6.7920 Fall 2025: Homework 4

Note: If you cannot do a part of a problem, you can assume the result of that part and proceed to the next one.

1 A Policy Iteration Variant

Consider a finite-state (with n states), finite-action, discounted infinite-horizon problem with discount factor γ and the following algorithm:

- Let V_0 be an arbitrary n-dimensional vector.
- The algorithm generates a sequence of vectors $V_1, V_2, \ldots$ and stationary policies $\pi_1, \pi_2, \ldots$.
- Each policy π_t is chosen to satisfy
$$\mathcal{T}_{\pi_t} V_t = \mathcal{T} V_t.$$
- The next vector V_{t+1} is computed according to
$$V_{t+1} = \mathcal{T}_{\pi_t}^2 V_t = \mathcal{T}_{\pi_t}(\mathcal{T}_{\pi_t} V_t).$$

Let's name $\mathcal{G}$ as the operator that performs this iteration, that is, $\mathcal{G}(V_t) := \mathcal{T}_{\pi_t}^2 V_t$. Based on proposed procedure, answer the following questions

1. Suppose $\mathcal{T} V_0 \geq V_0$. Show that $V_{t+1} \geq \mathcal{T} V_t$ for all t.
2. Suppose $\mathcal{T} V_0 \geq V_0$. Show that $\lim_{t\to\infty} V_t = V^*$.
3. For any given V_0, explain how you can choose a scalar d so that $\mathcal{T}\overline{V}_0 \geq \overline{V}_0$, where
$$\overline{V}_0(i) = V_0(i) + d, \forall\ i = 1, \ldots, n.$$
4. Show that
$$\lim_{t\to\infty} V_t = V^*$$
no matter how V_0 is chosen.

5. Suppose that the algorithm is stopped after a finite number of iterations and yields a policy π that satisfies (for some $\delta > 0$)

$$\mathcal{T}_\pi V_\pi(i) \geq \mathcal{T} V_\pi(i) - \delta, \forall\, i = 1, \ldots, n.$$

Show that

$$V^*(i) - V_\pi(i) \leq \frac{\delta}{1-\gamma}, \forall\, i = 1, \ldots, n.$$

2 Incremental Monte-Carlo as a TD Method

In this problem, we will show that the incremental Monte-Carlo method can be viewed as a TD method. The incremental Monte-Carlo method updates the estimated value function $\hat{V}^\pi$ after observing the $(n+1)$-th episode as follows:

$$\hat{V}^\pi_{n+1}(s_0) = (1 - \eta_{n+1})\hat{V}^\pi_n(s_0) + \eta_{n+1}\hat{R}_{n+1}(s_0),$$

where $\hat{R}_i(s_0)$ denotes the cumulative discounted reward obtained in episode i starting from s_0:

$$\hat{R}_i(s_0) := \sum_{t=0}^{T_i} \gamma^t r_{t,i}.$$

The temporal difference (TD) error at time t in episode i is defined as:

$$\delta_{t,i} := r_{t,i} + \gamma \hat{V}^\pi_{i-1}(s_{t+1,i}) - \hat{V}^\pi_{i-1}(s_{t,i})$$

Show that the incremental Monte-Carlo method update can be expressed as follows:

$$\hat{V}^\pi_{n+1}(s_0) = \hat{V}^\pi_n(s_0) + \eta_{n+1} \sum_{t=0}^{T_{n+1}} \gamma^t \delta_{t,n+1}.$$

3 Stepsize Conditions

Stochastic approximation algorithms often involve stepsize conditions of the form $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. This exercise is meant to give some insights into the role played by these two conditions.

Let w_t be independent random variables, with mean x^* and variance uniformly bounded by $b > 0$ (for all t). Consider the algorithm

$$x_{t+1} = x_t + \eta_t(w_t - x_t).$$

If the algorithm converges to some x, in the limit we should have $\mathbb{E}[w_t - x_t] \to 0$, or $x_t = \mathbb{E}[w_t] = x^*$. The question is whether we actually get this type of convergence, in the presence of noise. To simplify the analysis, let us assume that $x^* = 0$.

We assume that the stepsizes satisfy the given conditions and $\eta_t \in (0, 1)$ for all t.

1. Show that $\prod_{t=0}^{\infty}(1-\eta_t) = 0$.
2. Let us split the time axis into segments. The ith segment starts at some t_i and ends at t_{i+1}. We choose the segment lengths so that $\prod_{t=t_i}^{t_{i+1}-1}(1-\eta_t) \leq 1/2$.

   Show that $|\mathbb{E}[x_{t_{i+1}}]| \leq |\mathbb{E}[x_{t_i}]|/2$. (Thus, $\mathbb{E}[x_{t_i}]$ converges to zero.)
3. We now want to look at the variance of x_t, at least along times of the form $t = t_i$. Let $v_i = \mathrm{Var}(x_{t_i})$.

