CS 598 Stats RL Homework 2

September 7, 2020

- **Submission deadline: Sep. 23 before class.** Submission website: compass2g.
- **Your submission must be typeset with LaTeX.** Handwritten solutions will not be accepted. The LaTeX source file of this homework will be provided and you can use it as a template.
- This homework helps you familiarize with RL basics. It is very important that you work them out and clean up any confusion in basics before we move on to more advanced topics, as we will be using these techniques through the entire course.
- All homeworks will be considered in the final grade in a way similar to "participation scores", i.e., as long as you spend effort on the problems, not getting all the answers correct will not affect your grades. On a related note, homeworks *may not* be graded in a timely manner but we will release the solutions after the deadline.
- You can discuss with anyone and consult any material, but (1) you still need to write the homework on your own, and (2) if you get help from anyone other than the course instructors or any material other than the course notes, you will need to mention them in your homework.

1. Evaluation error to decision loss

(1) There are K items, $1, 2, \ldots, K$. The i-th item has value $v_i \in \mathbb{R}$. Let $v^\star := \max_{i\in[K]} v_i$, where $[K] := \{1, 2, \ldots, K\}$, and $i^\star = \arg\max_{i\in[K]} v_i$.

An agent chooses the j-th item, where $j = \arg\max_{i\in[K]} u_i$, and $\{u_i\}_{i=1}^K$ are K real numbers. Let

$$\epsilon := \max_{i\in[K]} |u_i - v_i|.$$

Upper-bound $v^\star - v_j$ as a function of ϵ. Prove your result.

(2) If we further have $\forall i,\ u_i \le v_i$, can you improve the bound? What about $\forall i,\ u_i \ge v_i$?

(3) State and prove an upper bound of $|u_j - v^\star|$.

(4) (Optional) Where in note1.pdf have we already used this result? Point out the theorem and the quantities that play the roles of u_i and v_i.

2. Policy-specific Bellman update operator

Recall that the Bellman update operator for policy π, $\mathcal{T}^\pi$, is defined as: $\forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,

$$(\mathcal{T}^\pi f)(s,a) := R(s,a) + \gamma \mathbb{E}_{s'\sim P(s,a)}[f(s', \pi(s'))].$$

1. Prove that similar to $\mathcal{T}$, $\mathcal{T}^\pi$ is also a γ-contraction under ℓ_∞.
2. Similar to value iteration, you can use an iterative method to approximately solve for Q^π by $f_0 := \mathbf{0}$, and $f_k := \mathcal{T}^\pi f_{k-1}$. How many iterations do you need so that $\|f_k - Q^\pi\|_\infty \le \epsilon$? Express it as a function of γ and $V_{\max} := R_{\max}/(1-\gamma)$.
3. Show that $\frac{\ln(V_{\max}/\epsilon)}{1-\gamma}$ is an upper bound on the expression in Question 2. (Hint: $(1-x)^{1/x} \to e^{-1}$.)

3. The use of non-stationary policy in VI

Recall that value iteration starts with $Q^{\star,0} := \mathbf{0}$ and keeps applying $\mathcal{T}$: $Q^{\star,h} := \mathcal{T}Q^{\star,h-1}$. Let $\pi_h := \pi_{Q^{\star,h}}$, that is, the greedy policy of $Q^{\star,h}$. In the lecture, we have hinted that instead of outputting the stationary π_H as the final policy, it can be beneficial to use the nonstationary policy $\pi_{H:1} := \pi_H \circ \pi_{H-1} \circ \ldots \circ \pi_1$. That is, $\pi_{H:1}(s_t) := \pi_{H-t+1}(s_t)$ for $t \le H$, and takes arbitrary actions when $t > H$. Prove a bound on $\|V^\star - V^{\pi_{H:1}}\|_\infty$, and compare it to the bound on $\|V^\star - V^{\pi_H}\|_\infty$ (the latter is a corollary of the results in note1.pdf).

Hint: you may directly use the fact that $\pi_{H:1}$ is the optimal policy for the objective $V^{\pi,H}$, i.e., $V^{\pi_{H:1},H}(s) = V^{\star,H}(s) := \max_\pi \mathbb{E}[\sum_{t=1}^H \gamma^{t-1} r_t | s_1 = s, \pi]$, where $\max_\pi$ maximizes over all policies, including non-stationary ones.

4. Loss of using a smaller γ

Suppose we are given an MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$. In the lecture we have seen that heavy discounting leads to faster convergence of planning, so we may run some planning algorithm using $\gamma' < \gamma$. The question is, how lossy is the obtained policy when we evaluate it in the original MDP?

More concretely, let's ignore the details of the planning algorithm and say we can compute the optimal policy for $M' = (\mathcal{S}, \mathcal{A}, P, R, \gamma')$, that is, a new MDP that is the same as M in all parameters except that it has a discount factor γ'. Let $\pi^\star_{\gamma'}$ denote the optimal policy of M'. Prove a bound on $\|V^\star_M - V_M^{\pi^\star_{\gamma'}}\|_\infty$. Here the subscript M is not necessary and only used to emphasize that both value functions are defined in the original M (i.e., w.r.t. γ, not γ').

Your bound should scale with $\gamma - \gamma'$, that is, when γ' is close to γ, the loss will be small.