Mirror Learning: A Unifying Framework of Policy Optimisation

Jakub Grudzien Kuba,

Christian Schroeder de Witt, Jakob Foerster

ICML 2022

Reinforcement Learning: Problem Formulation

Problem Formulation

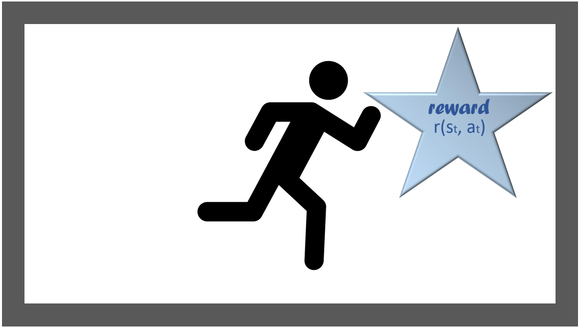
At time step t, the agent is at state s_t

Problem Formulation

The agent takes action $a_t \sim \pi(\cdot|s_t)$

Problem Formulation

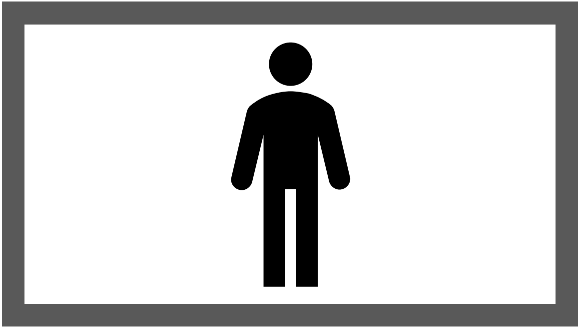
The environment emits the reward $r(s_t, a_t)$



Problem Formulation

The agent moves to the next state

$$s_{t+1} \sim P(\cdot|s_t, a_t)$$



Problem Formulation

The agent wants to maximise the expected return

$$\eta(\pi) = \mathbb{E}_{s_0 \sim d, a_{0:\infty} \sim \pi, s_{1:\infty} \sim P}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

Existing Frameworks of Policy Optimisation

- Generalised Policy Iteration (GPI)

$$\pi_{\text{new}}(\cdot|s) = \arg\max_{p \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{\text{a} \sim p}\left[Q_{\pi_{\text{old}}}(s, \text{a})\right]$$

Existing Frameworks of Policy Optimisation

- Generalised Policy Iteration (GPI)

$$\pi_{\text{new}}(\cdot|s) = \arg\max_{p \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{\mathrm{a} \sim p}\left[Q_{\pi_{\text{old}}}(s, \mathrm{a})\right]$$

Approximations: REINFORCE, A2C, DDPG.

Existing Frameworks of Policy Optimisation

- Generalised Policy Iteration (GPI)

$$\pi_{\text{new}}(\cdot|s) = \arg\max_{p \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{\mathrm{a} \sim p}\left[Q_{\pi_{\text{old}}}(s, \mathrm{a})\right]$$

Approximations: REINFORCE, A2C, DDPG.

- Trust Region Learning (TRL)

$$\pi_{\text{new}}(\cdot|s) = \arg\max_{\pi \in \Pi} \mathbb{E}_{\mathrm{s} \sim \rho_{\pi_{\text{old}}}, \mathrm{a} \sim \pi}\left[A_{\pi_{\text{old}}}(\mathrm{s}, \mathrm{a})\right] - C\mathrm{KL}_{\max}(\pi_{\text{old}}, \pi).$$

Existing Frameworks of Policy Optimisation

- Generalised Policy Iteration (GPI)

$$\pi_{\text{new}}(\cdot|s) = \arg\max_{p \in \mathcal{P}(\mathcal{A})} \mathbb{E}_{\mathrm{a} \sim p}\left[Q_{\pi_{\text{old}}}(s, \mathrm{a})\right]$$

Approximations: REINFORCE, A2C, DDPG.

