Neural Proximal/Trust Region Policy Optimization Attains Globally Optimal Policy

Presented by :
Atmani Hanan

University Mohammed VI Polytechnic

August 22, 2024

Presentation plan

1. Introduction and motivation
2. Neural PPO
3. Errors of Policy Improvement, Policy Evaluation and Propagation
4. Global Convergence of Neural PPO

Introduction

- PPO and TRPO have shown significant empirical success, their global convergence remains poorly understood due to the non-convexity of the policy space and neural network parametrization. To bridge this theory-practice gap, three key questions need to be addressed:
    1. How do PPO and TRPO converge to the optimal policy with infinite-dimensional updates?
    2. How does stochastic gradient descent improve the policy based on this approximate action-value function?

Neural Network Parametrization

- We consider the Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ is a campact state space, $\mathcal{A}$ is a finite action space.
- We denote by $v_k := v_{\pi_{\theta_k}}$: The stationary state distribution .
  $\sigma_k := \sigma_{\pi_{\theta_k}}$: The stationary state-action distribution.
  $\tilde{\sigma}_k := v_k \pi_0$: The auxiliary distribution
- We assume that $(s, a) \in \mathbb{R}^d$ for all $s \in \mathcal{S}$ and $a \in A$.
- We parametrize a functionn $u : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (policy π action-value function Q^π) by two-layer neural network, which is denoted by $NN(\alpha; m)$,

$$u_\alpha(s, a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i * \sigma([\alpha]_i^t (s, a)) \tag{1}$$

- m: The width of the neural network, $b_i \in [-1, 1] (i \in [m])$: the output weights. $\sigma(.)$: function activation (ReLU) $(\sigma(x) = max\{0, x\})$. $\alpha = ([\alpha]_1^t, ..., [\alpha]_m^t) \in \mathbb{R}^{md}$ with $[\alpha]_i \in \mathbb{R}^d$ $(i \in [m])$ are the input weights.
- We consider the random initialization

$$b_i \sim^{i.i.d} Unif([-1, 1]), [\alpha(0)]_i \sim^{i.i.d} \mathcal{N}(0, I_d/d), \quad \forall i \in [m]$$

- We restrict the input weights α to an L_2-ball centered at the initialization $\alpha(0)$ by the projection:

$$\Pi_{\mathcal{B}^0(R_\alpha)}(\alpha') = argmin_{\alpha \in \mathcal{B}^0(R_\alpha)}\{||\alpha - \alpha'||_2\},$$

where

$$\mathcal{B}^0(R_0) = \{\alpha : ||\alpha - \alpha(0)||_2 \leq R_\alpha\}$$

- Throughout training, we only update α, while keeping $b_i (i \in [m])$fixed at the initialization, we omit the the dependency on b_i $(i \in [m])$ in $NN(\alpha, m)$ and $u_\alpha(s, a)$.

Policy Improvement

- We consider the population version of the objective function:

$$L(\theta) = \mathbb{E}_{v_k}[\langle Q_{\omega_k}(s,.), \pi_\theta(.|s)\rangle - \beta_k KL(\pi_\theta(.|s)||\pi_{\theta_k}(.|s))]$$

- Where Q_{ω_k} is an estimator of $Q^{\pi_{\theta_k}}$
- We consider the energy-based policy $\pi(a|s) \propto \exp\{\tau^{-1} f\}$. Here $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the energy function and $\tau > 0$ is the temperature parameter.

Proposition

Let $\pi_{\theta_k} \propto \exp\left\{\tau_k^{-1} f_{\theta_k}\right\}$ be an energy-based policy. Given an estimator Q_{ω_k} of $Q^{\pi\theta_k}$, the update

$$\widehat{\pi}_{k+1} \leftarrow \operatorname{argmax}_\pi \left\{\mathbb{E}_{\nu_k}\left[\langle Q_{\omega_k}(s,\cdot), \pi(\cdot \mid s)\rangle - \beta_k \cdot \mathrm{KL}\left(\pi(\cdot \mid s) \| \pi_{\theta_k}(\cdot \mid s)\right)\right]\right\}$$

gives

$$\widehat{\pi}_{k+1} \propto \exp\left\{\beta_k^{-1} Q_{\omega_k} + \tau_k^{-1} f_{\theta_k}\right\} \quad (2)$$