   Show that
   $$v_{i+1} \leq \frac{1}{4}v_i + \epsilon_i,$$
   where $\epsilon_i \to 0$.

Note: From part 3, we obtain that the variance of x_t, for times of the form $t = t_i$, converges to zero. A similar argument also shows that the variance goes to zero for general times t. Besides the convergence of the variance to zero, it is also true that x_t converges to x^*, *with probability 1*. However, this latter statement requires more sophisticated mathematical machinery (the supermartingale convergence theorem).

4 Computational Problem: TD(λ) in Inventory Control

In this problem, we will study policy evaluation methods in a discounted inventory control setting with backlogs and ordering costs. The goal is to compare Monte Carlo evaluation with TD(λ) and to explore the sensitivity of performance to the parameter λ. We will consider a fixed base-stock policy and examine the bias–variance tradeoff across λ.

At each period $t = 0, 1, \ldots, T$, the state $s_t \in \mathbb{Z}$ represents the inventory level at the start of the period (negative values correspond to backorders). The action $a_t \in \{0, 1, \ldots, 20\}$ represents the order quantity placed at the beginning of the period. Demand D_t is i.i.d. uniform on $\{0, 1, \ldots, 10\}$. The holding cost, backlog cost, and ordering cost are denoted as h, b, and o, respectively. The inventory evolves as
$$s_{t+1} = s_t + a_t - D_t.$$
The per-period cost for $t < T$ is
$$c_t(s_t, a_t, D_t) = o\, a_t + \max(h\, s_{t+1},\, -b\, s_{t+1}),$$
and the terminal cost at $t = T$ is
$$c_T(s) = \max(h\, s, -b\, s).$$
We define the reward as the negative cost, $r_t = -c_t$, and include a discount factor $\gamma \in (0, 1]$.

We evaluate the fixed *base-stock policy* π_B with target level B. That is, order up to level B if inventory is below B, subject to the maximum order cap. The value function under π is

$$V_t^\pi(s) = \mathbb{E}_\pi\left[\sum_{\tau=t}^{T-1} \gamma^{\tau-t} r_{\tau+1} \;+\; \gamma^{T-t} r_{T+1} \;\middle|\; s_t = s\right].$$

Our goal is to compute and compare estimates of $V_0^\pi(s)$ without having explicit access to the dynamics model and probabilities.

For the rest of the problem, we will use $B = 3$, $(o, h, b) = (1, 4, 2)$, $\gamma = 0.95$ and an episode length of $T = 100$, where the terminal reward is collected at the final timestep. One "episode" corresponds to a trajectory $(s_0, a_0, r_0, \cdots, s_{t-1}, a_{t-1}, r_{t-1}, s_T)$. Let $\mathcal{S} = \{-10, -5, 0, 5, 10\}$, $\mathcal{L} = \{0, 0.3, 0.6, 0.8, 0.9, 1.0\}$, $\mathcal{LR} = \{0.001, 0.01, 0.1\}$. For V^π, feel free to only keep track of states in $[-10, 10]$.

1. Implement a simulator for the system under π_S. Each step should generate (s_t, a_t, r_t, s_{t+1}). To verify correctness, use start states $s \in \mathcal{S}$ and simulate 500 episodes. Report the average per-episode reward (i.e. Monte-Carlo approximation) for each start state. Which start state has the highest per-episode reward?

2. Implement the TD(0) update with constant step size $\alpha \in \mathcal{LR}$. Run 500 episodes with start states sampled uniformly from $\mathcal{S}$. Track the running estimates $\hat{V}_0^\pi(s)$ for all states and plot a separate figure for each $s \in \mathcal{S}$ showing 3 learning curves corresponding to each α. Comment on the convergence of the different learning rates. Why is the convergence of V^π different for various states?

3. Implement TD(λ) using *eligibility traces* with $\lambda \in \mathcal{L}$. Select the best α (briefly explain what is your criterion) and for each λ, run $10{,}000$ episodes with start states randomly sampled from $\mathcal{S}$ to estimate $\hat{V}_0^\pi(s)$. Plot $\hat{V}_0^\pi(0)$ vs number of TD updates, showing 6 learning curves corresponding to each λ. Comment on the convergence behavior of different λs.

4. Compute a reference solution $V_0^\pi(s)$ using backward dynamic programming with discount factor γ (you may use code from previous problem sets). For each $\lambda \in \mathcal{L}$, use your results from part 3 and compute the MSE across all $s \in \mathcal{S}$ between $V_0^\pi(s)$ and $\hat{V}_0^\pi(s)$ and plot MSE versus λ. Discuss the observed tradeoff.