STAT 234 – Sequential Decision Making April 9, 2021

Least-Squares Value Iteration

Instructor: Susan Murphy *Scribe: Kelly Zhang*

Contents

1 Overview

The focus of these notes is to understand Least-Squares Value Iteration, which is discussed in the paper by Osband, Van Roy, and Wen (`https://arxiv.org/pdf/1402.0635.pdf`); specifically see Algorithm 3 in the appendix. Future lectures will discuss RLSVI, which is the main algorithm presented in their paper.

Similar to Least Squares Policy Iteration, this Least Squares Value Iteration is a method for estimating the optimal policy from batch data. The two main differences are (1) LSVI uses a value iteration approach [a general method for learning the optimal policy that is an alternative to policy iteration], and (2) LSVI is developed for the finite horizon problem setting, rather than the infinite horizon, discounted reward setting.

2 Finite Horizon Problems

Here we discuss episodic, finite horizon Markov decision processes. We consider finite horizon problems with non-random, fixed horizon H. This means that each episode runs for exactly H timesteps every time. Note that in the finite horizon problem the optimal policy is the policy that maximizes the expected sum of rewards $E_\pi\left[\sum_{h=0}^{H-1} R_h\right]$, while in the infinite horizon discounted reward setting the optimal policy maximizes the infinite sum of discounted reward $E_\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t\right]$.

A very special property of in the infinite horizon discounted reward setting is that *the optimal action to take depends on the current state but not on the timestep t*. In other words if we visit state s at timestep $t = 1$ or $t = 1000$, the optimal action to take does not change. Let us illustrate this property mathematically. Recall that the value function of a policy π in the infinite horizon discounted reward setting is as follows:

$$V^\pi(s) = E_\pi\left[\sum_{s=t}^{\infty} \gamma^{s-t} R_t \middle| S_t = s\right] = E_\pi\left[R_t + \gamma V^\pi(S_{t+1}) \middle| S_t = s\right]$$

Note the Bellman optimality equation in this setting:

$$V^{\pi^*}(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \sum_{r,s'} P(r, s'|s, a)\left\{r + \gamma V^{\pi^*}(s')\right\}$$

Note that the value function does not change with t and in turn the optimal action to take in each timestep t also does not depend on t.

In contrast, in the finite horizon setting *the optimal action to take depends on both the current state and the timestep in the episode* $h \in [0\colon H-1]$. Intuitively, imagine that H is the number of years in my life and that all my money disappears at the end of year $H-1$. Near the end of my life, I have very little incentive to save money since my life will be over very soon. In fact at at timestep $H-1$, the optimal policy will probably to spend all my money to maximize the expected immediate reward $E[R_{H-1}]$, since there are no future rewards! However, at the start of my life, it may be optimal to save more money to increase reward in the future. Thus, generally, in finite horizon problems, the optimal policy becomes increasingly "greedy" as we get closer to the end of the episode, i.e., $h \in [0\colon H-1]$ increases.

Let us now write mathematically the optimal Q-functions and optimal policies in the finite horizon setting. First we write the Q-function for the last timestep in the episode $H-1$:

$$Q^*_{H-1}(s,a) = E\left[R_{H-1} | S_{H-1} = s, A_{H-1} = a\right]$$

Note that the expectation above does not depend on any policy. Thus the optimal Q-function, Q^*_{H-1}, is also the only Q-function. Thus the optimal policy at timestep $H-1$ is $\pi^*_{H-1}(s) = \text{argmax}_a\, Q^*_{H-1}(s,a)$. Now let us go one timestep back to $H-2$. The optimal Q-function for timestep $H-2$ chooses action A_{H-2}, assuming that we follow optimal policy π^*_{H-1} at the last timestep $H-1$:

$$Q^*_{H-2}(s,a) = E_{\pi^*_{H-1}}\left[R_{H-2} + R_{H-1} \middle| S_{H-2} = s, A_{H-2} = a\right]$$

$$= E\left[R_{H-2} + \operatorname*{argmax}_a Q^*_{H-1}(S_{H-1}, a) \middle| S_{H-2} = s, A_{H-2} = a\right]$$

Again the optimal policy in the $H-2$ timestep is defined as $\pi^*_{H-2}(s) = \text{argmax}_a\, Q^*_{H-2}(s,a)$.

