

## Algorithmic and Theoretical Foundations of RL

---

More on Policy Optimization

---

# Table of Contents

---

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Deep Deterministic Policy Gradient (DDPG)

Entropy Regularization

# Gradient Method for Optimization over Distributions

It is clear that policy optimization for RL is a special case of optimization over probability distributions:

$$\max_{\theta} J(\theta) = \mathbb{E}_{X \sim P_{\theta}} [f(X)].$$

The gradient ascent method for this problem is given by

$$\theta \leftarrow \theta + \underbrace{\alpha \cdot \nabla J(\theta)}_{\Delta \theta},$$

which can be interpreted as searching over the  $\ell_2$ -ball of the parameter space:

$$\Delta \theta = \operatorname{argmax}_{\|d\|_2 \leq \alpha} \{J(\theta) + \langle \nabla J(\theta), d \rangle\}.$$

**Question:** Is it more natural to search over the probability distribution space since  $J(\theta)$  essentially relies on  $P_{\theta}$ ? **YES** -> [Natural gradient method](#).

# Natural Gradient Method Optimization over Distributions

Natural gradient method conduct search based on KL divergence between probability distributions:

$$\begin{aligned}\Delta\theta &= \operatorname{argmax}_{\text{KL}(P_\theta \| P_{\theta+d}) \leq \alpha} \{J(\theta) + \langle \nabla J(\theta), d \rangle\} \\ &\approx \alpha \cdot F(\theta)^{-1} \nabla J(\theta),\end{aligned}$$

where  $F(\theta)$  is the Fisher information matrix at  $\theta$ , defined by

$$F(\theta) = \mathbb{E}_{X \sim P_\theta} \left[ \nabla_\theta \log p_\theta(X) (\nabla_\theta \log p_\theta(X))^T \right].$$

This yields the natural gradient method

$$\theta \leftarrow \theta + \alpha \cdot F(\theta)^{-1} \nabla J(\theta),$$

which can also be viewed as approximate Newton's method where  $F(\theta)$  acts as a precondition matrix.

---

For simplicity, we assume  $F(\theta)$  is invertible. If it is not the case we may consider using  $F(\theta) + \varepsilon \cdot I$  or using  $F(\theta)^\dagger \nabla J(\theta)$  which is minimal  $\ell_2$ -norm solution to  $\min_x \|F(\theta)x - \nabla J(\theta)\|_2$ . The natural gradient direction can be found by solving  $F(\theta)d = \nabla J(\theta)$  via CG method.

# Derivation of Natural Gradient Direction

First recall that given two probability distributions  $P$  and  $Q$  with pdf  $p(x)$  and  $q(x)$  respectively, the KL divergence is defined by

$$\text{KL}(P\|Q) = \mathbb{E}_P \left[ \log \frac{dP}{dQ} \right] = \mathbb{E}_P \left[ \log \frac{p(X)}{q(X)} \right].$$

It follows that

$$\begin{aligned} \text{KL}(P_\theta \| P_{\theta+d}) &= \mathbb{E}_{P_\theta} \left[ \log \frac{p_\theta(X)}{p_{\theta+d}(X)} \right] \\ &= -\mathbb{E}_{P_\theta} [\log p_{\theta+d}(X) - \log p_\theta(X)] \\ &\approx -d^T \underbrace{\mathbb{E}_{P_\theta} \left[ \frac{\nabla_\theta p_\theta(X)}{p_\theta(X)} \right]}_{l_1 = \mathbb{E}_{P_\theta} [\nabla_\theta \log p_\theta(X)]} - \frac{1}{2} d^T \underbrace{\mathbb{E}_{P_\theta} \left[ \frac{\nabla_\theta^2 p_\theta(X)}{p_\theta(X)} - \frac{\nabla_\theta p_\theta(X) (\nabla_\theta p_\theta(X))^T}{p_\theta(X)^2} \right]}_{l_2 = \mathbb{E}_{P_\theta} [\nabla_\theta^2 \log p_\theta(X)]} d. \end{aligned}$$