Policy Improvement

- To represent the ideal improved policy $\widehat{\pi}_{k+1}$ in Proposition using the energy-based policy $\pi_{\theta_{k+1}} \propto \exp\left\{\tau_{k+1}^{-1} f_{\theta_{k+1}}\right\}$, we solve the subproblem of minimizing the MSE,

$$\theta_{k+1} \leftarrow \underset{\theta\in\mathcal{B}^0(R_f)}{\operatorname{argmin}} \mathbb{E}_{\tilde{\sigma}_k}\left[\left(f_\theta(s,a) - \tau_{k+1}\cdot\left(\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a)\right)\right)\right. \tag{3}$$

- Here we use the neural network parametrization $f_\theta = \mathrm{NN}(\theta; m_f)$ defined in (1), where θ denotes the input weights and m_f is the width.
- To solve (3), we use the SGD update: $\theta(t+1/2) \leftarrow \theta(t) - \eta\cdot\left(f_{\theta(t)}(s,a) - \tau_{k+1}\cdot\left(\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a)\right)\right)\cdot \nabla_\theta f_{\theta(t)}(s,a)$
  where $(s,a)\sim\widetilde{\sigma}_k$ and $\theta(t+1) \leftarrow \Pi_{\mathcal{B}^\circ(R_f)}(\theta(t+1/2))$. Here η is the stepsize.

Policy Evaluation

- To obtain the estimator Q_{ω_k} of $Q^{\pi_{\theta_k}}$ in (3.3), we solve the subproblem of minimizing the MSBE (Mean Squared Bellman Error),

$$\omega_k \leftarrow \underset{\omega \in \mathcal{B}^0(R_Q)}{\operatorname{argmin}} \mathbb{E}_{\sigma_k} \left[\left(Q_\omega(s,a) - \left[\mathcal{T}^{\pi_{\theta_k}} Q_\omega\right](s,a) \right)^2 \right] \tag{4}$$

- The Bellman evaluation operator $\mathcal{T}^\pi$ of a policy π is defined as:
$[\mathcal{T}^\pi Q](s,a) =$
$\mathbb{E}\left[(1-\gamma) \cdot r(s,a) + \gamma \cdot Q(s',a') \mid s' \sim \mathcal{P}(\cdot \mid s,a), a' \sim \pi(\cdot \mid s')\right]$
- We use the neural network parametrization $Q_\omega = \mathrm{NN}(\omega; m_Q)$ defined in (1), where ω denotes the input weights and m_Q is the width.

Policy Evaluation

- To solve (4) we use the TD update:
  $\omega(t+1/2) \leftarrow$
  $\omega(t) - \eta \cdot \big(Q_{\omega(t)}(s,a) - (1-\gamma) \cdot r(s,a) - \gamma \cdot Q_{\omega(t)}(s',a')\big) \cdot$
  $\nabla_\omega Q_{\omega(t)}(s,a)$
- where $(s,a) \sim \sigma_k, s' \sim \mathcal{P}(\cdot \mid s,a), a' \sim \pi_{\theta_k}(\cdot \mid s')$, and
  $\omega(t+1) = \Pi_{\mathcal{B}^\circ(R_Q)}(\omega(t+1/2))$.

Neural PPO Algorithm

Require: MDP$(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, penalty parameter β, widths m_f and m_Q, number of SGD and TD iterations T, number of TRPO iterations K, and projection radii $R_f \geq R_Q$

- for $k = 0, \ldots, K-1$ do
  1. Set temperature parameter $\tau_{k+1} \leftarrow \beta\sqrt{K}/(k+1)$ and penalty parameter $\beta_k \leftarrow \beta\sqrt{K}$
  2. Sample $\left\{\left(s_t, a_t, a_t^0, s_t', a_t'\right)\right\}_{t=1}^{T}$ with $\left(s_t, a_t\right) \sim \sigma_k, a_t^0 \sim \pi_0\left(\cdot \mid s_t\right), s_t' \sim \mathcal{P}\left(\cdot \mid s_t, a_t\right)$ and $a_t' \sim \pi_{\theta_k}\left(\cdot \mid s_t'\right)$
  3. Solve for $Q_{\omega_k} = \mathrm{NN}\left(\omega_k; m_Q\right)$ in (4) (Algorithm 3)
  4. Solve for $f_{\theta_{k+1}} = \mathrm{NN}\left(\theta_{k+1}; m_f\right)$ in (3) (Algorithm 2)
  5. Update policy: $\pi_{\theta_{k+1}} \propto \exp\left\{\tau_{k+1}^{-1} f_{\theta_{k+1}}\right\}$
- end for

