Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Final Exam | 1404/07/13 | 180 Minutes

---

1 Value-Based Theory

1.1 Bellman Equation for the Q-Function

1.1.1 Part 1

Assume a stochastic reward function:

$$\Pr(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a), \quad \forall s, s' \in \mathcal{S},\ r \in \mathcal{R},\ a \in \mathcal{A}.$$

This is abbreviated as:

$$p(s', r \mid s, a).$$

The optimal action-value function $q^*(s, a)$ is defined as:

$$q^*(s, a) = \max_{\pi}\ \mathbb{E}[G_t \mid S_t = s, A_t = a].$$

Define the return G_t as:

$$G_t = \sum_{t'=t+1}^{\infty} \gamma^{t'-t-1} R_{t'}.$$

Prove that we can reach the following expression for the optimal action-value function:

$$q^*(s, a) = \sum_r r \sum_{s'} p(s', r \mid s, a) + \gamma \max_{\pi} \sum_{s'} p(s' \mid s, a) \sum_{a'} \pi(a' \mid s')\, q_\pi(s', a').$$

1.1.2 Part 2

Show that we can equivalently write the above equation as:

$$q^*(s, a) = \sum_r r \sum_{s'} p(s', r \mid s, a) + \gamma \max_{\pi} \sum_{s'} p(s' \mid s, a) \max_{a'} q_\pi(s', a').$$

and reach the following expression:

$$q^*(s, a) = \sum_{r, s'} p(s', r \mid s, a) \left[r + \gamma \max_{a'} q^*(s', a') \right].$$

1.2 Existence, Uniqueness, and Fixed-Point Property of the Bellman Equation for the Q-Function

Definition 1 (Contraction Mapping). Let (X, d) be a metric space. A map $T : X \to X$ is called a *contraction* if there exists a constant $c \in [0, 1)$ such that for all $x, y \in X$,

$$d(T(x), T(y)) \leq c\, d(x, y).$$

Definition 2 (Cauchy Sequence). A sequence $\{x_n\}$ in a metric space (X, d) is called a *Cauchy sequence* if for every $\epsilon > 0$ there exists an integer $N \geq 1$ such that

$$d(x_n, x_m) < \epsilon \quad \text{for all } n, m \geq N.$$

Definition 3 (Fixed Point). Let X be a set and $T : X \to X$ a mapping. A point $x^* \in X$ is called a *fixed point* of T if

$$T(x^*) = x^*.$$

1.2.1 Part 1

Show that if T is a contraction mapping and

$$Tx_{k-1} = x_k \quad \text{for } k = 1, 2, \ldots,$$

then the sequence $\{x_n\}$ is a Cauchy sequence, and determine an appropriate N such that

$$d(x_n, x_m) < \epsilon \quad \text{for all } n, m \geq N.$$

1.2.2 Part 2

Since T has the Cauchy sequence property, the sequence $\{x_n\}$ converges. Show that the limit of this sequence is a fixed point of T and that this fixed point is unique.

1.2.3 Part 3

Definition 4 (Bellman Optimality Operator). Define the operator T as

$$Tq(s, a) = \sum_{r, s'} p(r, s' \mid s, a) \left[r + \gamma \max_{a'} q(s', a') \right].$$

Prove for a finite Markov Decision Process (MDP), the mapping T defined above is a contraction mapping.

2 Advanced Theory

2.1 The Explore-Then-Commit (ETC) Algorithm and Its Regret Bound

This ETC algorithm is straightforward: it first explores by playing each arm a fixed number of times and then exploits by committing to the arm that appeared best during exploration. The ETC algorithm is parameterized by the number of exploration steps per arm, denoted by a natural number m. Given k possible actions, the algorithm performs $m \cdot k$ exploration rounds before selecting the best arm to commit to for the remaining rounds.
The pseudocode of this method is illustrated in the figure below:

Input m.
In round t choose action

$$A_t = \begin{cases} (t \bmod k) + 1\,, & \text{if } t \le mk\,; \\ \operatorname{argmax}_i \hat{\mu}_i(mk)\,, & t > mk\,. \end{cases}$$

(ties in the argmax are broken arbitrarily)

Figure 1: ETC Algorithm

Definition 1 (Suboptimality Gap). For a k-armed bandit problem, let μ_i denote the expected reward of arm i, and let $\mu^* = \max_{1 \le i \le k} \mu_i$ be the expected reward of the optimal arm. The *suboptimality gap* of arm i is defined as

$$\Delta_i = \mu^* - \mu_i.$$

Definition 2 (Subgaussianity). A random variable X with mean μ is called σ^2-subgaussian if for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\big[\exp(\lambda(X - \mu))\big] \le \exp\Big(\frac{\lambda^2 \sigma^2}{2}\Big).$$

Prove when the ETC algorithm interacts with any 1-subgaussian bandit and $1 \le m \le n/k$, the cumulative regret R_n satisfies:

$$R_n \le mk \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i \exp\left(- \frac{m\Delta_i^2}{4} \right),$$

where Δ_i denotes the suboptimality gap of arm i.