## Derivation of Natural Gradient Direction (Cont'd)

For  $l_1$ , there holds

$$\mathbb{E}_{p_\theta} \left[ \frac{\nabla_\theta p_\theta(X)}{p_\theta(X)} \right] = \int \nabla_\theta p_\theta(X) dx = 0.$$

For  $l_2$ , there holds

$$\mathbb{E}_{p_\theta} \left[ \frac{\nabla_\theta^2 p_\theta(X)}{p_\theta(X)} \right] = \int \nabla_\theta^2 p_\theta(X) dx = 0$$

and

$$\mathbb{E}_{p_\theta} \left[ \frac{\nabla_\theta p_\theta(X) (\nabla_\theta p_\theta(X))^T}{p_\theta(X)^2} \right] = \mathbb{E}_{p_\theta} \left[ \nabla_\theta \log p_\theta(X) (\nabla_\theta \log p_\theta(X))^T \right] = F(\theta).$$

It follows that

$$\begin{aligned} \Delta\theta &= \operatorname{argmax}_{\text{KL}(p_\theta \| p_{\theta+d}) \leq \alpha} \{J(\theta) + \langle \nabla J(\theta), d \rangle\} \\ &\approx \operatorname{argmax}_{d^T F(\theta) d \leq \alpha} \{J(\theta) + \langle \nabla J(\theta), d \rangle\} \\ &= \alpha \cdot F(\theta)^{-1} \nabla J(\theta). \end{aligned}$$

# Natural Policy Gradient (NPG)

Natural policy gradient is the natural gradient method applied to the policy optimization problem:

$$J(\theta) = \mathbb{E}_{s_0 \sim \mu} [V_{\pi_\theta}(s_0)] = \mathbb{E}_{\tau \sim P_\mu^{\pi_\theta}} [r(\tau)],$$

where given  $\tau = (s_t, a_t, r_t)_{t=0}^\infty$ ,

$$P_\mu^{\pi_\theta}(\tau) = \mu(s_0) \prod_{t=0}^{\infty} \pi_\theta(a_t | s_t) P(s_{t+1} | s_t, a_t) \quad \text{and} \quad r(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t.$$

The natural gradient search direction can be incorporated into different policy based methods (including REINFORCE, actor-critic) after statistical estimation of  $F(\theta)$  (e.g., using data from an episode). We only focus on expression for  $F(\theta)$ .

By the definition of  $F(\theta)$  and expression for  $P_\mu^{\pi_\theta}$ , we have

$$\begin{aligned} F(\theta) &= \mathbb{E}_{\tau \sim P_\mu^{\pi_\theta}} \left[ \left( \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t | s_t) \right) \left( \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t | s_t) \right)^\top \right] \\ &= \mathbb{E}_{\tau \sim P_\mu^{\pi_\theta}} \left[ \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t | s_t) (\nabla_\theta \log \pi_\theta(a_t | s_t))^\top \right]. \end{aligned}$$

## Two Common Expressions of $F(\theta)$ to Avoid Divergence

- Average case:

$$\begin{aligned} F(\theta) &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[ \sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) (\nabla_{\theta} \log \pi_{\theta}(a_t | s_t))^T \right] \\ &= \mathbb{E}_{s \sim d^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot | s)} \left[ \nabla_{\theta} \log \pi_{\theta}(a | s) (\nabla_{\theta} \log \pi_{\theta}(a | s))^T \right], \end{aligned}$$

where  $d^{\pi_{\theta}}(s) = \mathbb{E}_{s_0 \sim \mu} [\lim_{t \rightarrow \infty} P(s_t = s | s_0, \pi_{\theta})]$  is state stationary distribution.