Definition

For any constant $R > 0$, we define the function class

$$\mathcal{F}_{R,m} = \left\{ \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \cdot \nVdash \left\{ [\alpha(0)]_i^\top (s,a) > 0 \right\} \cdot [\alpha]_i^\top (s,a) : \|\alpha - \alpha(0)\|_2 \leq R \right\}$$

where $[\alpha(0)]_i$ and $b_i (i \in [m])$ are the random initialization

- Assumptions
  1. **Bounded Reward:** There exists a constant $R_{\max} > 0$ such that $R_{\max} = \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |r(s,a)|$, which implies $|V^\pi(s)| \leq R_{\max}$ and $|Q^\pi(s,a)| \leq R_{\max}$ for any policy π.
  2. **Action-Value Function Class:** It holds that $Q^\pi(s,a) \in \mathcal{F}_{R_Q, m_Q}$ for any π.
  3. **Regularity of Stationary Distribution:** There exists a constant $c > 0$ such that for any vector $z \in \mathbb{R}^d$ and $\zeta > 0$, it holds almost surely that $\mathbb{E}_{\sigma_n} \left[1 \left\{ \left| z^\top (s,a) \right| \leq \zeta \right\} \mid z \right] \leq c \cdot \zeta / \|z\|_2$ for any π.

Policy Improvement Error

Theorem

Suppose that Assumptions 1, 2, and 3 hold. We set $T \geq 64$ and the stepsize to be $\eta = T^{-1/2}$. Within the k-th iteration of Algorithm 1, the output $f_{\hat{\theta}}$ of Algorithm 2 satisfies

$$\mathbb{E}_{init,\ \bar{\sigma}_k}\left[\left(f_{\hat{\theta}}(s,a) - \tau_{k+1}\cdot\left(\beta_k^{-1}Q_{\omega_k}(s,a) + \tau_k^{-1}f_{\theta_k}(s,a)\right)\right)^2\right]$$
$$= O\left(R_f^2 T^{-1/2} + R_f^{5/2} m_f^{-1/4} + R_f^3 m_f^{-1/2}\right)$$

Policy Evaluation Error

Theorem

Suppose that Assumptions 1, 2, and 3 hold. We set $T \geq 64/(1-\gamma)^2$ and the stepsize to be $\eta = T^{-1/2}$. Within the k-th iteration of Algorithm 1, the output $Q_{\bar{\omega}}$ of Algorithm 3 satisfies

$$\mathbb{E}_{init,\ \sigma_k}\left[\left(Q_{\bar{\omega}}(s,a) - Q^{\pi_{\theta_k}}(s,a)\right)^2\right] = O\left(R_Q^2 T^{-1/2} + R_Q^{5/2} m_Q^{-1/4} + R_Q^3 m_Q\right.$$

Error Propagation

- π^*: Optimal policy.
- ν^*: Stationary state distribution under π^*.
- σ^*: Stationary state-action distribution under π^*.
- π_{k+1}: Improved policy based on $Q^{\pi_{\theta_k}}$, defined as: $\pi_{k+1} = \underset{\pi}{\operatorname{argmax}} \left\{\mathbb{E}_{\nu_k}\left[\langle Q^{\pi_{\theta_k}}(s,\cdot), \pi(\cdot, s)\rangle - \beta_k \cdot \mathrm{KL}\left(\pi(\cdot \mid s) \| \pi_{\theta_k}(\cdot \mid s)\right)\right]\right\}$
- Energy-based policy:

$$\pi_{k+1} \propto \exp\left\{\beta_k^{-1} Q^{\pi_{\theta_k}} + \tau_k^{-1} f_{\theta_k}\right\}$$

-

$$\phi_k^* = \mathbb{E}_{\tilde{\sigma}_k}\left[\left|\, \mathrm{d}\sigma^* / \mathrm{d}\widetilde{\sigma}_k - \mathrm{d}\left(\pi_{\theta_k}\nu^*\right) / \mathrm{d}\widetilde{\sigma}_k\right|^2\right]^{1/2}$$

$$\psi_k^* = \mathbb{E}_{\sigma_k}\left[\left|\, \mathrm{d}\sigma^* / \mathrm{d}\sigma_k - \mathrm{d}\nu^* / \mathrm{d}\nu_k\right|^2\right]^{1/2}$$

where $\mathrm{d}\sigma^* / \mathrm{d}\widetilde{\sigma}_k$, $\mathrm{d}\left(\pi_{\theta_k}\nu^*\right) / \mathrm{d}\widetilde{\sigma}_k$, $\mathrm{d}\sigma^* / \mathrm{d}\sigma_k$, and $\mathrm{d}\nu^* / \mathrm{d}\nu_k$ are the Radon-Nikodym derivatives.

