Algorithmic and Theoretical Foundations of RL

Policy Optimization II

Ke Wei
School of Data Science
Fudan University

Table of Contents

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Entropy Regularization

Deterministic Policy Gradient (DPG)

Gradient Method 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

$$\theta^{+} = \theta + \eta \cdot \nabla J(\theta),$$

where the search direction $\Delta\theta = \nabla J(\theta)$ satisfies

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

Question: Is it more natural to search over probability distribution space since $J(\theta)$ essentially relies on P_θ? YES –> Natural gradient method.

Natural Gradient over Distributions

Natural gradient method conducts search based on KL divergence between probability distributions ($F(\theta)^\dagger$ is pseudoinverse of $F(\theta)$):

$$\begin{aligned}\Delta\theta &\propto \underset{\mathrm{KL}\left(P_\theta \| P_{\theta+d}\right)\le\alpha}{\operatorname{argmax}} \{J(\theta) + \langle \nabla J(\theta), d\rangle\} \\ &\approx F(\theta)^\dagger \nabla J(\theta),\end{aligned}$$

where $F(\theta)$ is the Fisher information matrix at θ, 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 leads to natural gradient method:

$$\theta^+ = \theta + \eta \cdot F(\theta)^\dagger \nabla J(\theta),$$

which can also be viewed as preconditioned gradient method.

Derivation of Natural Gradient Direction

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

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

It follows that

$$\begin{aligned}
\mathrm{KL}(P_\theta\|P_{\theta+d}) &= \mathbb{E}_{P_\theta}\left[\log \frac{p_\theta(X)}{p_{\theta+d}(X)}\right] \\
&= -\mathbb{E}_{P_\theta}\left[\log p_{\theta+d}(X) - \log p_\theta(X)\right] \\
&\approx -d^T \underbrace{\mathbb{E}_{P_\theta}\left[\frac{\nabla_\theta p_\theta(X)}{p_\theta(X)}\right]}_{I_1 = \mathbb{E}_{P_\theta}\left[\nabla_\theta \log p_\theta(X)\right]} - \frac{1}{2} d^T \underbrace{\mathbb{E}_{P_\theta}\left[\frac{\nabla_\theta^2 p_\theta(X)}{p_\theta(X)} - \frac{\nabla_\theta p_\theta(X)(\nabla_\theta p_\theta(X))^T}{p_\theta(X)^2}\right]}_{I_2 = \mathbb{E}_{P_\theta}\left[\nabla_\theta^2 \log p_\theta(X)\right]} d.
\end{aligned}$$

Derivation of Natural Gradient Direction

For I_1, one has

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

For I_2, one has

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

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

$$\Delta\theta = \underset{\mathrm{KL}\left(P_\theta \| P_{\theta+d}\right) \leq \alpha}{\operatorname{argmax}} \{J(\theta) + \langle \nabla J(\theta), d\rangle\} \approx \underset{d^T F(\theta) d \leq 2\alpha}{\operatorname{argmax}} \{J(\theta) + \langle \nabla J(\theta), d\rangle\} \propto F(\theta)^\dagger \nabla J(\theta).$$

The pseudoinverse basically means that we won't consider the direction such $F(\theta)d = 0$ since in this case one has $\mathrm{KL}\left(P_\theta \| P_{\theta+d}\right) \approx d^T F(\theta) d = 0$ and the objective function roughly remains unchanged.

Natural Policy Gradient (NPG)

Natural policy gradient is natural gradient applied to RL optimization problem:

$$\max_{\theta} V^{\pi_\theta}(\mu) = \mathbb{E}_{s_0 \sim \mu}\left[V^{\pi_\theta}(s_0)\right] = \mathbb{E}_{\tau \sim P_\mu^{\pi_\theta}}\left[r(\tau)\right],$$

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

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

Natural gradient search direction can be incorporated into different policy optimization methods (including REINFORCE, actor-critic) after MC evaluation 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 $P_\mu^{\pi_\theta}$ (assuming $\pi_\theta(a|s) = 1$ for any θ),

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

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

- Average case:

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

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

- Discounted case:

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

  where $d_{\mu}^{\pi_\theta}(s) = \mathbb{E}_{s_0\sim\mu}\left[(1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(s_t = s|s_0, \pi_\theta)\right]$ is discounted state visitation measure.

