

Homework 1 – Theory Problems

Problem 1

In this problem we analyze the impact of two basic reward transformations on optimal policies.

Let $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mu, r, T, \gamma)$ be a finite MDP: \mathcal{X} and \mathcal{A} are finite space and action sets, $\mu \in \Delta(\mathcal{X})$ is the initial state distribution, $r: \mathcal{X} \times \mathcal{A} \rightarrow [-R, R]$ is a bounded reward function, $T(\cdot | x, a) \in \Delta(\mathcal{X})$ is the transition kernel, and $\gamma \in [0, 1)$ is the discount factor. The goal is to find a policy $\pi: \mathcal{X} \rightarrow \mathcal{A}$ that maximizes

$$J(\pi) := \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \right].$$

Let π^* be an optimal policy for \mathcal{M} . For each of the following reward function modifications, define a new MDP $\mathcal{M}' = (\mathcal{X}, \mathcal{A}, \mu, r', T, \gamma)$, and determine whether π^* must remain optimal in \mathcal{M}' . Prove your claims or give a counterexample.

- (i) $r'(x, a) = r(x, a) + \alpha$, where $\alpha \in \mathbb{R}$ is a constant.
- (ii) $r'(x, a) = \beta r(x, a)$, where $\beta \in \mathbb{R}$ is a constant.

Problem 2

In this problem we study a fundamental issue in sampling for Q-learning.

- (i) Suppose $X_a = x_a + \varepsilon_a$, where $a \in A$ for a finite set A , $x_a \in \mathbb{R}$, and $\{\varepsilon_a\}_{a \in A}$ are independent random variables with zero mean. Prove that

$$\mathbb{E} \left[\max_a X_a \right] \geq \max_a \mathbb{E}[X_a].$$

The inequality says that the max of unbiased noisy estimates is upward biased.

- (ii) Suppose $A = \{1, 2\}$, $x_1 = x_2 =: x$ and $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Compute the two terms $\mathbb{E}[\max_a X_a]$ and $\max_a \mathbb{E}[X_a]$.
- (iii) For $n \in \mathbb{N}$, let $A_n = \{1, 2, \dots, n\}$. Plot $f(n) = \mathbb{E}[\max_{a \in A_n} X_a]$ and $g(n) = \max_{a \in A_n} \mathbb{E}[X_a]$ as a function of n as n varies from 1 to 10^6 . Assume i.i.d. $\mathcal{N}(0, 1)$ noise and $x_i = 1$.
- (iv) Read the [double Q-learning paper](#) and explain the main idea behind the proposed algorithm.