- Discounted case:

$$\begin{aligned} F(\theta) &= \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[ \sum_{t=0}^{+\infty} \gamma^t \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) (\nabla_{\theta} \log \pi_{\theta}(a_t | s_t))^T \right] \\ &= \mathbb{E}_{T \sim \text{Geo}(1-\gamma)} \left[ \mathbb{E}_{\tau \sim p(\cdot | \theta)} \left[ \sum_{t=0}^{T-1} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) (\nabla_{\theta} \log \pi_{\theta}(a_t | s_t))^T \mid T \right] \right] \\ &= \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot | s)} \left[ \nabla_{\theta} \log \pi_{\theta}(a | s) (\nabla_{\theta} \log \pi_{\theta}(a | s))^T \right], \end{aligned}$$

where  $d_{\mu}^{\pi_{\theta}}(s) = \mathbb{E}_{s_0 \sim \mu} [(1 - \gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s | s_0, \pi_{\theta})]$  is state visitation measure, and  $T$  obeys the geometric distribution with parameter  $1 - \gamma$ .



# Table of Contents

---

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Deep Deterministic Policy Gradient (DDPG)

Entropy Regularization

## Performance Difference Lemma (PDL)

Given two policies  $\pi$  and  $\pi'$ , recall from Lecture 2 that

$$v_{\pi'} - v_{\pi} = (I - \gamma P^{\pi'})^{-1} (\mathcal{T}_{\pi'} v_{\pi} - v_{\pi}).$$

Thus,

$$\begin{aligned}\mathbb{E}_{s \sim \mu} [v_{\pi'}(s) - v_{\pi}(s)] &= (\mathcal{T}_{\pi'} v_{\pi} - v_{\pi})^T (I - \gamma (P^{\pi'})^T)^{-1} \mu \\ &= \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi'}} [\mathcal{T}_{\pi'} v_{\pi}(s) - v_{\pi}(s)] \\ &= \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi'}} \mathbb{E}_{a \sim \pi'(\cdot|s)} [A_{\pi}(s, a)],\end{aligned}$$

where  $A_{\pi}(s, a) = q_{\pi}(s, a) - v_{\pi}(s)$ , where second equality follows from a lemma from Lecture 7.

**Lemma 1.** For two policies  $\pi$  and  $\pi'$ , their difference in terms of state values is

$$\mathbb{E}_{s \sim \mu} [v_{\pi'}(s) - v_{\pi}(s)] = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi'}} \mathbb{E}_{a \sim \pi'(\cdot|s)} [A_{\pi}(s, a)].$$

# Trust Region Policy Optimization (TRPO)

## Overall Idea

Based on PDL, given a policy  $\pi_{\theta_{\text{old}}}$ , we can rewrite  $J(\theta)$  as

$$\max_{\theta} J(\theta) = J(\theta_{\text{old}}) + \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}} \mathbb{E}_{a \sim \pi_{\theta}(\cdot|s)} \left[ A_{\pi_{\theta_{\text{old}}}}(s, a) \right].$$

Since we do not have access to  $d_{\mu}^{\pi_{\theta}}$ , instead maximize the approximation:

$$\max_{\theta} L_{\theta_{\text{old}}}(\theta) = J(\theta_{\text{old}}) + \underbrace{\frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta_{\text{old}}}}} \mathbb{E}_{a \sim \pi_{\theta}(\cdot|s)} \left[ A_{\pi_{\theta_{\text{old}}}}(s, a) \right]}_{\text{surrogate advantage function}}.$$

►  $J(\theta)$  and  $L_{\theta_{\text{old}}}(\theta)$  matches at  $\theta_{\text{old}}$  up to first derivative.

# Trust Region Policy Optimization (TRPO)

## Overall Idea

It can be shown that  $L_{\theta_{\text{old}}}(\theta)$  can be used to provide a lower bound for  $J(\theta)$ :

$$J(\theta) \geq L_{\theta_{\text{old}}}(\theta) - C_{\epsilon} \cdot \text{KL}_{\text{max}}(\theta_{\text{old}} \parallel \theta),$$

where  $\epsilon = \max_s |\mathbb{E}_{a \sim \pi_{\theta}} [A_{\pi_{\theta_{\text{old}}}}(s, a)]|$ ,  $\text{KL}_{\text{max}}(\theta_{\text{old}} \parallel \theta) = \max_s \text{KL}(\pi_{\theta_{\text{old}}}(\cdot|s) \parallel \pi_{\theta}(\cdot|s))$ .

