ICML | 2022

Thirty-ninth International Conference on Machine Learning

THINK BIG WE DO

An Analytical Update Rule For General Policy Optimization

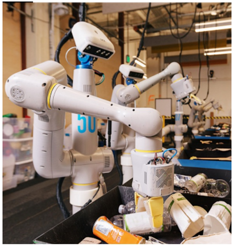
[*] **Hepeng Li**, [*] Nicholas Clavette, [*] Haibo He

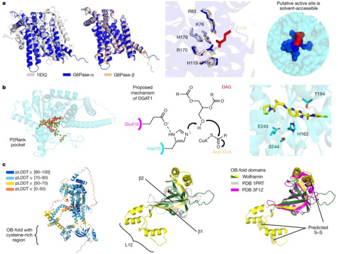
[*] Department of Electrical, Computer and Biomedical Engineering

University of Rhode Island

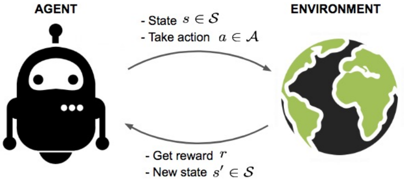
THE UNIVERSITY OF RHODE ISLAND

ICML International Conference On Machine Learning





RL Framework:



Value-Based Method

- PI: greedy policy: $a^* = \arg\max_a Q^{k-1}(s,a) \Rightarrow \pi^k$
- PE: $\mathcal{T}Q(s,a) \triangleq r(s,a) + \gamma \mathbb{E}_{s' \sim P, a' \sim \pi^k}[Q(s',a')] \Rightarrow Q^k$

$$\boxed{Q^0 \xrightarrow{\textbf{PI}} \pi^1 \xrightarrow{\textbf{PE}} Q^1 \xrightarrow{\textbf{PI}} \cdots \xrightarrow{\textbf{PE}} Q^*}$$

Policy Search Method

- Policy is parameterized by $\pi(a|s;\theta)$
- Policy update: $\theta^{k+1} \leftarrow \theta^k + \Delta\theta$ (policy gradient, random search, ...)

Advantages:

1. can learn stochastic policies
2. better convergence
3. effective for continuous actions

What are the limitations?

What are the limitations?

- only apply to parameterized policies

What are the limitations?

- only apply to parameterized policies
- difficult to integrate prior policy knowledge

What are the limitations?

- only apply to parameterized policies
- difficult to integrate prior policy knowledge
- sample inefficiency and high variance

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s, a)]$$

What are the limitations?

- only apply to parameterized policies
- difficult to integrate prior policy knowledge
- sample inefficiency and high variance

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s,a)]$$

- no improvement guarantee due to inappropriate choice of stepsize

What are the limitations?

- only apply to parameterized policies
- difficult to integrate prior policy knowledge
- sample inefficiency and high variance

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s,a)]$$

- no improvement guarantee due to inappropriate choice of stepsize

What if we directly search policy in a function space?

- optimize a functional

$$\max_\pi \; J(\pi), \;\; s.t. \;\; \pi \in \Pi$$

What are the limitations?

- only apply to parameterized policies
- difficult to integrate prior policy knowledge
- sample inefficiency and high variance

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s,a)]$$

- no improvement guarantee due to inappropriate choice of stepsize

What if we directly search policy in a function space?

- optimize a functional

$$\max_{\pi} J(\pi), \;\; s.t. \;\; \pi \in \Pi$$

- **A closed-form solution solving all these limitations?**

Modeling

Infinite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0\}$:

- $\mathcal{S}$ – state space $s \in \mathcal{S} \subseteq \mathbb{R}^m$ (continuous)
- $\mathcal{A}$ – action space $a \in \mathcal{A} \subseteq \mathbb{R}^n$ or $\mathcal{A} = \{a^1, \dots, a^n\}$
- P – transition kernel $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, \infty)$ (unknown)
- r – reward function $\mathcal{S} \times \mathcal{A} \to [r_{\min}, r_{\max}]$ (unknown)
- γ – discount factor $\gamma \to [0, 1)$
- ρ_0 – distribution of s_0 $\mathcal{S} \to [0, \infty)$

- objective: find an optimal policy π^* so that

$$\pi^* = \arg\max_{\pi} \; J(\pi) \;\; where \;\; J(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\textstyle\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

$$\tau = (s_0, a_0, s_1, \dots),\; s_0 \sim \rho_0(\cdot),\; s_{t+1} \sim P(\cdot|s_t, a_t),\; a_t \sim \pi(\cdot|s_t)$$

Modeling

Infinite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0\}$

Definitions and Notations:

- $V_\pi(s) = \mathbb{E}_{a_t, s_{t+1}, \ldots}\left[\sum_{l=t}^{\infty} \gamma^{l-t} r(s_l, a_l) | s_t = s, \pi\right]$
- $Q_\pi(s,a) = \mathbb{E}_{s_{t+1}, a_{t+1}, \ldots}\left[\sum_{l=t}^{\infty} \gamma^{l-t} r(s_l, a_l) | s_t = s, a_t = s, \pi\right]$
- $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$
- d^π: discounted state visitation density

$$d^\pi(s) = (1-\gamma)[\rho_0^\pi(s) + \gamma\rho_1^\pi(s) + \gamma^2\rho_2^\pi(s)] = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \rho_t^\pi(s)$$

where $\rho_t^\pi(\cdot)$ is the distribution of the state at step t.

