Reinforcement learning

Felix Crazzolara

22. Oktober 2021

1 Stochastic processes[1]

$x[n]$ is called a *discrete-state* process if its values are countable. Otherwise, it is a *continuous-state* process.

2 Dymamic programming

Dynamic programming (DP) is an algorithmic technique for solving an optimization problem by breaking it down into simpler subproblems and utilizing the fact that the optimal solution to the overall problem depends upon the optimal solution to its subproblems.

3 Policy gradient methods

3.1 Basics

The goal of reinforcement learning

$$\text{Infinite horizon:} \quad \theta^* \triangleq \arg\max_\theta \mathbb{E}_{(s,a)\sim p_\theta(s,a)}[r(s,a)]$$

$$\text{Finite horizon:} \quad \theta^* \triangleq \arg\max_\theta \sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\theta(s_t,a_t)}[r(s_t,a_t)]$$

The policy gradient (Finite horizon)

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\tau\sim p_\theta(\tau)}[\sum_{t=1}^{T} r(s_t,a_t)] = \int \nabla_\theta p_\theta(\tau) r(\tau)\,\mathrm{d}\tau \\ &= \mathbb{E}_{\tau\sim p_\theta(\tau)}[\nabla_\theta \log p_\theta(\tau) r(\tau)] \\ &= \mathbb{E}_{\tau\sim p_\theta(\tau)}[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=1}^{T} r(s_t,a_t)\right)]\end{aligned}$$

Note that the Markov property is not used here!

The policy gradient (Finite horizon with discount)

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{\tau\sim p_\theta(\tau)}[\sum_{t=1}^{T} \gamma^{t-1} r(s_t,a_t)] \\ &= \mathbb{E}_{\tau\sim p_\theta(\tau)}[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=1}^{T} \gamma^{t-1} r(s_t,a_t)\right)]\end{aligned}$$

Causality (Finite horizon)

$$\begin{aligned}\nabla_\theta J(\theta) &= \sum_{t=1}^{T} \mathbb{E}_{\substack{s_{1:t}\\ a_{1:t}}}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) \mathbb{E}_{\substack{s_{t+1:\infty}\\ a_{t+1:\infty}}}\left[\sum_{t'=1}^{T} r(s_{t'},a_{t'})\Big| s_{1:t}, a_{1:t}\right]\right] \\ &= (*)\end{aligned}$$

With

$$\mathbb{E}_{a_t|s_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)] = \int \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)\,\mathrm{d}a_t = 0,$$

it follows that

$$\begin{aligned}(*) &= \sum_{t=1}^{T} \mathbb{E}_{\substack{s_{1:t}\\ a_{1:t}}}\left[\nabla_\theta \log \pi_\theta(a_t|s_t) \mathbb{E}_{\substack{s_{t+1:\infty}\\ a_{t+1:\infty}}}\left[\sum_{t'=t}^{T} r(s_{t'},a_{t'})\Big| s_{1:t}, a_{1:t}\right]\right] \\ &= \mathbb{E}_{\substack{s_{1:\infty}\\ a_{1:\infty}}}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=t}^{T} r(s_{t'},a_{t'})\right)\right]\end{aligned}$$

Causality (Finite horizon with discount)

Similarly as before it holds that:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\substack{s_{1:\infty}\\ a_{1:\infty}}}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\left(\sum_{t'=t}^{T} \gamma^{t'-1} r(s_{t'},a_{t'})\right)\right]$$

The off-policy policy gradient

$$\begin{aligned}\nabla_{\theta'} J(\theta') = \mathbb{E}_{\tau\sim p_\theta(\tau)}\Bigg[&\sum_{t=1}^{T} \nabla_{\theta'} \log \pi_{\theta'}(a_t|s_t)\left(\prod_{t'=1}^{t} \frac{\pi_{\theta'}(a_{t'}|s_{t'})}{\pi_\theta(a_{t'}|s_{t'})}\right) \\ &\left(\sum_{t'=t}^{T} r(s_{t'},a_{t'})\left(\prod_{t''=t}^{t'} \frac{\pi_{\theta'}(a_{t''}|s_{t''})}{\pi_\theta(a_{t''}|s_{t''})}\right)\right)\Bigg]\end{aligned}$$

4 Papers

4.1 Introduction to Stochastic Dynamic Programming (1983)

Consider a countable state set $\mathcal{S}$ and a finite action set $\mathcal{A}$. The states are are labelled by nonnegative integers and the transition probabilities are given by $P_{ij}(a)$ for $a \in \mathcal{A}$. Furthermore, upon choosing an action, a reward $|R(i,a)| < B$ is credited to the agent. The value function is defined as

$$V_\pi(i) = \mathbb{E}_\pi\left[\sum_{n=0}^{\infty} R(X_n,a_n)\alpha^n | X_0 = i\right].$$