Given an estimator  $\theta_t$ , this inequality suggests that we may update it by solving

$$\max_{\theta} L_{\theta_t}(\theta) \quad \text{subject to} \quad \text{KL}_{\text{max}}(\theta_t \parallel \theta) \leq \delta.$$

In practice, replace  $\text{KL}_{\text{max}}(\theta_t \parallel \theta)$  by the average version and instead solve

$$\max_{\theta} L_{\theta_t}(\theta) \quad \text{subject to} \quad \overline{\text{KL}}(\theta_t \parallel \theta) \leq \delta,$$

where  $\overline{\text{KL}}(\theta_t \parallel \theta) = \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta_t}}} [\text{KL}(\pi_{\theta_t}(\cdot|s) \parallel \pi_{\theta}(\cdot|s))]$ .

# Trust Region Policy Optimization (TRPO)

## TRPO is Approximately NPG Plus Line Search

After linear approximation to  $L_{\theta_t}(\theta)$  and quadratic approximation to KL at  $\theta_t$ ,

$$L_{\theta_t}(\theta) \approx \nabla_{\theta} J(\theta_t)^T (\theta - \theta_t), \quad \overline{\text{KL}}(\theta_t \| \theta) \approx \frac{1}{2} (\theta - \theta_t)^T F(\theta_t) (\theta - \theta_t),$$

we arrive at the same problem as that for NPG,

$$\max_{\theta} \nabla_{\theta} J(\theta_t)^T (\theta - \theta_t) \quad \text{subject to} \quad \frac{1}{2} (\theta - \theta_t)^T F(\theta_t) (\theta - \theta_t) \leq \delta.$$

- TRPO is NPG with adaptive line search in implementations.

# Table of Contents

---

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Deep Deterministic Policy Gradient (DDPG)

Entropy Regularization

# Proximal Policy Optimization (PPO)

Recall from last section that (after omitting constant term  $J(\theta_t)$ )

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

which is surrogate function of true target in small region around  $\theta_t$  in terms of KL.

PPO keeps new policy close to old one by adopting two schemes:

- ▶ Adaptive KL penalty
- ▶ **Clipped objective**

## PPO with Clipped Objective

PPO-Clip neither has a KL-divergence term in the objective nor has a constraint. Instead, it relies on specialized clipping of the objective function to remove incentives for the new policy to get far from the old one. Let  $r(\theta) = \frac{\pi_{\theta}(a|s)}{\pi_{\theta_t}(a|s)}$ . Then  $r(\theta_t) = 1$ . The clipped objective function is given by

$$L_{\theta_t}^{\text{CLIP}}(\theta) = \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta_t}}} \mathbb{E}_{a \sim \pi_{\theta_t}(\cdot|s)} \left[ \min \left( r(\theta) A^{\pi_{\theta_t}}, \text{clip}(r(\theta), 1 - \epsilon, 1 + \epsilon) A^{\pi_{\theta_t}} \right) \right],$$

where

$$\text{clip}(r(\theta), 1 - \epsilon, 1 + \epsilon) = \begin{cases} 1 + \epsilon, & r(\theta) > 1 + \epsilon, \\ r(\theta), & r(\theta) \in [1 - \epsilon, 1 + \epsilon], \\ 1 - \epsilon, & r(\theta) < 1 - \epsilon. \end{cases}$$

- ▶ The  $\min$  operation ensure  $L_{\theta_t}^{\text{CLIP}}(\theta)$  provides a lower bound. Since a maximal point will be computed subsequently,  $\min$  will not cancel the effect of  $\text{clip}$ .
- ▶ PPO policy update (in expectation):  $\theta_{t+1} = \underset{\theta}{\operatorname{argmax}} L_{\theta_t}^{\text{CLIP}}(\theta)$ .



# Table of Contents

---

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Deep Deterministic Policy Gradient (DDPG)

Entropy Regularization

# Deterministic Policy and Parameterization

Consider the case where  $\mathcal{S}$  and  $\mathcal{A}$  are continuous state and action spaces, respectively. We use  $\pi$  to denote a deterministic policy:  $a = \pi(s)$  is an action.