Result 1: A Closed-From Policy Update Rule

Theorem (Monotonic Improvement Guarantee)

For any stochastic policies $\pi_{\text{new}}, \pi_{\text{old}}$ *that are continuously differentiable on the state space* $\mathcal{S}$*, the inequality*

$$J(\pi_{\text{new}}) \geq J(\pi_{\text{old}}) \text{ holds when } \pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a \sim \pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]}$$

where $\alpha_{\pi_{\text{old}}} = A_{\pi_{\text{old}}} / C_{\pi_{\text{old}}}$ *and* $C_{\pi_{\text{old}}}$ *is a constant*

$$C_{\pi_{\text{old}}} = \frac{\gamma^2 \epsilon}{(1-\gamma)^3}, \ \epsilon = \max_{s,a} |A_{\pi_{\text{old}}}(s,a)|, \gamma \in [0.5, 1).$$

Result 1: A Closed-From Policy Update Rule

Theorem (Monotonic Improvement Guarantee)

For any stochastic policies $\pi_{\text{new}}, \pi_{\text{old}}$ that are continuously differentiable on the state space $\mathcal{S}$, the inequality

$$J(\pi_{\text{new}}) \geq J(\pi_{\text{old}}) \text{ holds when } \pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{e^{\alpha \pi_{\text{old}}}}{\mathbb{E}_{a \sim \pi_{\text{old}}}\left[e^{\alpha \pi_{\text{old}}}\right]}$$

where $\alpha_{\pi_{\text{old}}} = A_{\pi_{\text{old}}} / C_{\pi_{\text{old}}}$ and $C_{\pi_{\text{old}}}$ is a constant

$$C_{\pi_{\text{old}}} = \frac{\gamma^2 \epsilon}{(1-\gamma)^3}, \; \epsilon = \max_{s,a} |A_{\pi_{\text{old}}}(s,a)|, \gamma \in [0.5, 1).$$

- The policy update rule is **off-policy**

Result 1: A Closed-From Policy Update Rule

Theorem (Monotonic Improvement Guarantee)

For any stochastic policies $\pi_{\text{new}}, \pi_{\text{old}}$ that are continuously differentiable on the state space $\mathcal{S}$, the inequality

$$J(\pi_{\text{new}}) \geq J(\pi_{\text{old}}) \quad \textit{holds when} \quad \pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]}$$

where $\alpha_{\pi_{\text{old}}} = A_{\pi_{\text{old}}} / C_{\pi_{\text{old}}}$ *and* $C_{\pi_{\text{old}}}$ *is a constant*

$$C_{\pi_{\text{old}}} = \frac{\gamma^2 \epsilon}{(1-\gamma)^3}, \ \epsilon = \max_{s,a} |A_{\pi_{\text{old}}}(s,a)|, \gamma \in [0.5, 1).$$

- The policy update rule is **off-policy**
- Derived from TRPO[1] based on a new bound on policy performance

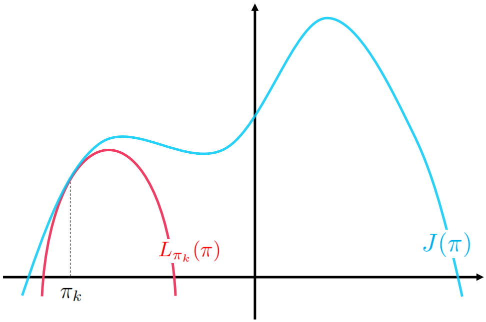
[1] J. Schulman et al. (2015). "Trust Region Policy Optimization". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, pp. 1889–1897

Trust Region Policy Optimization (TRPO)

1. Approximate $J(\pi)$ around π_k by a surrogate model

$$L_{\pi_k}(\pi) = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_k}, a\sim\pi}[A_{\pi_k}(s,a)]$$

2. Restrict policy search to the neighborhood of π_k



Trust Region Policy Optimization (TRPO)

1. Approximate $J(\pi)$ around π_k by a surrogate model

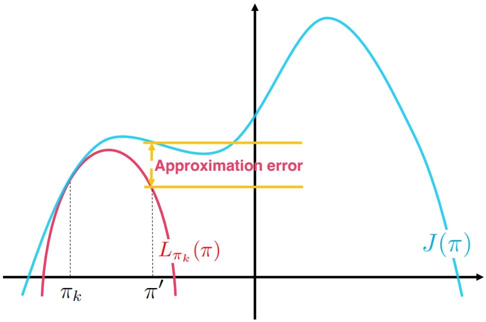
$$L_{\pi_k}(\pi) = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_k}, a\sim\pi}[A_{\pi_k}(s,a)]$$

2. Restrict policy search to the neighborhood of π_k

Bound of the approximation error:

$$\left|J(\pi') - L_{\pi_k}(\pi')\right| \leq C \max_s D_{\mathrm{KL}}[\pi' \| \pi_k](s),$$

$$\textit{where } C = \frac{4\gamma\epsilon}{(1-\gamma)^2}, \quad \epsilon = \max_{s,a} |A_{\pi_k}(s,a)|$$