Remark

- For the discounted case, it is not difficult to verify that the natural gradient direction $F(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\mu)$ satisfies

$$F(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\omega^*,$$

where ω^* is the (ℓ_2-minimal) solution to

$$\min_\omega L(\omega) = \mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\left[\left((\nabla_\theta \log \pi_\theta(a|s))^T \omega - A^{\pi_\theta}(s,a)\right)^2\right].$$

See "On the theory of policy gradient methods: Optimality, approximation, and distribution shift" by Agarwal et al. 2021 for details.

Remark

- For the softmax parameterization (i.e., $\pi_\theta(a|s) = \exp(\theta_{s,a})/(\sum_{a'} \exp(\theta_{s,a'}))$), it can be verified all the solutions to $\min_\omega L(\omega)$ has the following general form:

$$\omega^*_{s,a} = A^{\pi_\theta}(s,a) + c_s,$$

where c_s is a constant relying on s. Thus NPG in policy space is given by

$$\pi_{\theta^+}(a|s) = \frac{\pi_\theta(a|s) \cdot \exp\left(\frac{\eta}{1-\gamma} A^{\pi_\theta}(s,a)\right)}{\sum_{a'} \pi_\theta(a'|s) \cdot \exp\left(\frac{\eta}{1-\gamma} A^{\pi_\theta}(s,a')\right)},$$

which coincides with EQA in Lecture 7 (a policy mirror ascent method).

See "On the theory of policy gradient methods: Optimality, approximation, and distribution shift" by Agarwal et al. 2021 for details.

Table of Contents

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Entropy Regularization

Deterministic Policy Gradient (DPG)

Trust Region Policy Optimization (TRPO)

Overall Idea

Given a policy π_{θ_k}, by performance difference lemma, we can rewrite $V^{\pi_\theta}(\mu)$ as

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

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

$$\max_\theta \; V_k(\theta) = V^{\pi_{\theta_k}}(\mu) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_k}}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_{\theta_k}}(s,a)\right].$$

Trust Region Policy Optimization (TRPO)

Two Facts

- It is easy to see that $V^{\pi_\theta}(\mu)$ and $V_k(\theta)$ match at θ_t up to first derivative.
- It can be shown that

$$V^{\pi_\theta}(\mu) \geq V_k(\theta) - \frac{2\gamma\varepsilon_t}{(1-\gamma)^2} \max_s \mathrm{KL}(\pi_{\theta_k}(\cdot|s)\|\pi_\theta(\cdot|s)),$$

where $\varepsilon_t = \max_{s,a} |A^{\pi_{\theta_k}}(s,a)|$.

See "Trust region policy optimization" by Schulman et al. 2017 for derivation of second fact.

Trust Region Policy Optimization (TRPO)

TRPO is Approximately NPG Plus Line Search

The second fact suggests that we may seek a new estimator by maximizing $V_k(\theta)$ in a small neighborhood of θ_t:

$$\max_{\theta} \ V_k(\theta) \quad \text{subject to} \quad \max_{s} \mathrm{KL}(\pi_{\theta_k}(\cdot|s) \| \pi_\theta(\cdot|s)) \le \delta.$$

Moreover, replace constraint by the average version and instead solve

$$\max_{\theta} \ V_k(\theta) \quad \text{subject to} \quad \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_k}}} \left[\mathrm{KL}(\pi_{\theta_k}(\cdot|s) \| \pi_\theta(\cdot|s))\right] \le \delta.$$

Trust Region Policy Optimization (TRPO)

TRPO is Approximately NPG Plus Line Search

After linear approximation to $V_k(\theta)$ and quadratic approximation to KL at θ_k,

$$V_k(\theta) \approx (\nabla_\theta V^{\pi_{\theta_k}}(\mu))^T(\theta - \theta_k),$$

$$\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_k}}}\left[\mathrm{KL}(\pi_{\theta_k}(\cdot|s)\|\pi_\theta(\cdot|s))\right] \approx \frac{1}{2}(\theta - \theta_k)^T F(\theta_k)(\theta - \theta_t),$$

we arrive at the same problem as that for NPG,

$$\max_\theta (\nabla_\theta V^{\pi_{\theta_k}}(\mu))^T(\theta - \theta_k) \quad \text{subject to} \quad \frac{1}{2}(\theta - \theta_k)^T F(\theta_k)(\theta - \theta_k) \le \delta.$$

- 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)

Entropy Regularization

Deterministic Policy Gradient (DPG)

Proximal Policy Optimization (PPO)

Recall from last section that

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

serves as a surrogate function of true target in small region around θ_k.

PPO keeps new policy close to old one through clipped objective (i.e., based on quotient instead of a metric).

