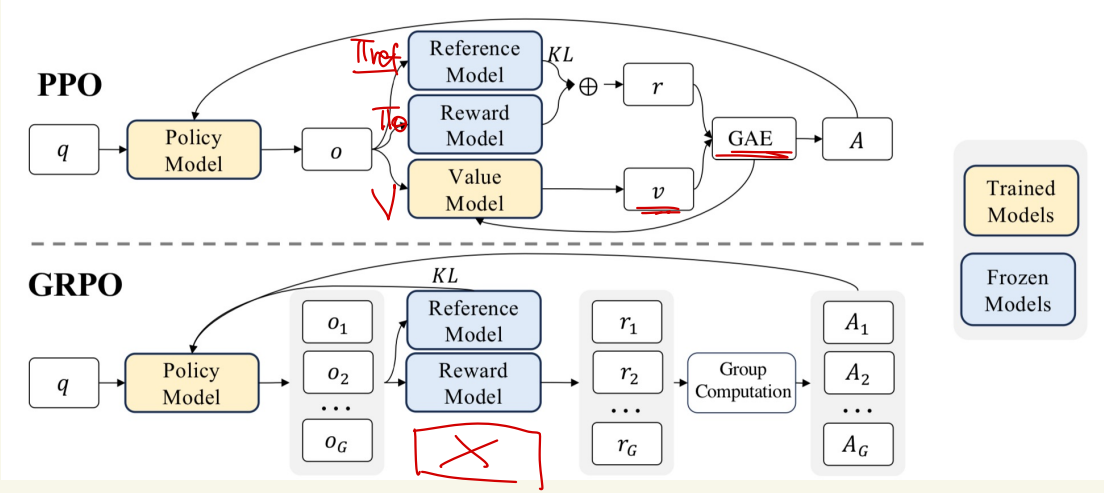
PPO in LLM:

$$\delta_t = R_t + \gamma\cdot V(s_{t+1}) - V(s_t) \quad \text{TD error}$$

$$\max_{\pi_\theta} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\min\left\{\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}\cdot A_t^{GAE}(s,a),\ \mathrm{Clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)};1-\varepsilon,1+\varepsilon\right)\cdot A_t^{GAE}(s,a)\right\}\right]$$

$$-\beta\cdot KL\left(\pi_\theta(\cdot|s)\,\|\,\pi_{ref}(\cdot|s)\right)$$

original SFT model.



GRPO: critic-free

"clipped version"

$$\max_{\pi_\theta} \mathbb{E}_{o \sim D(\pi_\theta; s)} \left[A^{Group}_{o_1, \cdots, o_G}(o) \right] - \beta \cdot KL(\pi_\theta \| \pi_{ref})$$

actor-model

$$A^{Group}_{o_1, \cdots, o_G}(o) = \frac{r_i - mean(r)}{std(r)}$$

reward model

RM

PRM: process RM (PPO)

ORM: outcome RM

DeepseekMath: Pushing the Limits of Mathematical Reasoning in Open LMs.

① ORM encourage LLM to think longer → longer CoT

② Even pure RL-based LLM (NO SFT) can behave well.