Trust Region Policy Optimization (TRPO)

1. Approximate $J(\pi)$ around π_k by a surrogate model

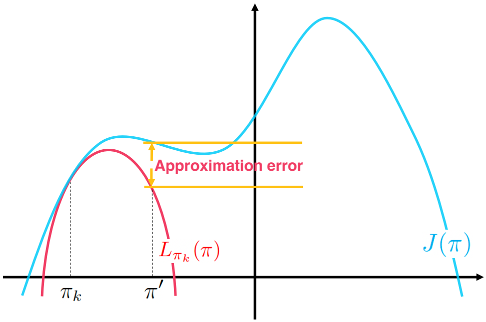
$$L_{\pi_k}(\pi) = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_k}, a\sim\pi}[A_{\pi_k}(s,a)]$$

2. Restrict policy search to the neighborhood of π_k

Bound of the approximation error:

$$\left|J(\pi') - L_{\pi_k}(\pi')\right| \leq C \max_s D_{\mathrm{KL}}[\pi' \| \pi_k](s),$$

where $C = \dfrac{4\gamma\epsilon}{(1-\gamma)^2}, \ \epsilon = \max_{s,a} |A_{\pi_k}(s,a)|$



Lower bound of policy performance:

$$J(\pi') \geq \underline{L_{\pi_k}(\pi') - C \max_s D_{\mathrm{KL}}[\pi' \| \pi_k](s)}$$

Maximizing the lower bound guarantees an improved policy

Trust Region Policy Optimization (TRPO)

1. Approximate $J(\pi)$ around π_k by a surrogate model

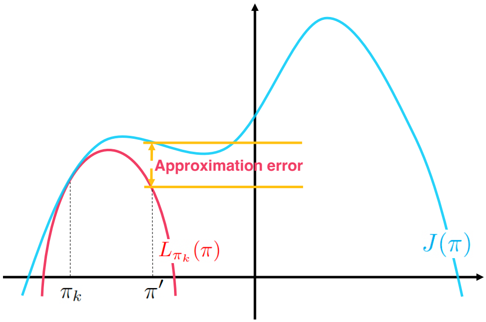
$$L_{\pi_k}(\pi) = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_k}, a\sim\pi}[A_{\pi_k}(s,a)]$$

2. Restrict policy search to the neighborhood of π_k

Bound of the approximation error:

$$\left|J(\pi') - L_{\pi_k}(\pi')\right| \leq C \max_s D_{\mathrm{KL}}[\pi' \| \pi_k](s),$$

where $C = \frac{4\gamma\epsilon}{(1-\gamma)^2}, \quad \epsilon = \max_{s,a}|A_{\pi_k}(s,a)|$



Lower bound of policy performance:

$$J(\pi') \geq \underline{L_{\pi_k}(\pi') - C\boxed{\max_s D_{\mathrm{KL}}[\pi' \| \pi_k](s)}}$$

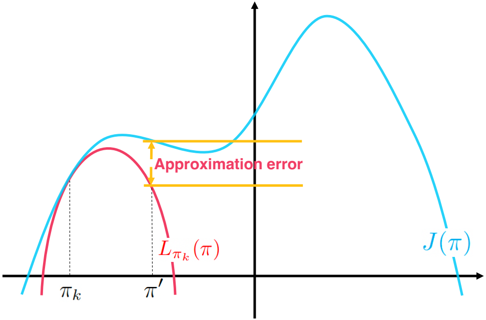
Maximizing the lower bound guarantees an improved policy

Trust Region Policy Optimization (TRPO)

1. Approximate $J(\pi)$ around π_k by a surrogate model

$$L_{\pi_k}(\pi) = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_k}, a\sim\pi}[A_{\pi_k}(s,a)]$$

2. Restrict policy search to the neighborhood of π_k



Bound of the approximation error:

$$\left|J(\pi') - L_{\pi_k}(\pi')\right| \le C \max_s D_{\mathrm{KL}}[\pi' \| \pi_k](s),$$

$$\textit{where } C = \frac{4\gamma\epsilon}{(1-\gamma)^2}, \quad \epsilon = \max_{s,a}|A_{\pi_k}(s,a)|$$

Lower bound of policy performance:

$$J(\pi') \ge \underline{L_{\pi_k}(\pi') - C\boxed{\max_s D_{\mathrm{KL}}[\pi' \| \pi_k](s)}} \quad \approx \quad \begin{array}{ll} \max_{\pi'} & L_{\pi_k}(\pi') \\ s.t. & \mathbb{E}_{s\sim d^{\pi_k}}[D_{\mathrm{KL}}[\pi' \| \pi_k](s)] \le \delta \end{array}$$

Result 2: A Tighter Lower Bound on Policy Performance

Theorem (Upper Bound on Surrogate Approximation Error)

For any stochastic policies π', π and discount factor $\gamma \in [0.5, 1)$, the following bound holds:

$$\left|J(\pi') - L_\pi(\pi')\right| \leq \frac{1}{1-\gamma} C_\pi \underline{\mathbb{E}_{s\sim d^\pi}\left[D_{\mathrm{KL}}[\pi' \| \pi](s)\right]},$$

$$\textit{where } C_\pi = \frac{\gamma^2 \epsilon}{(1-\gamma)^3}, \ \epsilon = \max_{s,a} |A_\pi(s,a)|.$$

A new lower bound on performance:

$$J(\pi') \geq L_\pi(\pi') - \frac{1}{1-\gamma} C_\pi \mathbb{E}_{s\sim d^\pi}\left[D_{\mathrm{KL}}[\pi' \| \pi](s)\right]$$

This result relates the bound to the expected KL, which is tighter than the max KL.