PPO with Clipped Objective

Let $r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}$. Then $r(\theta_t) = 1$. The clipped objective function is given by

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

where

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

- The $\min$ operation ensure $V_k^{\text{clip}}(\theta)$ provides a lower bound. Since a maximal point will be computed subsequently, $\min$ will not cancel the effect of clip.
- PPO policy update (in expectation): $\theta_{t+1} = \operatorname{argmax}_\theta V_k^{\text{clip}}(\theta)$.
- In flat region, gradient of $V_k^{\text{clip}}(\theta)$ is zero, thus won't move far from θ_t is using policy gradient type method to solve the sub-problem.

See "Proximal policy optimization algorithms" by Schulman et al. 2017 for details.

Prototype PPO-Clip

Algorithm 1: PPO-Clip

```
Initialization: policy parameters $\theta_0$, value function parameter $\omega_0$.
for $k = 0, 1, \cdots$ do
    Collect trajectories following $\pi_{\theta_k}$:
        $\tau_n = \{s_t, a_t, r_t, s_{t+1}\}_{t \geq 0}, \quad n = 1, \cdots, N$
    Construct the GAE($\lambda$) estimation of the advantage function at each $(s_t, a_t)$ and the
      corresponding value function at each $s_t$:
        $\delta_t(\omega_k) = r_t + \gamma V(s_{t+1}; \omega_k) - V(s_t; \omega_k)$
        $A^\lambda(s_t, a_t) = \sum_{\ell \geq 0} (\gamma\lambda)^\ell \delta_{t+\ell}(\omega_k), \quad G^\lambda(s_t) = A^\lambda(s_t, a_t) + V(s_t; \omega_k)$
    Update the policy parameter by maximizing the PPO-clip objective:
        $\theta_{k+1} = \operatorname*{argmax}_\theta \sum_{\tau_n} \sum_{t \geq 0} \min\left\{ \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^\lambda(s_t, a_t), \mathrm{clip}\left( \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) A^\lambda(s_t, a_t) \right\}$
    Fit value function by regression on mean-squared error
        $\omega_{k+1} = \operatorname*{argmin}_\omega \sum_{\tau_n} \sum_{t \geq 0} \left( V(s_t; \omega) - G^\lambda(s_t) \right)^2$
end
```

Remark

- In PPO-Clip, we use Generalized Advantage Estimation (GAE) to estimate the advantage which is based on the λ-return of the value function, see Lecture 5.
- It seems there is a mismatch between θ_{k+1} and ω_{k+1} and there are variants of PPO-Clip objective which include entropy regularization and a proximal term.
- In practice, the advantage estimation is often normalized.

Table of Contents

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Entropy Regularization

Deterministic Policy Gradient (DPG)

Entropy Regularized State Value

Given a policy π, the average entropy regularized state value is given by

$$\begin{aligned} V_\tau^\pi(\mu) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^\pi}\left\{\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s')\right] + \tau H(\pi(\cdot|s))\right\} \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^\pi}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') - \tau\log\pi(a|s)\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t,s_{t+1}) - \tau\log\pi(a_t|s_t)\right)\ \Big|\ s_0\sim\mu,\pi\right], \end{aligned}$$

where $H(p) = -\sum_a p_a \log p_a$ is the entropy of a probability distribution.

- Entropy regularized state value at s, denoted $V_\tau^\pi(s)$, can be similarly defined.
- In addition to the perspective based on entropy regularization for more exploration, it can also be interpreted as encouraging exploration via revising the reward (the third equation).

In this section, we will use τ to denote the regularization parameter, which should be distinguished from the trajectory.

Bellman Equation and Operator

It is clear that $V_\tau^\pi(\mu)$ satisfies the following Bellman equation

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

Define the Bellman operator as follows

$$\mathcal{T}_\tau^\pi V(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') - \tau\log\pi(a|s) + \gamma V(s')\right].$$

It is easy to see that $\mathcal{T}_\tau^\pi$ is of γ-contraction and V_τ^π is a fixed point of $\mathcal{T}_\tau^\pi$.