- Trust Region Learning (TRL)

$$\pi_{\text{new}}(\cdot|s) = \arg\max_{\pi \in \Pi} \mathbb{E}_{\mathrm{s} \sim \rho_{\pi_{\text{old}}}, \mathrm{a} \sim \pi}\left[A_{\pi_{\text{old}}}(\mathrm{s}, \mathrm{a})\right] - C\mathrm{KL}_{\max}(\pi_{\text{old}}, \pi).$$

Loose Approximations: TRPO, PPO.

Mirror Learning

Drift

The *drift* $\mathfrak{D}_{\pi_{\text{old}}}(\pi_{\text{new}}|s)$ between two policies

Drift

The *drift* $\mathfrak{D}_{\pi_{\text{old}}}(\pi_{\text{new}}|s)$ between two policies

- Non-negative everywhere and zero at identity

$$\mathfrak{D}_{\pi_{\text{old}}}(\pi_{\text{new}}|s) \geq \mathfrak{D}_{\pi_{\text{old}}}(\pi_{\text{old}}|s) = 0$$

Drift

The *drift* $\mathfrak{D}_{\pi_{\text{old}}}(\pi_{\text{new}}|s)$ between two policies

- Non-negative everywhere and zero at identity

$$\mathfrak{D}_{\pi_{\text{old}}}(\pi_{\text{new}}|s) \geq \mathfrak{D}_{\pi_{\text{old}}}(\pi_{\text{old}}|s) = 0$$

- Zero Gâteaux derivative at identity

$$\delta_{\pi}\mathfrak{D}_{\pi_{\text{old}}}(\pi|s)|_{\pi=\pi_{\text{old}}} = 0$$

Neighbourhood Operator

The *neighbourhood* $\mathcal{N}(\pi)$ is a subset of Π that

Neighbourhood Operator

The *neighbourhood* $\mathcal{N}(\pi)$ is a subset of Π that

- Is continuous as a function of π

Neighbourhood Operator

The *neighbourhood* $\mathcal{N}(\pi)$ is a subset of Π that

- Is continuous as a function of π
- Is always compact

Neighbourhood Operator

The *neighbourhood* $\mathcal{N}(\pi)$ is a subset of Π that

- Is continuous as a function of π
- Is always compact
- It containts a closed ball for some metric

Distributions

The drift distribution $\nu_{\pi_{\text{old}}}^{\pi} \in \mathcal{P}(\mathcal{S})$

- Such that $\mathbb{E}_{s \sim \nu_{\pi_{\text{old}}}^{\pi}}[\mathfrak{D}_{\pi_{\text{old}}}(\pi|s)]$ is continuous in π.

The sampling distribution $\beta_{\pi} \in \mathcal{P}(\mathcal{S})$

- Continuous in π.

Mirror Learning

At every step, let

$$[\mathcal{M}^{\pi}_{\mathfrak{D}} V_{\pi_{\text{old}}}](s) = \mathbb{E}_{\mathrm{a}\sim\pi}[A_{\pi_{\text{old}}}(s, \mathrm{a})] - \frac{\beta_{\pi_{\text{old}}}(s)}{\nu^{\pi}_{\pi_{\text{old}}}(s)} \mathfrak{D}_{\pi_{\text{old}}}(\pi|s)$$

Mirror Learning

At every step, let

$$[\mathcal{M}_{\mathfrak{D}}^{\pi} V_{\pi_{\text{old}}}](s) = \mathbb{E}_{\mathrm{a}\sim\pi}[A_{\pi_{\text{old}}}(s, \mathrm{a})] - \frac{\beta_{\pi_{\text{old}}}(s)}{\nu_{\pi_{\text{old}}}^{\pi}(s)} \mathfrak{D}_{\pi_{\text{old}}}(\pi|s)$$

Mirror Learning updates the policy by

$$\pi_{\text{new}} = \underset{\pi\in\mathcal{N}(\pi_{\text{old}})}{\arg\max}\, \mathbb{E}_{\mathrm{s}\sim\beta_{\pi_{\text{old}}}}\left[[\mathcal{M}_{\mathfrak{D}}^{\pi} V_{\pi_{\text{old}}}](\mathrm{s})\right]$$