Let $V(i) = \sup_\pi V_\pi(i)$. A policy π^* is said to be α-optimal if $V_{\pi^*}(i) = V(i)$ for all $i \geq 0$. the optimal value function is satisfies an optimality equation:

Theorem 2.1 (The Optimality Equation):

$$V(i) = \max_a[R(i,a) + \alpha \sum_j P_{ij}(a) V(j)], \quad i \geq 0.$$

Finally, the next theorem establishes the existance of an optimal policy:

Theorem 2.2: Let f be a mapping $f : \mathcal{S} \to \mathcal{A}$ which satisfies

$$R(i,f(i)) + \alpha \sum_j P_{ij}(f(i)) V(j) = \max_a[R(i,a) + \alpha \sum_j P_{ij}(a) V(j)], \quad i \geq 0.$$

Then $V_f(i) = V(i)$ for all $i \geq 0$, and hence f is α-optimal.

4.2 Learning Without State-Estimation in Partially Observable Markovian Decision Processes

Points out a few very interesting issues that can arise when applying RL methods to POMDPs. Hello **??**

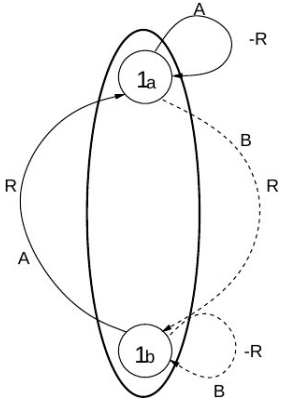

Abbildung 1: POMDP with two states (1a&1b), two actions (A&B), but only one possible observation.

Issue 1: Any policy will incur a cost of R per step, whereas the optimal behavior results in a reward of R per step. *Issue 2:* The optimal policy for the POMDP is non-stationary.

4.3 An Analysis of Temporal-Difference Learning with Function Approximation

Assumption 1: 1) The Markov chain i_t is irreducible and aperiodic. Furthermore, there is a unique distribution π that satisfies $\pi' P = \pi'$ with $\pi(i) > 0$ for all i; here, π is a finite of infinite vector, depending on the cardinality of S. Let $E_0[\cdot]$ stand for expectation with respect to this distribution.
2) Transition costs $g(i_t, i_{t+1})$ satisfy $E_0[g^2(i_t, i_{t+1})] < \infty$ *Assumption 2:* 1) the matrix Φ has full column rank; that is, the basis functions $\{\phi_k | k = 1, \dots, K\}$ are linearly independent.
2) For every k, the basis function ϕ_k satisfies $E_0[\phi_k^2(i_t)] < \infty$. *Assumption 3:* There exists a function $f : S \to \mathbb{R}_+$ satisfying the following requirements:
1) For all i_0 and $m \geq 0$

$$\sum_{\tau=0}^{\infty} ||E[\phi(i_\tau)\phi'(i_{\tau+m})|i_0] - E_0[\phi(i_t)\phi'(i_{t+m})]|| \leq f(i_0)$$

and

$$\sum_{\tau=0}^{\infty} ||E[\phi(i_\tau)g(i_{\tau+m}, i_{\tau+m+1})|i_0] - E_0[\phi(i_t)g(i_{t+m}, i_{t+m+1})]|| \leq f(i_0).$$

2) For any $q > 1$, there exists a constant μ_q such that for all i_0, t $E[f^q(i_t)|i_0] \leq \mu_q f^q(i_0)$. *Assumption 4:* The step sizes γ_t are positive, nonincreasing, and predetermined. Furthermore, they satisfy $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$.

Theorem 1: Under assumption 1-4, the following hold:
1) The cost-to-go function J^* is in $L_2(S, D)$.
2) For any $\lambda \in [0, 1]$, the TD(λ) algorithm with linear function approximators converges with probability one.
3) The limit of convergence r^* is the unique solution of the equation $\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*$.
4) Furthermore, r^* satisfies

$$||\Phi r^* - J^*||_D \leq \frac{1 - \lambda\alpha}{1 - \alpha} |\Pi J^* - J^*|_D.$$

4.4 Policy Gradient Methods for Reinforcement Learning with Function Approximation

Establishes the policy gradient theorem and shows convergence of policy gradient algorithm for arbitrary policy classes.

5 Tipps and tricks

5.1 Value function normalization

Consider a reward function $R : \mathcal{S} \times \mathcal{A} \to [-R, R]$. If we define the normalized value function as $V_\pi^n(s) \triangleq (1 - \gamma)V_\pi(s)$, where $V_\pi(s) \triangleq \mathbb{E}_{\substack{s_{0:\infty} \\ a_{0:\infty}}}[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)]$ is the usual value function and $\gamma \in (0, 1)$ is the discount factor, then we are guarantueed that $V_\pi^n(s) \in [-R, R]$.

Literatur

[1] Athanasios Papoulis. Probability, random variables and stochastic processes. 1965.