- State value and action value:

$$v_{\pi}(s) = q_{\pi}(s, \pi(s)) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s \right],$$

where  $r_t = r(s_t, \pi(s_t), s_{t+1})$  and  $s_{t+1} \sim P(\cdot | s_t, \pi(s_t))$ .

- Given a parameterized policy  $\pi_{\theta}$ , we have

$$\begin{aligned} J(\theta) &= \int_{\mathcal{S}} \mu(s_0) v_{\pi_{\theta}}(s_0) ds_0 \\ &= \frac{1}{1-\gamma} \int_{\mathcal{S}} d_{\mu}^{\pi_{\theta}}(s) \int_{\mathcal{S}} p(s' | s, \pi_{\theta}(s)) r(s, \pi_{\theta}(s), s') ds' ds \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}} \mathbb{E}_{s' \sim P(\cdot | s, \pi_{\theta}(s))} [r(s, \pi_{\theta}(s), s')], \end{aligned}$$

where  $d_{\mu}^{\pi_{\theta}}(s) = \mathbb{E}_{s_0 \sim \mu} [(1-\gamma) \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t P(s_t = s | s_0, \pi_{\theta}) ds]$  is state visitation measure and  $\mu$  is initial state distribution.

# Deterministic Policy Gradient Theorem

---

**Theorem 1.** Suppose that  $\nabla_{\theta}\pi_{\theta}(s)$  and  $\nabla_a q_{\pi_{\theta}}(s, a)$  exist. Then,

$$\nabla_{\theta}J(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}(s)} \left[ \nabla_{\theta}\pi_{\theta}(s) \nabla_a q_{\pi_{\theta}}(s, a)|_{a=\pi_{\theta}(s)} \right].$$

# Proof of Theorem 1

It suffices to consider  $\nabla_{\theta} v_{\pi_{\theta}}(s_0)$ , which can be rewritten as

$$v_{\pi_{\theta}}(s_0) = \sum_{t=0}^{\infty} \int \gamma^t r(s_t, \pi_{\theta}(s_t), s_{t+1}) \prod_{k=0}^{\infty} P(s_{k+1}|s_k, \pi_{\theta}(s_k)) d\tau.$$

Thus,

$$\begin{aligned} \nabla_{\theta} v_{\pi_{\theta}}(s_0) &= \sum_{t=0}^{\infty} \int \gamma^t \nabla_a r(s_t, \pi_{\theta}(s_t), s_{t+1}) \nabla_{\theta} \pi_{\theta}(s_t) \prod_{k=0}^{\infty} P(s_{k+1}|s_k, \pi_{\theta}(s_k)) d\tau \\ &+ \sum_{t=0}^{\infty} \int \gamma^t r(s_t, \pi_{\theta}(s_t), s_{t+1}) \left( \sum_{k=0}^{\infty} \nabla_a \log P(s_{k+1}|s_k, \pi_{\theta}(s_k)) \nabla_{\theta} \pi_{\theta}(s_k) \right) \prod_{k=0}^{\infty} P(s_{k+1}|s_k, \pi_{\theta}(s_k)) d\tau \\ &= \sum_{t=0}^{\infty} \mathbb{E}_{\tau} \left[ \gamma^t \nabla_a r(s_t, \pi_{\theta}(s_t), s_{t+1}) \nabla_{\theta} \pi_{\theta}(s_t) \right] \\ &+ \sum_{t=0}^{\infty} \mathbb{E}_{\tau} \left[ \nabla_a \log P(s_{t+1}|s_t, \pi_{\theta}(s_t)) \nabla_{\theta} \pi_{\theta}(s_t) \sum_{k=0}^{t-1} \gamma^k r(s_k, \pi_{\theta}(s_k), s_{k+1}) \right] \\ &+ \sum_{t=0}^{\infty} \mathbb{E}_{\tau} \left[ \nabla_a \log P(s_{t+1}|s_t, \pi_{\theta}(s_t)) \nabla_{\theta} \pi_{\theta}(s_t) \sum_{k=t}^{\infty} \gamma^k r(s_k, \pi_{\theta}(s_k), s_{k+1}) \right]. \end{aligned}$$

