6.7950 Fall 2022: - Recitation 6 Handout

1 Recap and contextualization

In the previous recitation, we used the supermartingale theorem to prove the convergence of a simplified version of value iteration. In reality, the derivation presented was more general in order to better integrate with the contents of the lectures. As you may recall, the previous lecture mentioned that the max-norm contraction of Q-learning relies on bounding the error with two terms: one which is deterministic, D_k, and one which is stochastic: W_t.

The deterministic part is centered around the idea that $\mathcal{T}$ is a γ-contraction on the L_∞ norm. If we ignore the learning rate η_t for a moment, consider some value x_t that we want to to converge to $x^* = 0$. Then after applying an operator with such contractive properties, the largest element will be shrunken by a factor γ. Since there may be many components to x (let's say there are n of them) of them, then in order to be sure that all of them were shrunken by γ one would need n iterations of the algorithm. So the deterministic upper-bound D_k would shrink by factor of γ every n steps and thus $t = nk$.

When we consider that effect of the learning steps η_t, they change the rate of this contraction operation, so the number of iterations required to guarantee shrinkage all components of x may change, but the same dynamics will be preserved because η_t is non negative and $\sum \eta_t = \infty$. In other words, even if the steps are small, when taking a sufficient number of them their contribution sums to at least 1 and thus guaranteeing that the bound eventually decreases. Predicting when this decrease happens might not be easy, but it's not necessary.

The stochastic part W_t is the one where our study of supermartingales can be applied. Recall that we proved that convergence under certain assumptions for an iteration of the form

$$x_{t+1} = x_t + \eta_t g(x_t, w_t)$$

and the noise part of the upper bound followed

$$W_{t+1}(s) = (1 - \eta_t)W_t(s) + \eta_t w_t(s) = W_t(s) + \eta_t(-W_t(s) + w_t(s))$$

If we take the update rule $g(W_t, w_t) = W_t + w$, then the convergence result can apply for $x_t = W_t$ and the loss $f(W_t) = \|W_t\|_2^2$. The pseudo gradient property is satisfied since $W^* = 0$ and

$$(W^* - W_t)^\top \mathbb{E}_w[-W_t + w|\mathcal{F}_t] = W_t^\top W_t = \|W_t\|_2^2 = 1\|W^* - W_t\|_2^2$$

while the stepsizes η_t satisfy the Robbins-Monro condition by construction. Since it was shown in lecture that the variance of w_t is bounded, then we have that

$$\mathbb{E}_w[\|-W_t + w_t\|_2^2 \,|\mathcal{F}_t] \leq Var(w_t) + \|W_t\|_2^2$$

Thus, all the required conditions are satisfied and the result applies, that is, $W_t \to 0$ and the stochastic part of the upper bounded indeed does not interfere with the deterministic one.

2 Approximate Value Iteration

In the next lectures, we are going to explore the topic of approximate methods for dynamic programming and reinforcement learning. So for this part of the recitation, we are going to study some results related approximate value iteration that may not be covered in lecture and can serve as an introduction and motivation for the next topic and as an illustration between tabular and non-tabular methods.

2.1 Algorithm

In the standard value iteration, we start with an initial value function $V_0(s)$ estimate and repeatedly apply the optimal Bellman operator

$$V_{t+1} = \mathcal{T}V_t$$

Approximate value iteration, as the name implies, modifies this procedure by introducing an approximation operator $\mathcal{A}$. For example, we can consider a class of functions (e.g., neural networks, polynomials, radial basis functions) that have a function space $\mathcal{F}$. In other words, $\mathcal{F}$ consists of all the values that can be represented in this approximation. Thus, we can typically represent $\mathcal{A}$ as

$$\mathcal{A}(x) = \arg\inf_{x \in \mathcal{F}}(\|x\|)$$

for some norm function, which we will soon learn is an important consideration for the method. Thus, the whole approximate value iteration algorithm becomes

$$V_{t+1} = \mathcal{A}\mathcal{T}V_t = \arg\inf_{V \in \mathcal{F}}(\|\mathcal{T}V_t - V\|)$$

2.2 Motivational example

Let's consider an interesting example by Tsitsiklis and Van Roy adapted from Reinforcement Learning: An Introduction and displayed in Figure 1

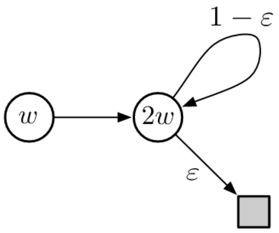

Figure 1: MRP and approximation used in the example.

In this problem, the value of the leftmost state is approximated as w, while the value for the other is approximated with the same parameter w as $2w$. All rewards are zero and we can see that the parameter choice $w = 0$ exactly captures the true value of the process. The transition is deterministic from the first state to the second and from there it can move to terminal state with probability ϵ or remain with probability $1 - \epsilon$. In other words, our approximation can, in principle, learn the true value function. Consider that the we initialize $w_0 \neq 0$.

1. Show that approximate value iteration with an approximation with respect to the L_2 norm can diverge.
2. Show that approximate value iteration with an approximation with respect to the L_∞ norm will converge.

Solution:

1. Let's find w_{t+1} that minimizes the desired norm given the approximate value function V_t generated from w_t

$$V_{t+1} = \arg\min_{V \in \mathcal{F}} \|\mathcal{T} V_t - V\|_2 = \arg\min_{V \in \mathcal{F}} \|\mathbb{E}_{s'}[r + \gamma V_t(s') - V(s)|s]\|_2$$
$$V_{t+1} = \arg\min_{V \in \mathcal{F}} \|\mathbb{E}'_s[V(s) - \gamma V_t(s')|s]\|_2$$
$$\therefore w_{t+1} = \arg\min_{w \in \mathcal{R}} (w - \gamma 2 w_t)^2 + (2w - \gamma(1-\epsilon)2w_t)^2$$
$$w_{t+1} = \frac{6 - 4\epsilon}{5} \gamma w_t$$

Since $w_0 \neq 0$, then if $\gamma > \frac{5}{6-4\epsilon}$, then each w_{t+1} will be larger than the w_t and the iteration will diverge. This combination of parameters is valid if ϵ is small enough (i.e., $\gamma < 1$).

