

Reinforcement Learning

Our formulation of the problem based on [4] is to find $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$ (\mathcal{P} is a p-measure) such that

$$\max_{\pi} J(\pi) = \max_{\pi} \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{i=0}^{\infty} \gamma^i \underbrace{R(s_i, a_i, s_{i+1})}_{r_i} \right]$$

$$p(\tau | \pi) = p(s_0) \cdot \prod_{t=0}^{\infty} \pi(a_t | s_t) \cdot T_{a_t}(s_{t+1}, s_t)$$

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

Where τ is a trajectory of the system, which consists of the states, actions and rewards at each time step.

$$\tau_t = \left\{ s_{t+i}, a_{t+i}, r_{t+i}, s_{t+i+1} \right\}_{i=0}^{\infty} \Rightarrow J(\textcolor{red}{\pi}) = \mathbb{E} \left[R(\tau_0) \right]$$

Parameterizing the policy π with θ leads to:

Infinite dimensional	Finite dimensional
$\Pi = \{\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})\}$	$\theta \in \mathbb{R}^d$ and $\pi_{\theta} \in \Pi$
$J_{\pi} : \Pi \rightarrow \mathbb{R}$	$J_{\theta} : \mathbb{R}^d \rightarrow \mathbb{R}$
$J(\pi) = \mathbb{E}_{\tau \sim p(\tau \pi)} \left[R(\tau) \right]$	$J(\theta) = \mathbb{E}_{\tau \sim p(\tau \theta)} \left[R(\tau) \right]$
$\nabla_{\pi} J = \sum_{t=0}^{T-1} p(s_t) \cdot Q^{\pi}(\mathbf{s}, a)$	$\nabla_{\theta} \mathbb{E}_{\tau \sim p(\tau \theta)} [R(\tau)]$

$$\min_{u \in \mathcal{U}} \int_0^T L(X_t, u) dt + \Phi(X_T),$$

$$dX_t = f(X_t, u(t)) dt + \sigma dW_t, \quad X_0 = x_0$$

Relaxation of the Problem:

$$\mathcal{U}_{stoch} = \{\pi : [0, T] \rightarrow \mathcal{P}(\mathbb{R}) \mid \pi(\cdot | t) \text{ ist W-Maß}\}$$

$$\mathcal{U} \subset \mathcal{U}_{stoch} \quad u : [0, T] \rightarrow \mathbb{R} \Rightarrow \pi(\cdot | t) = \delta_{u(t)}$$

Diskretisierung: $\mathbf{u} = \{u_i\}_{i=0}^{N-1} \quad u_i \sim \pi(u_i | t_i)$

Parametrisierung: $\theta \rightarrow \pi_{\theta}(u_i | t_i)$ und $\theta \in \mathbb{R}^n$

$$\log [\pi_{\theta}(\mathbf{u})] = \log \left[\prod_{i=0}^{N-1} \pi_{\theta}(u_i | t_i) \right] = \sum_{i=0}^{N-1} \log [\pi_{\theta}(u_i | t_i)]$$

$$\int_0^T L(X_t, u_t) dt + \Phi(X_T) \approx C(\mathbf{u}) = C(u_0, \dots, u_{N-1})$$

Optimierung: Wahl spezieller Verteilung $\pi_{\theta} = \mathcal{N}(\mu(\theta, t), \sigma^2)$

$$\nabla_{\theta} \mathbb{E} [C(\mathbf{u})] = \mathbb{E} \left[\sum_{i=0}^{N-1} \nabla_{\theta} \log (\pi_{\theta}(u_i | t_i)) \cdot C(u_i, \dots, u_{N-1}) \right]$$

$$\approx \frac{1}{M} \sum_{m=1}^M \left[\sum_{i=0}^{N-1} \nabla_{\theta} \log (\pi_{\theta}(u_i^{(m)} | t_i)) \cdot C(u_i^{(m)}, \dots, u_{N-1}^{(m)}) \right]$$

Gradientenverfahren:

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_{\theta} J(\theta_k)$$

Simulation des Zustands $X_{t_{i+1}}$

$$X_{t_{i+1}} = X_{t_i} + f(X_{t_i}, u_i) \Delta t + \sigma Z_i,$$

$$Z_i \sim \mathcal{N}(0, \Delta t)$$

Projektion auf \mathcal{U} :

$$u^*(t_i) = \mu(\theta^*, t_i) \in \mathcal{U}$$

Policy Gradient Methods

Policy Gradient Theorem:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p(\tau|\theta)} \left[\sum_{t=0}^{T-1} Q^{\pi_{\theta}}(s_t, a_t) \cdot \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$$

Optimal Control Problems

Control Problem	RL Problem
$\min_u \int_0^1 u^2 dt + (x(1) - 1)^2$	$\mathbb{E}_{\pi} \sum_{t=0}^{T-1} u_t^2 + (x_T - 1)^2$
$\sum_{i=1}^N u_i^2 + (x_N - 1)^2$	$- \sum_{t=0}^{T-1} u_t^2 - (x_T - 1)^2$

Further Directions

Further improvements based on [3], [2] and [1].

References

- [1] Tuomas Haarnoja et al. "Soft Actor-Critic Algorithms and Applications". In: *arXiv preprint arXiv:1812.05905* (2018). arXiv: 1812 . 05905 [cs.LG].
- [2] Tuomas Haarnoja et al. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *Proceedings of the 35th International Conference on*

Machine Learning (ICML). 2018. arXiv: 1801.01290 [cs.LG].

- [3] Timothy P. Lillicrap et al. “Continuous control with deep reinforcement learning”. In: *arXiv preprint arXiv:1509.02971* (2015). arXiv: 1509.02971 [cs.LG].
- [4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.