## Proof of Theorem 1 (Cont'd)

Note that the second term is equal to 0 since

$$\mathbb{E}_{s_{t+1} \sim P(\cdot | s_t, \pi_\theta(s_t))} [\nabla_\theta \log P(s_{t+1} | s_t, \pi_\theta(s_t))] = 0.$$

Moreover,

$$\begin{aligned} & \mathbb{E}_\tau \left[ \nabla_a \log P(s_{t+1} | s_t, \pi_\theta(s_t)) \nabla_\theta \pi_\theta(s_t) \sum_{k=t}^{\infty} \gamma^k r(s_k, \pi_\theta(s_k), s_{k+1}) \right] \\ &= \mathbb{E}_\tau \left[ \gamma^t \nabla_a \log P(s_{t+1} | s_t, \pi_\theta(s_t)) \nabla_\theta \pi_\theta(s_t) (r(s_t, \pi_\theta(s_t), s_{t+1}) + \gamma v_{\pi_\theta}(s_{t+1})) \right] \end{aligned}$$

Further note that

$$\begin{aligned} \mathbb{E}_{s_{t+1}} [\nabla_a q_{\pi_\theta}(s_t, a)] &= \nabla_a \int (r(s_t, a, s) + \gamma v_{\pi_\theta}(s)) P(s | s_t, a) ds \\ &= \int \nabla_a r(s_t, a, s) P(s | s_t, a) ds + \int \nabla_a \log P(s | s_t, a) (r(s_t, a, s) + \gamma v_{\pi_\theta}(s)) P(s | s_t, a) ds \\ &= \mathbb{E}_{s_{t+1}} [\nabla_a r(s_t, a, s_{t+1}) + \nabla_a \log P(s_{t+1} | s_t, a) (r(s_t, a, s_{t+1}) + \gamma v_{\pi_\theta}(s_{t+1}))]. \end{aligned}$$

## Proof of Theorem 1 (Cont'd)

---

Submitting the last two results into the expressions for  $\nabla_{\theta} v_{\pi_{\theta}}(s_0)$  yields that

$$\begin{aligned}\nabla_{\theta} v_{\pi_{\theta}}(s_0) &= \sum_{t=0}^{\infty} \gamma^t \nabla_{\theta} \pi_{\theta}(s_t) \nabla_a q_{\pi_{\theta}}(s_t, a)|_{a=\pi_{\theta}(s_t)} \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_{\theta}}(s)} [\nabla_{\theta} \pi_{\theta}(s) \nabla_a q_{\pi_{\theta}}(s, a)|_{a=\pi_{\theta}(s)}] .\end{aligned}$$

Averaging over all  $s_0$  completes the proof of Theorem 1.

# Deep Deterministic Policy Gradient (DDPG)

- ▶ DDPG is a policy gradient method which concurrently a deterministic policy  $\pi_\theta$  and an action value function  $q_\omega(s, a) \approx q_{\pi_\theta}(s, a)$ . It is an actor-critic algorithm.
- ▶ Policy of DDPG is deterministic, need to add random noisy when collecting data; experience replay buffer is also used to break statistical dependence.
- ▶ Update of  $\omega$  for action value function is overall the same to Fitted Q-learning.
  - Note that the TD target for updating  $\omega$  would be

$$r(s, a, s') + \gamma q_{\pi_\theta}(s', \pi_\theta(s')) \approx r(s, a, s') + \gamma q_\omega(s', \pi_\theta(s')),$$

which also relies on  $\theta$ . Thus, two target networks for both  $\omega$  and  $\theta$  are used for stable training (Since we use  $q_\omega(s, a)$  to approximate action values of different policies,  $\theta$  should not vary too much during training).