Entropy Regularized Action Value

The entropy regularized action value is defined as

$$Q_\tau^\pi(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V_\tau^\pi(s')\right].$$

Note that we choose not to include $-\tau \log \pi(a|s)$ here. One immediately has

$$V_\tau^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_\tau^\pi(s,a) - \tau \log \pi(a|s)\right].$$

- Action value is state value where initial policy is deterministic, thus entropy 0.
- It is convenient to give the maximum improvement policy (similar to PI policy). That is, the solution to

  $$\max_\pi \mathcal{T}_\tau^\pi V(s) = \max_\pi \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') - \tau \log \pi(a|s) + \gamma V(s')\right]$$

  is $\pi(\cdot|s) \propto \exp(Q^V(s,\cdot)/\tau)$, where $Q^V(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V(s')\right]$. Entropy regularization moves the maxima to the interior so that it has an explicit solution in terms of softmax representation.

Performance Difference Lemma

Define the advantage function

$$A_\tau^\pi(s,a) = Q_\tau^\pi(s,a) - \tau \log \pi(a|s) - V_\tau^\pi(s).$$

It is evident that $\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A_\tau^\pi(s,a)\right] = 0$.

Lemma 1
One has

$$\mathcal{T}_\tau^{\pi_1} V_\tau^{\pi_2}(s) - V_\tau^{\pi_2}(s) = \mathbb{E}_{a\sim\pi_1(\cdot|s)}\left[A_\tau^{\pi_2}(s,a)\right] - \tau \mathrm{KL}(\pi_1(\cdot|s)\|\pi_2(\cdot|s)).$$

Lemma 2 (Performance Difference Lemma)
There holds

$$V_\tau^{\pi_1}(\mu) - V_\tau^{\pi_2}(\mu) = \frac{1}{1-\gamma}\sum_s d_\mu^{\pi_1}(s)\left(\mathcal{T}_\tau^{\pi_1} V_\tau^{\pi_2}(s) - V_\tau^{\pi_2}(s)\right).$$

Optimality

Define the Bellman optimality operator $\mathcal{T}_\tau$ as follows:

$$\mathcal{T}_\tau V(s) = \max_\pi \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') - \tau\log\pi(a|s) + \gamma V(s')\right].$$

Then $\mathcal{T}_\tau$ is monotone and γ-contraction with respect to $\|\cdot\|_\infty$.

Theorem 1 (Optimality)
Let V_τ^* be the solution to the Bellman optimality equation $\mathcal{T}_\tau V(s) = \mathcal{T}_\tau V(s)$. Then

$$V_\tau^*(s) = \max_\pi V_\tau^\pi(s).$$

Moreover, there exists an optimal policy π^* such that $V_\tau^{\pi^*} = V_\tau^*$.

Optimality

Proposition 1
Define $Q_\tau^*(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma V_\tau^*(s')\right]$. *It is evident that*

$$Q_\tau^*(s,a) = \max_\pi Q_\tau^\pi(s,a), \quad \forall s,\ a.$$

Moreover, one has $\pi^*(\cdot|s) \propto \exp\left(Q_\tau^*(s,\cdot)/\tau\right)$ *and*

$$V_\tau^*(s) = Q_\tau^*(s,a) - \tau\log\pi^*(a|s) \Leftrightarrow A_\tau^*(s,a) = 0, \quad \forall a.$$

- Recall that for the non-regularized case, one has $A^*(s,a) \leq 0,\ \forall a$. Moreover, $A_\tau^*(s,a) = 0,\ \forall a$ guarantees $\mathbb{E}_{a\sim\pi^*(\cdot|s)}\left[A_\tau^*(s,a)\right] = 0$ even $\pi^*(\cdot|s) > 0,\ \forall a$.

Lemma 3 (Sub-Optimality Lemma)
There holds

$$V_\tau^*(\mu) - V_\tau^\pi(\mu) = \frac{\tau}{1-\gamma}\sum_s d_\mu^\pi(s)\mathrm{KL}(\pi(\cdot|s)\|\pi^*(\cdot|s)).$$

Reverse Direction

Theorem 2
If

$$V(s) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma V(s') \right] - \tau \log \pi(a|s), \quad \forall s, a,$$

then $V = V_\tau^*$ and $\pi = \pi_\tau^*$.

Proof. Taking expectation with respect to $\pi(\cdot|s)$ on both sides yields $V = V_\tau^\pi$. Thus, V is a value function. By Lemma 5 in Lecture 7, the condition also means

$$\pi(\cdot|s) = \operatorname*{argmax}_{\tilde{\pi}(\cdot|s)} \mathbb{E}_{a \sim \tilde{\pi}(\cdot|s)} \left[\mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s,a,s') + \gamma V(s') \right] - \tau \log \tilde{\pi}(a|s) \right],$$

which implies $\mathcal{T}_\tau V(s) = V(s)$.

