Reinforcement Learning

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

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

$$p(\tau \mid \pi) = p(\boldsymbol{s}_0) \cdot \prod_{t=0}^{\infty} \pi(a_t|\boldsymbol{s}_t) \cdot P_{a_t}(\boldsymbol{s}_{t+1}, \boldsymbol{s}_t)$$

$$\boldsymbol{s}_{t+1} \sim P_{a_t}(\cdot, \boldsymbol{s}_t) \quad a_t \sim \pi(\cdot|\boldsymbol{s}_t)$$

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

$$\boldsymbol{\tau}_t = \left\{\boldsymbol{s}_{t+i}, a_{t+i}, r_{t+i}, \boldsymbol{s}_{t+i+1}\right\}_{i=0}^{\infty} \Rightarrow J(\pi) = \mathsf{E}\left[R(\boldsymbol{\tau}_0)\right]$$

Policy Gradient Methods

Let π_θ be a parameterized policy with $\theta \in \mathbb{R}^d$.

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \mathsf{E}_{\tau\sim p(\tau|\theta)}\left[R(\boldsymbol{\tau})\right] \\ &= \int_{\mathcal{T}} R(\boldsymbol{\tau}) \cdot \underbrace{\nabla_\theta \log(p(\boldsymbol{\tau} \mid \theta)) \cdot p(\boldsymbol{\tau} \mid \theta)}_{\nabla_\theta p(\boldsymbol{\tau}|\theta)}\, \mathrm{d}\boldsymbol{\tau} \\ &= \mathsf{E}_{\tau\sim p(\tau|\theta)}\left[R(\boldsymbol{\tau}) \cdot \nabla_\theta \log(p(\boldsymbol{\tau} \mid \theta))\right]\end{aligned}$$

The Logarithm applied to the distribution $p(\boldsymbol{\tau} \mid \theta)$:

$$\begin{aligned}&\nabla_\theta \log\left[p(\boldsymbol{s}_0) \cdot \prod_{k=0}^{T} \pi_\theta(a_k|\boldsymbol{s}_k) \cdot P_{a_k}(\boldsymbol{s}_{k+1}, \boldsymbol{s}_k)\right] \\ &= \sum_{k=0}^{T} \nabla_\theta \log\left[\pi_\theta(a_k|\boldsymbol{s}_k)\right]\end{aligned}$$

Let $R(\boldsymbol{\tau}) = \sum_{i=0}^{T-1} r_i + r_T$ where $r_i = R(\boldsymbol{s}_i, a_i, \boldsymbol{s}_{i+1})$

$$\begin{aligned}&\mathsf{E}_{\tau\sim p(\tau|\theta)}\left[R(\boldsymbol{\tau}) \cdot \sum_{k=0}^{T} \nabla_\theta \log\left[\pi_\theta(a_k|\boldsymbol{s}_k)\right]\right] \\ &= \sum_{i=0}^{T}\sum_{k=0}^{T} \mathsf{E}_\tau\left[r_i \cdot \nabla_\theta \log\left[\pi_\theta(a_k|\boldsymbol{s}_k)\right]\right]\end{aligned}$$

For fixed $i < k$ the law of total expectation yields:

$$\begin{aligned}&\mathsf{E}_{a_k,\boldsymbol{s}_k}\left[\mathsf{E}_\tau\left[r_i \cdot \nabla_\theta \log\left[\pi_\theta(a_k|\boldsymbol{s}_k)\right]\right] \mid a_k, \boldsymbol{s}_k\right] \\ &= \mathsf{E}_{a_k,\boldsymbol{s}_k}\left[\nabla_\theta \log\left[\pi_\theta(a_k|\boldsymbol{s}_k)\right] \cdot \underbrace{\mathsf{E}_\tau\left[r_i \mid a_k, \boldsymbol{s}_k\right]}_{\mathsf{E}_\tau[r_i|\boldsymbol{s}_k]=g(\boldsymbol{s}_k)}\right] \\ &= \mathsf{E}_{\boldsymbol{s}_k}\left[\mathsf{E}_{a_k|\boldsymbol{s}_k}\left[\left[g(\boldsymbol{s}_k) \cdot \nabla_\theta \log\left(\pi_\theta(a_k|\boldsymbol{s}_k)\right)\right] \mid \boldsymbol{s}_k\right]\right]\end{aligned}$$