# Table of Contents

---

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Deep Deterministic Policy Gradient (DDPG)

Entropy Regularization



# Motivation

## *Enhance exploration by adding entropy regularization*

First consider a general policy  $\pi$ . Recalling definition of visitation measure  $d_{\mu}^{\pi}$ , entropy regularized objective function is define by

$$\begin{aligned} J(\pi) &= \mathbb{E}_{s \sim d_{\mu}^{\pi}} \left( \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(\cdot|s,a)} [r(s, a, s')] + \tau H(\pi(\cdot|s)) \right) \\ &= \mathbb{E}_{s \sim d_{\mu}^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(\cdot|s,a)} [r(s, a, s') - \tau \log \pi(a|s)] , \end{aligned}$$

where  $H(\pi(\cdot|s))$  denotes the entropy of the probability distribution  $\pi(\cdot|s)$ :

$$H(\pi(\cdot|s)) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ \log \frac{1}{\pi(a|s)} \right] .$$

We can rewrite  $J(\pi)$  in terms of state values based on a regularized reward

$$J(\pi) = \mathbb{E}_{s \sim \mu} [v_{\pi}^{\tau}(s)] ,$$

where  $v_{\pi}^{\tau}(s) = \mathbb{E}_{\pi} [\sum_{t=0}^{\infty} \gamma^t r_{\tau}(s_t, a_t, s_{t+1}) | s_0 = s]$  with

$$r_{\tau}(s, a, s') = r(s, a, s') - \tau \log \pi(a|s).$$

# Soft Bellman Equation

- Soft state value  $v_\pi^\tau$  :

$$v_\pi^\tau(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_\tau(s_t, a_t, s_{t+1}) | s_0 = s \right]$$

- Soft action value  $q_\pi^\tau(s, a)$ : [ $a_0$  is chosen, thus entropy equal to 0]

$$q_\pi^\tau(s, a) = \mathbb{E}_\pi \left[ r(s_0, a_0, s_1) + \sum_{t=1}^{\infty} \gamma^t r_\tau(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a \right]$$

- Relation between  $q_\pi^\tau$  and  $v_\pi^\tau$ :

$$v_\pi^\tau(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} [-\tau \log \pi(a|s) + q_\pi^\tau(s, a)]$$

- Soft Bellman equation:

$$\begin{aligned} v_\pi^\tau(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(\cdot|s, a)} [r_\tau(s, a, s') + \gamma v_\pi^\tau(s')] \\ q_\pi^\tau(s, a) &= \mathbb{E}_{s' \sim P(\cdot|s, a)} [r(s, a, s') + \gamma v_\pi^\tau(s')] \\ &= \mathbb{E}_{s' \sim P(\cdot|s, a)} [r(s, a, s') + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s')} [q_\pi^\tau(s', a') - \tau \log \pi(a'|s')]] \end{aligned}$$

# Soft Bellman Operator

- For state value, soft Bellman operator  $\mathcal{T}_\pi^\tau$  under a policy  $\pi$  is defined by

$$[\mathcal{T}_\pi^\tau v](s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(\cdot|s,a)} [r_\tau(s, a, s') + \gamma v(s')].$$

It can be shown that  $\mathcal{T}_\pi^\tau$  is  $\gamma$ -contraction with respect to  $\ell_\infty$ -norm. Thus,  $v_\pi^\tau$  is a unique fixed point of  $\mathcal{T}_\pi^\tau$ .

- For action value, soft Bellman operator  $\mathcal{F}_\pi^\tau$  under a policy  $\pi$  is given by

$$[\mathcal{F}_\pi^\tau q](s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} [r(s, a, s') + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s)} [q(s', a') - \tau \log \pi(a'|s')]] ,$$

It can also be shown that  $\mathcal{F}_\pi^\tau$  is  $\gamma$ -contraction with respect to  $\ell_\infty$ -norm. Thus,  $q_\pi^\tau$  is a unique fixed point of  $\mathcal{F}_\pi^\tau$ .

# Soft Bellman Optimal Equation

For any  $v \in \mathbb{R}^{|S|}$ , the soft Bellman optimality operator  $\mathcal{T}^\tau$  is defined by

$$\begin{aligned} [\mathcal{T}^\tau v](s) &= \max_{\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(\cdot|s,a)} [r_\tau(s, a, s') + \gamma v(s')] \\ &= \max_{\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ \underbrace{\mathbb{E}_{s' \sim P(\cdot|s,a)} [r(s, a, s') + \gamma v(s')]}_{:=q(s,a)} - \tau \log \pi(a|s) \right] \\ &= \tau \log \left( \|\exp(q(s, \cdot) / \tau)\|_1 \right), \end{aligned}$$

where the maximum value is attained iff  $\pi(\cdot|s) \propto \exp(q(s, \cdot) / \tau)$ .