Derive the Update Rule using Calculus of Variation

The lower bound around π_k:

$$\underline{J(\pi')} = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}, a\sim \pi'}\left[A_{\pi_k}(s,a) - C_{\pi_k}\log\frac{\pi'(a|s)}{\pi_k(a|s)}\right]$$

Derive the Update Rule using Calculus of Variation

The lower bound around π_k:

$$\underline{J(\pi')} = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}, a\sim \pi'}\left[A_{\pi_k}(s,a) - C_{\pi_k}\log\frac{\pi'(a|s)}{\pi_k(a|s)}\right]$$

Maximizing $\underline{J(\pi')}$ is equivalent to:

$$\max_{\pi'} \iint d^{\pi_k}(s)\pi'(a|s)\left[A_{\pi_k}(s,a) - C_{\pi_k}\log\frac{\pi'(a|s)}{\pi_k(a|s)}\right]dsda$$

$$s.t. \int \pi'(a|s)da = 1$$

Derive the Update Rule using Calculus of Variation

The lower bound around π_k:

$$\underline{J(\pi')} = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi},a\sim\pi'}\left[A_{\pi_k}(s,a) - C_{\pi_k}\log\frac{\pi'(a|s)}{\pi_k(a|s)}\right]$$

Maximizing $\underline{J(\pi')}$ is equivalent to:

$$\begin{aligned}\max_{\pi'}\ &\iint d^{\pi_k}(s)\pi'(a|s)\left[A_{\pi_k}(s,a) - C_{\pi_k}\log\frac{\pi'(a|s)}{\pi_k(a|s)}\right]dsda\\ s.t.\ &\int \pi'(a|s)da = 1\end{aligned}$$

Euler-Lagrange equation: $A_{\pi_k} - C_{\pi_k}\log\pi' - C_{\pi_k} + C_{\pi_k}\log\pi_k - \lambda = 0$

Result 3: The Update Rule for Multi-Agent RL

Corollary

For any stochastic policies $\pi^i_{\text{new}}, \pi^i_{\text{old}}$ *of agent* i *that are continuously differentiable on the local observation space* $\mathcal{O}^i$*, the inequality,*

$$J(\pi_{\text{new}}) \geq J(\pi_{\text{old}}) \quad \textit{holds when}$$

$$\pi^i_{\text{new}} = \pi^i_{\text{old}} \cdot \frac{e^{\alpha \pi_{\text{old}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha\pi_{\text{old}}}\right]} \quad \text{and} \quad \pi^{-i}_{\text{new}} = \pi^{-i}_{\text{old}},$$

where $\pi^{-i}_{\text{new}}, \pi^{-i}_{\text{old}}$ *are the joint policies of all agents except* i*.*

Result 3: The Update Rule for Multi-Agent RL

Corollary

For any stochastic policies $\pi^i_{\text{new}}, \pi^i_{\text{old}}$ of agent i that are continuously differentiable on the local observation space $\mathcal{O}^i$, the inequality,

$$J(\pi_{\text{new}}) \geq J(\pi_{\text{old}}) \quad \textit{holds when}$$

$$\pi^i_{\text{new}} = \pi^i_{\text{old}} \cdot \frac{e^{\alpha \pi_{\text{old}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha \pi_{\text{old}}}\right]} \quad \text{and} \quad \underline{\pi^{-i}_{\text{new}} = \pi^{-i}_{\text{old}},}$$

Environment becomes stationary for agent i

where $\pi^{-i}_{\text{new}}, \pi^{-i}_{\text{old}}$ are the joint policies of all agents except i.

Connections to Prior Work

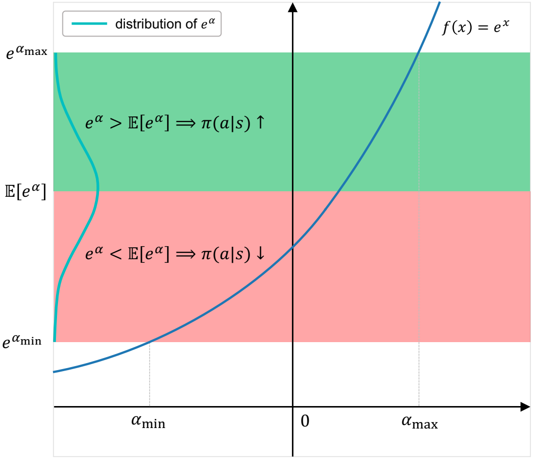
- Proximal Policy Optimization
- Value-based Methods
- Relative Entropy Policy Search
- Soft Actor-Critic

Proximal Policy Optimization (PPO)