Error Propagation

Lemma

Suppose that the policy improvement error in Line 4 of Algorithm 1 satisfies

$$\mathbb{E}_{\tilde{\sigma}_k}\left[\left(f_{\theta_{k+1}}(s,a)-\tau_{k+1}\cdot\left(\beta_k^{-1}Q_{\omega_k}(s,a)-\tau_k^{-1}f_{\theta_k}(s,a)\right)\right)^2\right]\leq\epsilon_{k+1}$$

and the policy evaluation error in Line 3 of Algorithm 1 satisfies

$$\mathbb{E}_{\sigma_k}\left[\left(Q_{\omega_k}(s,a)-Q^{\pi_{\theta_k}}(s,a)\right)^2\right]\leq\epsilon_k'$$

For π_{k+1} and $\pi_{\theta_{k+1}}$ obtained in Line 5 of Algorithm 1, we have

$$\left|\mathbb{E}_{\nu^*}\left[\left\langle\log\left(\pi_{\theta_{k+1}}(\cdot\mid s)/\pi_{k+1}(\cdot\mid s)\right),\pi^*(\cdot\mid s)-\pi_{\theta_k}(\cdot\mid s)\right\rangle\right]\right|\leq\varepsilon_k$$

where $\varepsilon_k=\tau_{k+1}^{-1}\epsilon_{k+1}\cdot\phi_{k+1}^*+\beta_k^{-1}\epsilon_k'\cdot\psi_k^*$.

Stepwise Energy Difference

Lemma

Under the same conditions of last Lemma , we have

$$\mathbb{E}_{\nu^*}\left[\left\|\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,\cdot) - \tau_k^{-1} f_{\theta_k}(s,\cdot)\right\|_\infty^2\right] \leq 2\varepsilon_k' + 2\beta_k^{-2} M$$

where $\varepsilon_k' = |\mathcal{A}| \cdot \tau_{k+1}^{-2} \epsilon_{k+1}^2$ *and*
$M = 2\mathbb{E}_{\nu^*}\left[\max_{a\in\mathcal{A}} \left(Q_{\omega_0}(s,a)\right)^2\right] + 2R_f^2.$

Convergence of Neural PPO

Theorem

Suppose that Assumptions 1, 2 and 3 hold. For the policy sequence $\{\pi_{\theta_k}\}_{k=1}^{K}$ attained by neural PPO in Algorithm 1, we have

$$\min_{0\leq k\leq K}\{\mathcal{L}(\pi^*)-\mathcal{L}(\pi_{\theta_k})\}\leq\frac{\beta^2\log|\mathcal{A}|+M+\beta^2\sum_{k=0}^{K-1}(\varepsilon_k+\varepsilon_k')}{(1-\gamma)\beta\cdot\sqrt{K}}$$

Here $\varepsilon_k=\tau_{k+1}^{-1}\epsilon_{k+1}\cdot\phi_k^*+\beta_k^{-1}\epsilon_k'\cdot\psi_k^*$ *and* $\varepsilon_k'=|\mathcal{A}|\cdot\tau_{k+1}^{-2}\epsilon_{k+1}^2$, *where* $\epsilon_{k+1}=O\left(R_f^2T^{-1/2}+R_f^{5/2}m_f^{-1/4}+R_f^3m_f^{-1/2}\right)$, $\epsilon_k'=O\left(R_Q^2T^{-1/2}+R_Q^{5/2}m_Q^{-1/4}+R_Q^3m_Q^{-1/2}\right)$. *Also, we have* $M=2\mathrm{E}_{\nu^*}\left[\max_{a\in\mathcal{A}}(Q_{\omega_0}(s,a))^2\right]+2R_f^2$.

Iteration Complexity

Corollary

Suppose that Assumptions 1, 2 and 3 hold. Let $m_f = \Omega\left(K^6 R_f^{10} \cdot \phi_k^{*4} + K^4 R_f^{10} \cdot |\mathcal{A}|^2\right)$, $m_Q = \Omega\left(K^2 R_Q^{10} \cdot \psi_k^{*4}\right)$, *and* $T = \Omega\left(K^3 R_f^4 \cdot \phi_k^{*2} + K^2 R_f^4 \cdot |\mathcal{A}| + K R_Q^4 \cdot \psi_k^{*2}\right)$ *for any* $0 \le k \le K$. *We have*

$$\min_{0 \le k \le K} \left\{\mathcal{L}\left(\pi^*\right) - \mathcal{L}\left(\pi_{\theta_k}\right)\right\} \le \frac{\beta^2 \log|\mathcal{A}| + M + O(1)}{(1-\gamma)\beta \cdot \sqrt{K}}$$