- $\mathcal{T}^\tau$  is  $\gamma$ -contraction with respect to  $\ell_\infty$ -norm.
- It is worth noting that

$$\max_{\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} [q(s, a) - \tau \log \pi(a|s)] = \min_{\pi} \text{KL} \left( \pi(\cdot|s) \parallel \frac{\exp(q(s, a) / \tau)}{Z(s)} \right),$$

where  $Z(s) = \|\exp(q(s, \cdot) / \tau)\|_1$  is the normalization factor.

## Soft Bellman Optimal Equation

For  $q \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ , the soft Bellman optimality operator  $\mathcal{F}^\tau$  is defined by

$$\begin{aligned} [\mathcal{F}^\tau q](s, a) &= \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ r(s, a, s') + \gamma \max_{\pi} \mathbb{E}_{a' \sim \pi(\cdot|s')} [q(s', a') - \tau \log \pi(a'|s')] \right] \\ &= \mathbb{E}_{s' \sim P(\cdot|s, a)} \left[ r(s, a, s') + \gamma \left[ \tau \log \left( \|\exp(q(s', \cdot)/\tau)\|_1 \right) \right] \right], \end{aligned}$$

where the maximum value is attained iff  $\pi(\cdot|s) \propto \exp(q(s, \cdot)/\tau)$ .

►  $\mathcal{F}^\tau$  is  $\gamma$ -contraction with respect to  $\ell_\infty$ -norm.

**Theorem 2.** Let  $v_*^\tau$  and  $q_*^\tau$  be the fixed points of  $\mathcal{T}^\tau$  and  $\mathcal{F}^\tau$ , respectively. It is easy to see that  $q_*^\tau = \mathbb{E}_{s' \sim P(\cdot|s,a)}[r(s, a, s') + \gamma v_*^\tau(s')]$ . Then,

$$v_*^\tau(s) = \max_{\pi} v_{\pi}^\tau(s), \quad \forall s \in \mathcal{S},$$

and the equality is achieved by the optimal policy given by

$$\pi_*^\tau(a|s) = \frac{\exp(q_*^\tau(s, a)/\tau)}{\|\exp(q_*^\tau(s, \cdot)/\tau)\|_1}.$$

# Soft Policy Iteration

- Soft policy evaluation:

$$q_{\pi_k}^\tau = \mathcal{F}_\pi^\tau q_{\pi_k}^\tau$$

- Soft policy improvement:

$$\pi_{k+1} = \arg \min_{\pi' \in \Pi} \text{KL} \left( \pi'(\cdot|s) \parallel \frac{\exp(q_{\pi_k}^\tau(s, \cdot)/\tau)}{Z_{\pi_k}(s)} \right),$$

where  $Z_{\pi_k}(s)$  is the normalization factor.

**Theorem 3 (Convergence of Soft Policy Iteration).** Repeated application of soft policy evaluation and soft policy improvement from any  $\pi \in \Pi$  converges to a policy  $\pi_*^\tau$  such that  $q_{\pi_*^\tau}^\tau(s, a) \geq q_\pi^\tau(s, a)$  for all  $\pi \in \Pi$  and  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .

---

See “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor” by Haarnoja et al. 2018 for details.

# Soft Actor Critic (SAC)

SAC is a policy based or actor-critic method for solving

$$\max_{\theta} J(\theta) = \mathbb{E}_{s \sim \mu} [v_{\pi_{\theta}}^{\tau}(s)] .$$

In addition to typical ways for updating value function and policy parameters,

- ▶ Reparametrization trick is used in the computation of policy gradient;
- ▶ Both state values and action values have been parametrized for stable training.

---

See “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor” by Haarnoja et al. 2018 for details.



Questions?