Recall our policy update rule:

$$\pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]} \quad \text{where}$$

$$\alpha_{\pi_{\text{old}}} = \frac{A_{\pi_{\text{old}}}(s,a)}{\max_{s,a}|A_{\pi_{\text{old}}}(s,a)|} \cdot \frac{(1-\gamma)^3}{\gamma^2}$$



An explanation of the policy update rule

[2] J. Schulman et al. (2017). *Proximal Policy Optimization Algorithms*. DOI: 10.48550/ARXIV.1707.06347

Proximal Policy Optimization (PPO)

Recall our policy update rule:

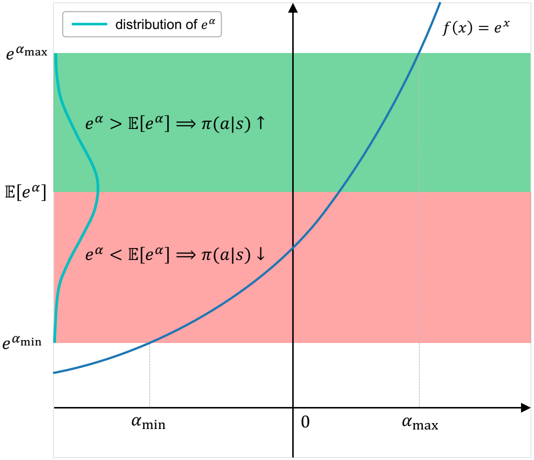
$$\pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]} \quad \text{where}$$

$$\alpha_{\pi_{\text{old}}} = \frac{A_{\pi_{\text{old}}}(s,a)}{\max_{s,a}|A_{\pi_{\text{old}}}(s,a)|} \cdot \frac{(1-\gamma)^3}{\gamma^2}$$

Assume $\alpha_{\pi_{\text{old}}} \in [\alpha_{\min}, \alpha_{\max}]$, then we have

$$\frac{\pi_{\text{new}}}{\pi_{\text{old}}} \in \left[\frac{e^{\alpha_{\min}}}{Z}, \frac{e^{\alpha_{\max}}}{Z}\right] = [1-\epsilon_1, 1+\epsilon_2]$$

where $Z = \mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]$ and $\epsilon_1, \epsilon_2 \geq 0, \epsilon_1 < 1$.



An explanation of the policy update rule

[2] J. Schulman et al. (2017). *Proximal Policy Optimization Algorithms.* DOI: 10.48550/ARXIV.1707.06347

Proximal Policy Optimization (PPO)

Recall our policy update rule:

$$\pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a \sim \pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]} \quad \text{where}$$

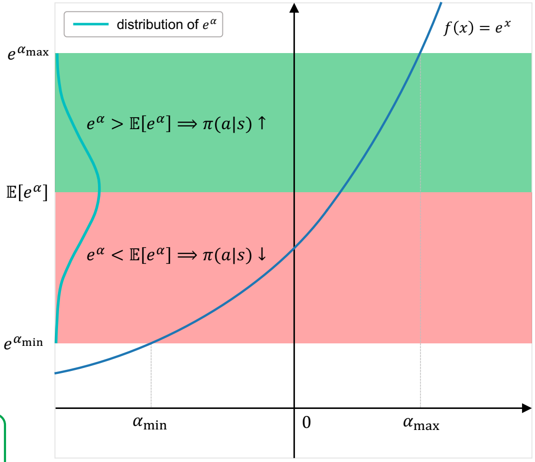
$$\alpha_{\pi_{\text{old}}} = \frac{A_{\pi_{\text{old}}}(s,a)}{\max_{s,a}|A_{\pi_{\text{old}}}(s,a)|} \cdot \frac{(1-\gamma)^3}{\gamma^2}$$

Assume $\alpha_{\pi_{\text{old}}} \in [\alpha_{\min}, \alpha_{\max}]$, then we have

$$\frac{\pi_{\text{new}}}{\pi_{\text{old}}} \in \left[\frac{e^{\alpha_{\min}}}{Z}, \frac{e^{\alpha_{\max}}}{Z}\right] = [1-\epsilon_1, 1+\epsilon_2]$$

where $Z = \mathbb{E}_{a \sim \pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]$ and $\epsilon_1, \epsilon_2 \geq 0, \epsilon_1 < 1$.

This helps explain why clipping policy ratio works and closes the gap between theory and practice in PPO[2].



An explanation of the policy update rule

[2] J. Schulman et al. (2017). *Proximal Policy Optimization Algorithms*. DOI: 10.48550/ARXIV.1707.06347

Proximal Policy Optimization (PPO)

Recall our policy update rule:

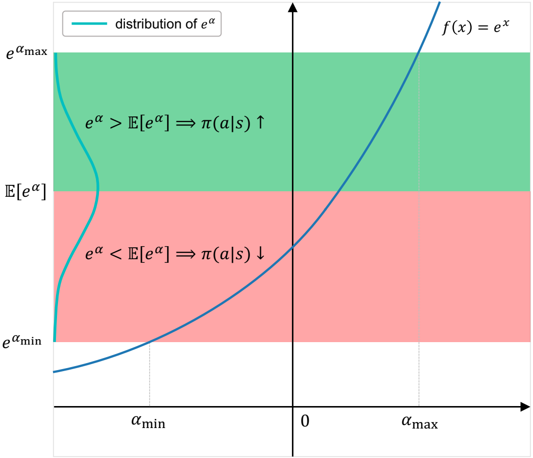
$$\pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]} \quad \text{where}$$

$$\alpha_{\pi_{\text{old}}} = \frac{A_{\pi_{\text{old}}}(s,a)}{\max_{s,a}|A_{\pi_{\text{old}}}(s,a)|} \cdot \frac{(1-\gamma)^3}{\gamma^2}$$