2. Following analogously to the previous item, but with the L_∞ norm, we will have that

$$w_{t+1} = \arg\min_{w \in \mathcal{R}} [\max(|w - \gamma 2 w_t|, |2w - \gamma(1-\epsilon)2w_t|)]$$

Since this is a linear program, we know that the minimizer is achieved when both of these terms are equal, though we don't yet know the signs of the terms inside each. Alternatively, we can argue geometrically by realizing that each term is the function $|x|$ shifted and scaled by some factors. Namely, the first term is shifted by $\overline{w}^1 = \gamma 2 w_t$ and the second one is shifted by $\overline{w}^2 = \gamma(1-\epsilon)w_t$ and scaled by 2.. From these expressions we notice that $|\overline{w}^1| > |\overline{w}^2|$ and the minimum must be realized between $\overline{w}^1$ and $\overline{w}_2$ (otherwise the point of intersection between the two terms would be further away from either $\overline{w}^1$ and $\overline{w}^2$). Thus, we get that the minimizer satisfies $|\overline{w}^1| \geq |w| \geq |\overline{w}^2|$ and we have that

$$2w - \gamma(1-\epsilon)2w_t = -w + \gamma 2 w_t$$
$$\therefore w_{t+1} = \frac{\gamma(2-\epsilon)}{3} w_t$$

So w_t will shrink to $w_t = 0$m which corresponds to the true value function.

This simple problem shows a marked distinction between tabular and non-tabular methods. Even an algorithm that was guaranteed to converge on the dynamic programming settings can fail at even a simple example. Thus, it's important to study when convergence can be guaranteed and when it cannot

2.3 Convergence and properties

We can show that approximate value iteration converges to a unique fixed point and its error can be bounded if we consider an approximate projection $\mathcal{A}$ using the L_∞ norm. Ultimately, we want to arrive at the following result

For an approximate projection $\mathcal{A}$ using the L_∞ norm, the AVI algorithm converges to the fixed point solution $\overline{V} = \mathcal{A}\mathcal{T}\overline{V}$, where a greedy policy $\overline{\pi}$ over $\overline{V}$ achieves the value $V^{\overline{\pi}}$ satisfying

$$\|V^* - V^{\overline{\pi}}\|_\infty \leq \frac{2}{(1-\gamma)^2} \inf_{V \in \mathcal{F}} \|V^* - V\|_\infty \tag{1}$$

In other words, the function space $\mathcal{F}$ may not be able to represent the optimal value function exactly, but even so the policy we extract from AVI won't be arbitrarily far from the optimal. More optimistically, if the approximation can be made exact, then the policy will be optimal. Let's prove this result in a series of subproblems

1. Show that the operator $\mathcal{A}\mathcal{T}$ is a γ-contractive, that is,

$$\|\mathcal{A}\mathcal{T}V_1 - \mathcal{A}\mathcal{T}V_2\|_\infty \leq \gamma\|V_1 - V_2\|_\infty$$

2. Show that after T iterations of AVI, we have the value function V_T with greedy policy π_T satisfying

$$\|V^* - V^{\pi_T}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq t \leq T} \|\mathcal{T}V_t - \mathcal{A}\mathcal{T}V_t\|_\infty + \frac{2\gamma^{T+1}}{1-\gamma}\|V^* - V_0\|_\infty$$

Hint: Use the property from the HW2 and the lecture notes when using a greedy policy π w.r.t. a value function V:

$$\|V^* - V^\pi\|_\infty \leq \frac{2\gamma}{1-\gamma}\|V^* - V\|_\infty$$

3. Use the results from the previous items to prove the bound presented in Equation 1.

Solution:

1. We notice that $\mathcal{A}$ is a normed projection operator, so it's non-expansive (may not contract, but won't expand). Since $\mathcal{T}$ is γ-contractive, combining both factors we get the desired result.

2. Let's first bound $\|V^* - V_{t+1}\|_\infty$ since a similar expression appear in the RHS of the hint

$$\begin{aligned} \|V^* - V_{t+1}\| &\leq \|V^* - \mathcal{T}V_t\|_\infty + \|\mathcal{T}V_t - V_{t+1}\|_\infty \\ &\leq \gamma\|V^* - V_t\| + \epsilon_t \end{aligned}$$

where we named the last term $\epsilon_t = \|\mathcal{T}V_t - V_{t+1}\|_\infty = \|\mathcal{T}V_t - \mathcal{A}\mathcal{T}V_t\|_\infty$ as the approximation error for convenience. Repeatedly applying the previous inequality and for convenience writing $\epsilon = \max_{t\in\{0,\ldots,T\}} \epsilon_t$ we get

$$\begin{aligned}\|V^* - V_T\|_\infty &\leq (1+\gamma+,\ldots,+\gamma^{T-1})\epsilon + \gamma^T\|V^* - V_0\|_\infty \\ &\leq \frac{1}{1-\gamma}\epsilon + \gamma^T\|V^* - V_0\|_\infty\end{aligned}$$

As we have bounded the RHS side of the hint, we desired result immediately follows form its application

3. First, we use the fact to that $\mathcal{A}\mathcal{T}$ is γ-contractive to prove that the AVI algorithm has a fixed point $\overline{V}$. Applying that the bound from the second item for $V_0 = \overline{V}$ and $T \to \infty$, we get that