The Mirror Learning Theorem

Let policies $(\pi_n)_{n=0}^{\infty}$ be generated by a Mirror Learning algorithm. Then,

The Mirror Learning Theorem

Let policies $(\pi_n)_{n=0}^{\infty}$ be generated by a Mirror Learning algorithm. Then,

- They attain the monotonic improvement property,

$$\eta(\pi_{n+1}) \geq \eta(\pi_n), \ \forall n \in \mathbb{N}$$

The Mirror Learning Theorem

Let policies $(\pi_n)_{n=0}^{\infty}$ be generated by a Mirror Learning algorithm. Then,

- They attain the monotonic improvement property,

$$\eta(\pi_{n+1}) \geq \eta(\pi_n), \ \forall n \in \mathbb{N}$$

- Their value functions converge to the optimal value function,

$$V_{\pi_n} \to V^*, \text{ as } n \to \infty$$

The Mirror Learning Theorem

Let policies $(\pi_n)_{n=0}^{\infty}$ be generated by a Mirror Learning algorithm. Then,

- They attain the monotonic improvement property,

$$\eta(\pi_{n+1}) \geq \eta(\pi_n), \ \forall n \in \mathbb{N}$$

- Their value functions converge to the optimal value function,

$$V_{\pi_n} \to V^*, \text{ as } n \to \infty$$

- Their returns converge to the optimal return,

$$\eta(\pi_n) \to \eta^*, \text{ as } n \to \infty$$

The Mirror Learning Theorem

Let policies $(\pi_n)_{n=0}^{\infty}$ be generated by a Mirror Learning algorithm. Then,

- They attain the monotonic improvement property,

$$\eta(\pi_{n+1}) \geq \eta(\pi_n), \ \forall n \in \mathbb{N}$$

- Their value functions converge to the optimal value function,

$$V_{\pi_n} \rightarrow V^*, \text{ as } n \rightarrow \infty$$

- Their returns converge to the optimal return,

$$\eta(\pi_n) \rightarrow \eta^*, \text{ as } n \rightarrow \infty$$

- Their ω-limit set consists of optimal policies

Mirror Learning Instances

Existing instances of Mirror Learning include

Mirror Learning Instances

Existing instances of Mirror Learning include

- GPI

$$\mathcal{N} \equiv \Pi \qquad \mathfrak{D} \equiv 0$$

Mirror Learning Instances

Existing instances of Mirror Learning include

- GPI
- TRL

$$\mathcal{N} \equiv \Pi \qquad \mathfrak{D}_{\pi}(\bar{\pi}|s) = C\mathrm{KL}\big(\pi(\cdot|s), \bar{\pi}(\cdot|s)\big)$$

Mirror Learning Instances

Existing instances of Mirror Learning include

- GPI
- TRL
- TRPO

$$\mathcal{N}(\pi) = \left\{\bar{\pi} \in \Pi \mid \overline{\mathrm{KL}}(\pi, \bar{\pi}) \leq \delta\right\} \qquad \mathfrak{D} \equiv 0$$

Mirror Learning Instances

Existing instances of Mirror Learning include

- GPI
- TRL
- TRPO
- PPO

$$\mathcal{N} \equiv \Pi \quad \mathfrak{D}_{\pi}(\bar{\pi}|s) = \mathbb{E}_{\mathrm{a}\sim\pi}\Big[\mathrm{ReLU}\Big(\Big[\frac{\bar{\pi}(\mathrm{a}|s)}{\pi(\mathrm{a}|s)} - \mathrm{clip}\Big(\frac{\bar{\pi}(\mathrm{a}|s)}{\pi(\mathrm{a}|s)}, 1 \pm \epsilon\Big)\Big]A_{\pi}(s, \mathrm{a})\Big)\Big]$$

Mirror Learning Instances

Existing instances of Mirror Learning include

- GPI
- TRL
- TRPO
- PPO

Thus, the convergence guarantees of these algorithms follow by the Mirror Learning Theorem.

Thank you for your attention!

- *Jakub Grudzien Kuba*
- Christian Schroeder de Witt
- Jakob Foerster