Objective of TRPO/PPO[2]:

$$\max_{\pi} \ \mathbb{E}_{s\sim d^{\pi_{\text{old}}}, a\sim\pi_{\text{old}}}\left[\frac{\pi(a|s)}{\pi_{\text{old}}(a|s)} A_{\pi_{old}}(s,a)\right]$$

- $\pi(a|s)\uparrow$ to gain weights for large A values
- $\pi(a|s)\downarrow$ to lose weights for small A values



An explanation of the policy update rule

[2] J. Schulman et al. (2017). *Proximal Policy Optimization Algorithms*. DOI: 10.48550/ARXIV.1707.06347

Value-Based Methods

For discrete actions, the update rule can be written as:

$$\pi_{\text{new}}(a^i|s) = \pi_{\text{old}}(a^i|s) \cdot \frac{e^{A_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{A_{\pi_{\text{old}}}(s,a^j)/C_{\pi_{\text{old}}}}}$$

Value-Based Methods

For discrete actions, the update rule can be written as:

$$\pi_{\text{new}}(a^i|s) = \pi_{\text{old}}(a^i|s) \cdot \frac{e^{A_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{A_{\pi_{\text{old}}}(s,a^j)/C_{\pi_{\text{old}}}}}$$

$$= \pi_{\text{old}}(a^i|s) \cdot \frac{e^{\left[Q_{\pi_{\text{old}}}(s,a^i)\cancel{-V_{\pi_{\text{old}}}(s)}\right]/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{\left[Q_{\pi_{\text{old}}}(s,a^j)\cancel{-V_{\pi_{\text{old}}}(s)}\right]/C_{\pi_{\text{old}}}}}$$

Value-Based Methods

For discrete actions, the update rule can be written as:

$$\pi_{\text{new}}(a^i|s) = \pi_{\text{old}}(a^i|s) \cdot \frac{e^{A_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{A_{\pi_{\text{old}}}(s,a^j)/C_{\pi_{\text{old}}}}}$$

$$= \pi_{\text{old}}(a^i|s) \cdot \frac{e^{\left[Q_{\pi_{\text{old}}}(s,a^i) \cancel{- V_{\pi_{\text{old}}}(s)}\right]/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{\left[Q_{\pi_{\text{old}}}(s,a^j) \cancel{- V_{\pi_{\text{old}}}(s)}\right]/C_{\pi_{\text{old}}}}}$$

$$= \frac{\pi_{\text{old}}(a^i|s)\omega^i_{\pi_{\text{old}}}}{\sum_j \pi_{\text{old}}(a^j|s)\omega^j_{\pi_{\text{old}}}}, \text{ where } \omega^i_{\pi_{\text{old}}} = e^{Q_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}}$$

Value-Based Methods

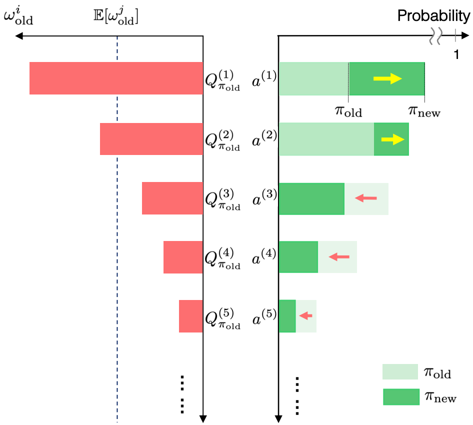
For discrete actions, the update rule can be written as:

$$\pi_{\text{new}}(a^i|s) = \pi_{\text{old}}(a^i|s) \cdot \frac{e^{A_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{A_{\pi_{\text{old}}}(s,a^j)/C_{\pi_{\text{old}}}}}$$

$$= \pi_{\text{old}}(a^i|s) \cdot \frac{e^{\left[Q_{\pi_{\text{old}}}(s,a^i) \cancel{- V_{\pi_{\text{old}}}(s)}\right]/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{\left[Q_{\pi_{\text{old}}}(s,a^j) \cancel{- V_{\pi_{\text{old}}}(s)}\right]/C_{\pi_{\text{old}}}}}$$

$$= \frac{\pi_{\text{old}}(a^i|s)\omega^i_{\pi_{\text{old}}}}{\sum_j \pi_{\text{old}}(a^j|s)\omega^j_{\pi_{\text{old}}}}, \text{ where } \omega^i_{\pi_{\text{old}}} = e^{Q_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}}$$

A softmax function of $Q_{\pi_{\text{old}}}$ weighted by π_{old}



An explanation for discrete actions

Value-Based Methods

For discrete actions, the update rule can be written as:

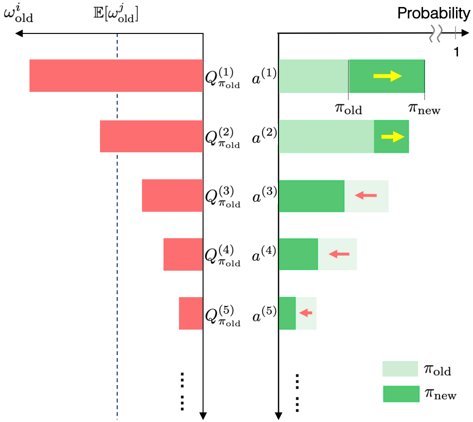
$$\pi_{\text{new}}(a^i|s) = \pi_{\text{old}}(a^i|s) \cdot \frac{e^{A_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{A_{\pi_{\text{old}}}(s,a^j)/C_{\pi_{\text{old}}}}}$$

$$= \pi_{\text{old}}(a^i|s) \cdot \frac{e^{\left[Q_{\pi_{\text{old}}}(s,a^i) \cancel{- V_{\pi_{\text{old}}}(s)}\right]/C_{\pi_{\text{old}}}}}{\sum_j \pi_{\text{old}}(a^j|s) e^{\left[Q_{\pi_{\text{old}}}(s,a^j) \cancel{- V_{\pi_{\text{old}}}(s)}\right]/C_{\pi_{\text{old}}}}}$$

$$= \frac{\pi_{\text{old}}(a^i|s)\omega^i_{\pi_{\text{old}}}}{\sum_j \pi_{\text{old}}(a^j|s)\omega^j_{\pi_{\text{old}}}}, \text{ where } \omega^i_{\pi_{\text{old}}} = e^{Q_{\pi_{\text{old}}}(s,a^i)/C_{\pi_{\text{old}}}}$$

A softmax function of $Q_{\pi_{\text{old}}}$ weighted by π_{old}

1. Actions with larger Q will be more likely to be selected
2. The policy acts like a stochastic analogy of ϵ-greedy



An explanation for discrete actions

Relative Entropy Policy Search (REPS)

A similar update rule was derived in REPS[3]:

$$\begin{aligned}&\max_{\pi} J(\pi)\\ &s.t.\ D_{KL}(p^{\pi}||q) \leq \epsilon\end{aligned} \quad \Longrightarrow \quad \pi(a|s) = \frac{q(s,a)\exp\left(\frac{1}{\eta}\delta_\theta(s,a)\right)}{\sum_b q(s,b)\exp\left(\frac{1}{\eta}\delta_\theta(s,b)\right)}$$

- $p^{\pi}(s,a) = d^{\pi}(s)\pi(a|s)$ is the state-action distribution generated by π
- $q(s,a)$ is the observed data distribution
- $\delta_\theta(s,a)$ is the Bellman error

[3] J. Peters et al. (2010). "Relative Entropy Policy Search". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI'10. Atlanta, Georgia: AAAI Press, pp. 1607–1612

Relative Entropy Policy Search (REPS)

A similar update rule was derived in REPS[3]:

$$\begin{aligned}&\max_{\pi} J(\pi)\\ &s.t.\ D_{KL}(p^{\pi}||q) \leq \epsilon\end{aligned} \quad \Longrightarrow \quad \pi(a|s) = \frac{q(s,a)\exp\left(\frac{1}{\eta}\delta_{\theta}(s,a)\right)}{\sum_b q(s,b)\exp\left(\frac{1}{\eta}\delta_{\theta}(s,b)\right)}$$

- $p^{\pi}(s,a) = d^{\pi}(s)\pi(a|s)$ is the state-action distribution generated by π
- $q(s,a)$ is the observed data distribution
- $\delta_{\theta}(s,a)$ is the Bellman error

If q is generated by π_{old}, i.e. $q(s,a) = d^{\pi_{\text{old}}}(s)\pi_{\text{old}}(a|s)$, then our update rule is obtained by replacing $\delta_{\theta}(s,a)$ with $A_{\pi_{\text{old}}}(s,a)$.

[3] J. Peters et al. (2010). "Relative Entropy Policy Search". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI'10. Atlanta, Georgia: AAAI Press, pp. 1607–1612

Relative Entropy Policy Search (REPS)

A similar update rule was derived in REPS[3]:

$$\begin{aligned}&\max_{\pi} J(\pi)\\ &\; s.t.\; D_{KL}(p^{\pi}||q) \leq \epsilon\end{aligned} \quad \Longrightarrow \quad \pi(a|s) = \frac{q(s,a)\exp\left(\frac{1}{\eta}\delta_\theta(s,a)\right)}{\sum_b q(s,b)\exp\left(\frac{1}{\eta}\delta_\theta(s,b)\right)}$$

However, REPS

- only applies to discrete actions
- needs to optimize the dual problem to determine the Lagrange multiplier η
- no monotonic improvement guarantee

[3] J. Peters et al. (2010). "Relative Entropy Policy Search". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI'10. Atlanta, Georgia: AAAI Press, pp. 1607–1612

Soft Actor-Critic (SAC)