3 Exploration Methods

3.1 UCB and Sub-Logarithmic Regret

Consider a stochastic K-armed bandit with rewards in $[0, 1]$ and unknown means $\mu_1, \dots, \mu_K$. Let $\mu^\star = \max_i \mu_i$ and $\Delta_i = \mu^\star - \mu_i$. The *UCB1* algorithm selects arm

$$A_t \;=\; \arg\max_i \Big[\widehat{\mu}_i(t-1) \;+\; c\sqrt{\tfrac{\ln t}{N_i(t-1)}}\Big],$$

where $\widehat{\mu}_i(t-1)$ is the empirical mean and $N_i(t-1)$ the pull count for arm i.

(a) Using Hoeffding's inequality, derive the confidence bonus $c\sqrt{\ln t / N_i(t)}$ (choose c so the failure probability is at most t^{-2}). Explain how "optimism in the face of uncertainty" balances exploration and exploitation.

(b) Prove that, for any sub-optimal arm i,

$$\mathbb{E}\big[N_i(T)\big] \;\le\; \frac{8\,\ln T}{\Delta_i^2} + 1 \quad \text{for UCB1 with } c = \sqrt{2}.$$

Outline the major steps (high-probability concentration, stopping-time argument, union bound).

(c) Using part (b), derive

$$\mathbb{E}\big[\mathrm{Reg}(T)\big] \;=\; \mathcal{O}\big(K\,\ln T\big).$$

Explain why this is a *no-regret* guarantee and compare it to the minimax lower bound of $\Theta\big(\sqrt{T}\big)$.

4 Imitation & Inverse RL

4.1 Feature Matching

Question 1: What is the difference between *Forward Reinforcement Learning* and *Inverse Reinforcement Learning (IRL)*?

Question 2: What is *Feature Matching* in IRL? Write down its mathematical formulation, explain its relationship to the reward function, and discuss the main limitations or problems associated with this approach.

5 Offline RL

5.1 Implicit Policy Constraints (AWR)

Consider the offline RL setup and the Advantage-Weighted Regression (AWR) algorithm. The constrained optimisation problem is

$$\max_{\pi} J(\pi) \quad \text{s.t.} \quad \mathbb{E}_{s\sim d_\beta}\left[D_{\mathrm{KL}}(\pi(\cdot|s)\|\beta(\cdot|s))\right] \le \epsilon,$$

with the corresponding Lagrangian

$$\mathcal{L}(\pi,\eta) = J(\pi) + \eta\Big(\epsilon - \mathbb{E}_{s\sim d_\beta}[D_{\mathrm{KL}}(\pi(\cdot|s)\|\beta(\cdot|s))]\Big),$$

and let the advantage function be $A_\beta(s,a) = Q_\beta(s,a) - V_\beta(s)$.

5.1.1 Part 1

Derive the closed-form expression for the optimal policy $\pi^\star(a|s)$ by taking the functional derivative of $\mathcal{L}(\pi,\eta)$ with respect to $\pi(a|s)$. Show clearly how the exponential advantage term appears, and define α in terms of the Lagrange multiplier η.

5.1.2 Part 2

Show that the resulting AWR policy update reduces to a weighted behavior cloning problem of the form

$$\max_{\pi} \mathbb{E}_{(s,a)\sim D}[w(s,a)\log\pi(a|s)],$$

and derive the explicit form of the weights $w(s,a)$. Discuss the effect of α on these

6 Multi-Agent RL

Show that this game has no pure Nash equilibrium:

		Player 2	
		Heads	Tails
Player 1	Heads	1,-1	-1,1
	Tails	-1,1	1,-1

7 Meta RL

Generic meta learning could be described mathematically as follows:
"Generic" learning:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta, D^{tr}) = f_{\text{learn}}(D^{tr})$$

"Generic" meta-learning:

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \mathcal{L}(\phi_i, D_i^{\text{ts}}), \quad \text{where } \phi_i = f_\theta(D_i^{\text{tr}})$$

(a) How to extend this formulation to the meta reinforcement learning? Describe the math precisely.
(b) Based on your explanation in part (a), explain why "exploration" plays an important role in identifying ϕ_i.