- This result essentially states that if $A_\tau^\pi(s,a) = 0, \forall\, s, a$, then π is the optimal policy. It is parallel to the non-regularized case: if $A^\pi(s,a) \leq 0, \forall\, s, a$, then π is an optimal policy. Note this result can also be proved using performance difference lemma.

Remark

- The optimal policy is unique with entropy regularization.
- It is evident that as $\tau \to 0$, $\pi_\tau^*(a|s) \to 0$ for $a \notin \operatorname{argmax} Q^*(s,a)$.
- Since one has

$$\max_a Q_\tau^*(s,a) \leq \tau \log\left(\left\|\exp\left(Q_\tau^*\left(s,\cdot\right)/\tau\right)\right\|_1\right) \leq \tau \log|\mathcal{A}| + \max_a Q_\tau^*(s,a),$$

it is easy to see that $V_\tau^*(s) \to \max_a Q^*(s,a) = V^*(s)$ as $\tau \to 0$.

Soft Policy Iteration

Soft Policy Iteration:

$$\pi_{k+1}(\cdot|s) = \operatorname*{argmax}_{\pi} \mathcal{T}_\tau^\pi V_\tau^{\pi_k} = \frac{\exp\left(Q_\tau^{\pi_k}(s,\cdot)/\tau\right)}{\|\exp\left(Q_\tau^{\pi_k}(s,\cdot)/\tau\right)\|_1}.$$

- γ-rate convergence, with local quadratic convergence.

"Elementary Analysis of Policy Gradient Methods" by Jiacai Liu, Wenye Li, and Ke Wei, 2024.

Policy Gradient Theorem

Theorem 3 (Policy Gradient Theorem)
Assume $\forall\theta, \sum_a \pi_\theta(a|s) = 1$ *for simplicity. One has*

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

- For softmax parameterization,

$$\nabla_{\theta_s} V_\tau^{\pi_\theta}(\mu) = \frac{d_\mu^{\pi_\theta}(s)}{1-\gamma}\pi_\theta(\cdot|s)A_\tau^{\pi_\theta}(s,\cdot).$$

Policy Gradient Methods

- Entropy softmax PG: in the parameter space,

$$\theta_{s,a}^{+} = \theta_{s,a} + \eta \frac{d_{\mu}^{\pi_\theta}(s)}{1-\gamma} \pi_\theta(a|s) A_{\tau}^{\pi_\theta}(s,a).$$

  In the policy space,

$$\pi_{s,a}^{+} \propto \pi_{s,a} \exp\left(\eta \frac{d_{\mu}^{\pi}(s)}{1-\gamma} \pi_\theta(a|s) A_{\tau}^{\pi_\theta}(s,a)\right).$$

- Entropy softmax NPG, in the parameter space,

$$\theta_{s,a}^{+} = \theta_{s,a} + \frac{\eta}{1-\gamma} A_{\tau}^{\pi_\theta}(s,a).$$

  In the policy space,

$$\pi_{s,a}^{+} \propto \pi_{s,a} \exp\left(\frac{\eta}{1-\gamma} A_{\tau}^{\pi}(s,a)\right) \propto (\pi_{s,a})^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta}{1-\gamma} Q_{\tau}^{\pi}(s,a)\right).$$

For linear convergence of entropy softmax PG and NPG, see “On the Global Convergence Rates of Softmax Policy Gradient Methods” by Jincheng Mei et al., 2020 and “Fast global convergence of natural policy gradient methods with entropy regularization” by Cen et al., 2022.

Table of Contents

Natural Policy Gradient (NPG)

Trust Region Policy Optimization (TRPO)

Proximal Policy Optimization (PPO)

Entropy Regularization

Deterministic Policy Gradient (DPG)

Deterministic Policy Parameterization

Consider the case where $\mathcal{S}$ and $\mathcal{A}$ are continuous, and use π_θ to denote a deterministic policy: $a = \pi_\theta(s)$ is an action.

- Average state value:

$$V^{\pi_\theta}(\mu) = \int_{\mathcal{S}} V^{\pi_\theta}(s_0)\mu(s_0)\mathrm{d}s_0 = \mathbb{E}_{\tau \sim p_\mu^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi_\theta(s_t), s_{t+1}) \right],$$

  where given trajectory $\tau = (s_t, \pi_\theta(s_t), s_{t+1})_{t=0}^{\infty}$,

$$p_\mu^{\pi_\theta}(\tau) = \mu(s_0) \prod_{t=0}^{\infty} p(s_{t+1}|s_t, \pi_\theta(s_t))$$

  is the probability density over τ. Note that there is no probability over action space since $\pi_\theta(s)$ selects a deterministic action.