We can derive SAC[4,5] as a special case of our update rule. Note that

$$\pi_{\text{new}}(a|s) = \pi_{\text{old}}(a|s) \cdot \frac{e^{(Q_{\pi_{\text{old}}}(s,a) \cancel{- V_{\pi_{\text{old}}}(s)})/C_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{(Q_{\pi_{\text{old}}}(s,a) \cancel{- V_{\pi_{\text{old}}}(s)})/C_{\pi_{\text{old}}}}\right]}$$

$$= \frac{1}{Z}\exp\left(Q_{\pi_{\text{old}}}(s,a)/C_{\pi_{\text{old}}} + \log \pi_{\text{old}}(a|s)\right),$$

where $Z = \mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{Q_{\pi_{\text{old}}}(s,a)/C_{\pi_{\text{old}}}}\right]$. To optimize a policy π, we can minimize the KL of π and π_{new}:

$$\min_{\pi} D_{\text{KL}}\left(\pi(\cdot|s) \middle\| \frac{\exp\left(\frac{1}{C_{\pi_{\text{old}}}}\widetilde{Q}_{\pi_{\text{old}}}(s,\cdot)\right)}{Z}\right), \tag{1}$$

where $\widetilde{Q}_{\pi_{\text{old}}} = Q_{\pi_{\text{old}}}(s,a) + C_{\pi_{\text{old}}} \log \pi_{\text{old}}(a|s)$ is the soft Q-function.

---

[4] T. Haarnoja et al. (Oct. 2018b). “Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1861–1870

[5] T. Haarnoja et al. (2018a). *Soft Actor-Critic Algorithms and Applications*. DOI: 10.48550/ARXIV.1812.05905

Conclusion & Future Work

1. Monotonic Guarantee and Function Approximation

2. Tightness of the Bound in Terms of γ

3. Simultaneous Update for Multi-Agent RL

Conclusion & Future Work

1. Monotonic Guarantee and Function Approximation

$$\pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{\exp\{A_{\pi_{\text{old}}}/C_{\pi_{\text{old}}}\}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[\exp\{A_{\pi_{\text{old}}}/C_{\pi_{\text{old}}}\}\right]}, \text{ where } C_{\pi_{\text{old}}} = \frac{\gamma^2}{(1-\gamma)^3} \cdot \max_{s,a} |A_{\pi_{\text{old}}}(s,a)|.$$

2. Tightness of the Bound in Terms of γ

3. Simultaneous Update for Multi-Agent RL

Conclusion & Future Work

1. Monotonic Guarantee and Function Approximation

$$\pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{\exp\{A_{\pi_{\text{old}}}/C_{\pi_{\text{old}}}\}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[\exp\{A_{\pi_{\text{old}}}/C_{\pi_{\text{old}}}\}\right]}, \text{ where } C_{\pi_{\text{old}}} = \frac{\gamma^2}{(1-\gamma)^3} \cdot \max_{s,a} |A_{\pi_{\text{old}}}(s,a)|.$$

2. Tightness of the Bound in Terms of γ

$$\left|J(\pi') - L_\pi(\pi')\right| \le \frac{1}{1-\gamma} C_\pi \mathbb{E}_{s\sim d^\pi}\left[D_{\text{KL}}[\pi' \| \pi](s)\right]$$

3. Simultaneous Update for Multi-Agent RL

Conclusion & Future Work

1. Monotonic Guarantee and Function Approximation

$$\pi_{\text{new}} = \pi_{\text{old}} \cdot \frac{\exp\{A_{\pi_{\text{old}}}/C_{\pi_{\text{old}}}\}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[\exp\{A_{\pi_{\text{old}}}/C_{\pi_{\text{old}}}\}\right]}, \text{ where } C_{\pi_{\text{old}}} = \frac{\gamma^2}{(1-\gamma)^3} \cdot \max_{s,a} |A_{\pi_{\text{old}}}(s,a)|.$$

2. Tightness of the Bound in Terms of γ

$$\left|J(\pi') - L_\pi(\pi')\right| \le \frac{1}{1-\gamma} C_\pi \mathbb{E}_{s\sim d^\pi}\left[D_{\text{KL}}[\pi' \| \pi](s)\right]$$

3. Simultaneous Update for Multi-Agent RL

$$\pi^i_{\text{new}} = \pi^i_{\text{old}} \cdot \frac{e^{\alpha_{\pi_{\text{old}}}}}{\mathbb{E}_{a\sim\pi_{\text{old}}}\left[e^{\alpha_{\pi_{\text{old}}}}\right]} \text{ and } \pi^{-i}_{\text{new}} = \pi^{-i}_{\text{old}}$$

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. ECCS 1917275. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Thank you for your attention!

THE UNIVERSITY OF RHODE ISLAND

Part I

Appendix

References I

Haarnoja, T. et al. (2018a). *Soft Actor-Critic Algorithms and Applications*. DOI: 10.48550/ARXIV.1812.05905.

Haarnoja, T. et al. (Oct. 2018b). “Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1861–1870.

Peters, J., K. Mülling, and Y. Altün (2010). “Relative Entropy Policy Search”. In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. AAAI’10. Atlanta, Georgia: AAAI Press, pp. 1607–1612.

Schulman, J. et al. (2015). “Trust Region Policy Optimization”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML’15. Lille, France: JMLR.org, pp. 1889–1897.

Schulman, J. et al. (2017). *Proximal Policy Optimization Algorithms*. DOI: 10.48550/ARXIV.1707.06347.