We can apply this repeatedly for other values of h and get the following Q function:

$$Q^*_h(s,a) = E_{\{\pi^*_{h'}\}_{h'=h+1}^{H-1}}\left[\sum_{h'=h}^{H-1} R_{h'} \middle| S_h = s, A_h = a\right]$$

Note in the expectation above, we assume that we follow optimal policies $\{\pi^*_{h'}\}_{h'=h+1}^{H-1}$ for the timesteps $h+1$ through end of episode $H-1$.

$$= E_{\{\pi^*_{h'}\}_{h'=h+1}^{H-1}}\left[R_h + \operatorname*{argmax}_a Q^*_{h+1}(S_{h+1}, a) \middle| S_h = s, A_h = a\right]$$

The optimal policy $\pi^*_h(s) = \text{argmax}_a\, Q^*_h(s,a)$. In the end we have sequence of optimal policies $\{\pi^*_h\}_{h=0}^{H-1}$.

3 Least Squares Value Iteration (LSVI)

Throughout, we assume that we are in the off-policy MDP setting, in which we have access to data collected with a potentially unknown policy. In other words, we have access to data of the form $\mathcal{D}_L = (S_i, A_i, R_i, S'_i)_{i=1}^L$, which is collected by some policy. Our goal is to use this data to estimate an optimal policy, which maximizes expected sum of rewards in a finite horizon, episodic setting.

In the previous notes on Least Squares Policy Iteration, we discussed in generalized **policy iteration**, a method to estimate the optimal policy from batch data. Recall that in policy iteration, we alternate between *policy evaluation* and *policy improvement*. An alternative approach to estimating the optimal policy from batch data is **value iteration**. See Sutton and Barto section 4.4 for more details about value iteration (specifically in the infinite horizon setting). In this section we discuss Least Squares Value Iteration, which applies the idea of value iteration for finite horizon problems. Note that value iteration typically implies learning an optimal value function, while in LSVI, we are learning an optimal Q-function. We will use value iteration in the sense that we will first estimate Q^*_{H-1}. Then will use this estimate of Q^*_{H-1} to estimate Q^*_{H-2}. Then we will use our estimate of Q^*_{H-2} to help us estimate Q^*_{H-3}. And so on.

As we did for LSPI, we will parameterize our Q-function estimators with linear models: $Q_h(s,a) = \phi(s,a)^\top \theta_h$, where $\phi(s,a)$ is a feature mapping that is assumed to be known. (Although we do not do this here we could allow the feature $\phi(s,a)$ to depend on h as well.)

First we choose $\hat{\theta}_{H-1}$ to be the minimizer of the following least squares criterion:

$$\sum_{i=1}^{L} \left(\underbrace{R_i}_{\text{target}} - \phi(S_i, A_i)^\top \hat{\theta}_{H-1} \right)^2 \tag{1}$$

This means that $\hat{\theta}_{H-1} = \left(\sum_{i=1}^{L} \phi(S_i, A_i) \phi(S_i, A_i)^\top \right)^{-1} \sum_{i=1}^{L} R_i \phi(S_i, A_i)$.

However, in the paper the estimator is actually slightly different because they add an L2 regularization term. So rather than using the least squares criterion (1), we add an L2 regularization to get the following criterion:

$$\sum_{i=1}^{L} \left(\underbrace{R_i}_{\text{target}} - \hat{\theta}_{H-1}^\top \phi(S_i, A_i) \right)^2 + \lambda \hat{\theta}_{H-1}^\top \hat{\theta}_{H-1}$$

Above λ is some constant greater than zero that controls the degree of regularization. This means that $\hat{\theta}_{H-1} = \left(I\lambda + \sum_{i=1}^{L} \phi(S_i, A_i) \phi(S_i, A_i)^\top \right)^{-1} \sum_{i=1}^{L} R_i \phi(S_i, A_i)$.