- It is worth noting that $V^{\pi_\theta}(s) = Q^{\pi_\theta}(s, \pi_\theta(s))$.

Deterministic Policy Parameterization

- Similarly, we can express $V^{\pi_\theta}(\mu)$ over state space

$$\begin{aligned} V^{\pi_\theta}(\mu) &= \frac{1}{1-\gamma}\int_{\mathcal{S}} d_\mu^{\pi_\theta}(s)\mathrm{d}s \int_{\mathcal{S}} p(s'|s,\pi_\theta(s))r(s,\pi_\theta(s),s')\mathrm{d}s' \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{s'\sim p(\cdot|s,\pi_\theta(s))}\left[r(s,\pi_\theta(s),s')\right], \end{aligned}$$

where $d_\mu^{\pi_\theta}(s) = \mathbb{E}_{s_0\sim\mu}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_t(s|s_0,\pi_\theta)\right]$ is state visitation density, and $p_t(s|s_0,\pi_\theta)$ is the density over state space after transitioning t time steps. Note there is no expectation over action space since $\pi_\theta(s)$ is deterministic.

Deterministic Policy Gradient Theorem

Theorem 4 (Deterministic Policy Gradient Theorem)

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

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

Proof of Theorem 4

First note that

$$\begin{aligned} V^{\pi_\theta}(s_0) &= Q^{\pi_\theta}(s_0, \pi_\theta(s_0)) \\ &= \int_{\mathcal{S}} \big(r(s_0, \pi_\theta(s_0), s_1) + \gamma V^{\pi_\theta}(s_1)\big) p(s_1|s_0, \pi_\theta(s_0)) \mathrm{d}s_1. \end{aligned}$$

Therefore, one has

$$\begin{aligned} \nabla_\theta V^{\pi_\theta}(s_0) &= \int_{\mathcal{S}} \nabla_a r(s_0, a, s_1)|_{a=\pi_\theta(s_0)} \nabla_\theta \pi_\theta(s_0) p(s_1|s_0, \pi_\theta(s_0)) \mathrm{d}s_1 \\ &+ \int_{\mathcal{S}} r(s_0, \pi_\theta(s_0), s_1) \; \nabla p(s_1|s_0, a)|_{a=\pi_\theta(s_0)} \nabla_\theta \pi_\theta(s_0) \mathrm{d}s_1 \\ &+ \gamma \int_{\mathcal{S}} V^{\pi_\theta}(s_1) \; \nabla p(s_1|s_0, a)|_{a=\pi_\theta(s_0)} \nabla_\theta \pi_\theta(s_0) \mathrm{d}s_1 \\ &+ \gamma \int_{\mathcal{S}} \nabla_\theta V^{\pi_\theta}(s_1) p(s_1|s_0, \pi_\theta(s_0)) \mathrm{d}s_1. \end{aligned}$$

Proof of Theorem 4 (Cont'd)

Moreover, it is easy to verify that the sum of the first three terms is equal to

$$\nabla_\theta \pi_\theta(s_0)\ \nabla_a Q^{\pi_\theta}(s,a)\big|_{a=\pi_\theta(s_0)}\,.$$

Therefore,

$$\begin{aligned}\nabla_\theta V^{\pi_\theta}(s_0) &= \nabla_\theta \pi_\theta(s_0)\ \nabla_a Q^{\pi_\theta}(s,a)\big|_{a=\pi_\theta(s_0)} + \gamma \int_{\mathcal{S}} \nabla_\theta V^{\pi_\theta}(s_1) p(s_1|s_0,\pi_\theta(s_0))\mathrm{d}s_1 \\ &= \ldots \\ &= \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \pi_\theta(s_t) \nabla_a Q^{\pi_\theta}(s_t,a)|_{a=\pi_\theta(s_t)}|s_0,\pi_\theta\Big] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}}\left[\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s,a)|_{a=\pi_\theta(s)}\right].\end{aligned}$$

Averaging over all s_0 completes the proof of Theorem 1.

Deep Deterministic Policy Gradient (DDPG)

- DDPG is a policy gradient method which learns a deterministic policy π_θ 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 ω for action value function is overall the same to Fitted Q-learning.

See “Continuous control with deep reinforcement learning” by Lillicrap et al. 2016 for details.

Questions?