Now using $g(\boldsymbol{s}_k)$ is independent of a_k and therefore

$$\begin{aligned}&\mathsf{E}_{\boldsymbol{s}_k}\left[g(\boldsymbol{s}_k)\mathsf{E}_{a_k|\boldsymbol{s}_k}\left[\left[\nabla_\theta \log\left(\pi_\theta(a_k|\boldsymbol{s}_k)\right)\right] \mid \boldsymbol{s}_k\right]\right] \\ &= \mathsf{E}_{\boldsymbol{s}_k}\left[g(\boldsymbol{s}_k) \cdot \int_{\mathcal{A}} \underbrace{\nabla_\theta \log\left(\pi_\theta(a|\boldsymbol{s}_k)\right) \cdot \pi_\theta(a|\boldsymbol{s}_k)}_{\nabla_\theta \pi_\theta(a|\boldsymbol{s}_k)}\, \mathrm{d}a\right] \\ &= \nabla_\theta \underbrace{\int_{\mathcal{A}} \pi_\theta(a|\boldsymbol{s}_k)\mathrm{d}a}_{=1} = 0\end{aligned}$$

We conclude that for $i < k$ the expectation is zero and we denote $R_t = \sum_{k=t}^{T} r_k = Q(\boldsymbol{s}_t, a_t)$ and obtainthe formula:

$$\nabla_\theta J(\theta) = \mathsf{E}_{\tau\sim p(\tau|\theta)}\left[\sum_{t=0}^{T} \nabla_\theta \log\left(\pi_\theta(a_t|\boldsymbol{s}_t)\right) \cdot Q(\boldsymbol{s}_t, a_t)\right]$$

Parameterizing the policy π with θ leads to:

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

Optimal Control Problems

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

$$\min_{u \in \mathcal{U}} \int_0^T L(X_t, u)\,\mathrm{d}t + \Phi(X_T),$$
$$\mathrm{d}X_t = f(X_t, u(t))\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad X_0 = x_0$$

Relaxation of the Problem:
$\mathcal{U}_{stoch} = \{\pi : [0,T] \to \mathcal{P}(\mathbb{R}) \mid \pi(\cdot|t) \text{ ist W-Maß}\}$
$\mathcal{U} \subset \mathcal{U}_{stoch}$ $u : [0,T] \to \mathbb{R} \Rightarrow \pi(\cdot|t) = \delta_{u(t)}$

Diskretisierung: $\boldsymbol{u} = \{u_i\}_{i=0}^{N-1}$ $u_i \sim \pi(u_i \mid t_i)$
Parametrisierung: $\theta \to \pi_\theta(u_i \mid t_i)$ und $\theta \in \mathbb{R}^n$

$$\log\left[\pi_\theta(\boldsymbol{u})\right] = \log\left[\prod_{i=0}^{N-1} \pi_\theta(u_i \mid t_i)\right] = \sum_{i=0}^{N-1} \log\left[\pi_\theta(u_i \mid t_i)\right]$$
$$\int_0^T L(X_t, u_t)\,\mathrm{d}t + \Phi(X_T) \approx C(\boldsymbol{u}) = C(u_0, \dots, u_{N-1})$$

Optimierung: Wahl spezieller Verteilung $\pi_\theta = \mathcal{N}(\mu(\theta, t), b)$
$$\nabla_\theta \mathsf{E}\left[C(\boldsymbol{u})\right] = \mathsf{E}\left[\sum_{i=0}^{N-1} \nabla_\theta \log\left(\pi_\theta(u_i \mid t_i)\right) \cdot C(u_i, \dots, u_{N-1})\right]$$
$$\approx \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{i=0}^{N-1} \nabla_\theta \log\left(\pi_\theta(u_i^{(m)} \mid t_i)\right) \cdot C(u_i^{(m)}, \dots, u_{N-1}^{(m)})\right]$$

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

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.