Note that the L2 regularized least squares estimator can also be interpreted as the maximum a posteriori estimator for θ when we have multivariate normal priors on θ. In particular, the above corresponds to a $N(0, \lambda I)$ prior on θ_{H-1}.

To estimate θ_{H-2}, we minimize the following criterion, which is a function of the previously defined $\hat{\theta}_{H-1}$ (we also include an L2 regularization):

$$0 = \sum_{i=1}^{L} \left(\underbrace{R_i + \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ \hat{\theta}_{H-1}^\top \phi\left(S_i', a\right) \right\}}_{\text{target}} - \hat{\theta}_{H-2} \phi(S_i, A_i) \right)^2 + \lambda \hat{\theta}_{H-2}^\top \hat{\theta}_{H-2}$$

This means that $\hat{\theta}_{H-2} = \left(I\lambda + \sum_{i=1}^{L} \phi(S_i, A_i) \phi(S_i, A_i)^\top \right)^{-1} \sum_{i=1}^{L} \left(R_i + \operatorname{argmax}_{a \in \mathcal{A}} \left\{ \hat{\theta}_{H-1}^\top \phi\left(S_i', a\right) \right\} \right) \phi(S_i, A_i)$.

Similarly for other values of h, we can estimate θ_h by minimizing the least squares criterion,

assuming we already have estimate $\hat{\theta}_{h+1}$:

$$0 = \sum_{i=1}^{L} \left(\underbrace{R_i + \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left\{ \hat{\theta}_{h+1}^\top \phi\left(S_i', a\right) \right\}}_{\text{target}} - \hat{\theta}_h \phi(S_i, A_i) \right)^2 + \lambda \hat{\theta}_h^\top \hat{\theta}_h$$

This means that $\hat{\theta}_h = \left(I\lambda + \sum_{i=1}^{L} \phi(S_i, A_i)\phi(S_i, A_i)^\top \right)^{-1} \sum_{i=1}^{L} \left(R_i + \operatorname{argmax}_{a \in \mathcal{A}} \left\{ \hat{\theta}_{h+1}^\top \phi\left(S_i', a\right) \right\} \right) \phi(S_i, A_i)$.

It is interesting to note that in this finite horizon problem we can use least squares minimization whereas in the infinite horizon problem considering least squares led to the Bellman residual minimizer which is less preferable than the fixed point approximation(see the LSPI notes for why we favor the fixed point approximation). In this setting the fixed point approximation is obtained via least squares minimization. Why do you think this is the case?

Through this procedure we get estimates $\{\hat{\theta}_h\}_{h=0}^{H-1}$, which we can use to get estimates of Q-functions $\{Q_h^*\}_{h=0}^{H-1}$.

Algorithm 3 Least-Squares Value Iteration

Input: Data $\Phi(s_{i0},a_{i0}),r_{i0},..,\Phi(s_{iH-1},a_{iH-1}),r_{iH} : i < L$
Parameter $\lambda > 0$

Output: $\theta_{l0},...,\theta_{l,H-1}$

1: $\theta_{lH} \leftarrow 0$, $\Phi_H \leftarrow \mathbf{0}$
2: **for** $h = H-1,...,1,0$ **do**
3: Generate regression problem $A \in \mathbb{R}^{l \times K}$, $b \in \mathbb{R}^l$:

$$A \leftarrow \begin{bmatrix} \Phi_h(s_{0h},a_{0h}) \\ \vdots \\ \Phi_h(s_{l-1,h},a_{l-1,h}) \end{bmatrix}$$

$$b_i \leftarrow \begin{cases} r_{ih} + \max_\alpha \left(\Phi_{h+1} \tilde{\theta}_{l,h+1} \right)(s_{i,h+1},\alpha) & \text{if } h < H-1 \\ r_{ih} + r_{i,h+1} & \text{if } h = H-1 \end{cases}$$

4: Linear regression for value function

$$\theta_{lh} \leftarrow (A^\top A + \lambda I)^{-1} A^\top b$$

5: **end for**

14

Typo: $L = l$, $\theta_{Lh} = \theta_{lh}$