ECE239AS: Special Topics Course, Spring Quarter, 2023

Report: Model-Based Reinforcement Learning with Value-Targeted Regression

Yunqi Hong, Haochen Zhao, Deyi Wang

June 16, 2023

Abstract

In this report we focus on the core part of the paper *Model-Based Reinforcement Learning with Value-Targeted Regression*. In this part, the paper proposed a value-targeted regression algorithm for a linear mixture transition model-based reinforcement learning problem. The paper also proposed the sublinear regret bound of this algorithm. Our report is arranged as follows: *Section 1* introduces the background and motivation of the problem we focus on. *Section 2* illustrates the algorithm and the regret bound proposed by the paper. *Section 3 & 4* focus on the proof of the regret bound. We first decompose the regret into two parts, and then derive the upper bound for each part. This sublinear upper bound sheds light on solving this special case of model-based reinforcement learning problem.

1 Background & Motivation

In this paper, the authors focus on a special case of finite-horizon episodic reinforcement learning, where the unknown probability transition model P belongs to a known family of models $\mathcal{P}$ and the models in $\mathcal{P}$ is the linear mixture of known fixed basis models. We can write $\mathcal{P}$ as

$$\mathcal{P} = \{P_\theta | P_\theta = \sum_{i=1}^{d} \theta_i P_i\}$$

where $P_1, P_2, ..., P_d$ are fixed, known basis models and $\theta = (\theta_1, \theta_2, ..., \theta_d)$ are unknown, real-valued parameters. $P \in \mathcal{P}$ if and only if there exists θ in $\mathbb{R}^d$ such that

$$P(ds'|s,a) = \sum_{i=1}^{d} \theta_i P_i(ds'|s,a) = P_{\cdot}(ds'|s,a)^T \theta_*$$

Assumption 1 (Known Transition Model Family). *The unknown transition model P belongs to a family of models $\mathcal{P}$ which is available to the learning agent. The elements of $\mathcal{P}$ are transition kernels mapping state-action pairs to signed distributions over $\mathcal{S}$.*

Different from previous practice of selecting models based on their ability to predict next states or raw observations, the main novelty of this paper is their criterion to select models that are deemed consistent with past data, where the algorithm aims to select models based on their ability to produce small losses in a value-targeted regression problem.

2 Algorithm & Regret Bound

2.1 Algorithm Outline

The pseudo code of the Upper Confidence Reinforcement Learning - Value-Targeted Regression algorithm (UCRL-VTR) is shown in Algorithm 1:

Algorithm 1: UCRL-VTR

```
Input: Family of MDP models $\mathcal{P}$, $d$, $H$, $T = KH$;
Initialize: pick the sequence $\{\beta_k\}$ as in Eq. (2)
$B_1 = \mathcal{P}$
for $k = 1, 2, ..., K$ do
    Observe the initial state $s_1^k$ of episode $k$
    Optimistic Planing:
        $P_k = \text{argmax}_{P' \in B_k} V^*_{P',1}(s_1^k)$
        Compute $Q_{1,k}, ..., Q_{H,k}$ for $P_k$ using (3)
    for $h = 1, 2, ..., H$ do
        Choose the next action greedily with respect to $Q_{h,k}$:
            $a_h^k = \text{argmax}_{a \in \mathcal{A}} Q_{h,k}(s_h^k, a)$
        Observe state $s_{h+1}^k$
        Compute and store value predictions: $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$
    end
    Construct confidence set using value-targeted regression (1)
end
```

In the value-targeted regression, the model is estimated by regression using the estimated value functions as target.

$$\hat{P}_{k+1} = \text{argmin}_{P' \in \mathcal{P}} \sum_{k'=1}^{k} \sum_{h=1}^{H} (\langle P'(\cdot | s_h^{k'}, a_h^{k'}), V_{h+1,k'} \rangle - y_{h,k'})^2 \tag{1}$$

where $y_{h,k'} = V_{h+1,k'}(s_{h+1}^{k'})$. The regret target keeps changing as the algorithm constructs increasingly accurate value estimates.

In each episode, the subset B_{k+1} of the model class $\mathcal{P}$, also called as the confidence set that contains models consistent with the past collected data, is constructed. The confidence set B_{k+1} is centered at $\hat{P}_{k+1}$. The criterion of consistency is based on the total squared error the model incurs on the task of predicting values determined by the last value estimate along the transitions.

$$L_{k+1}(P, \hat{P}_{k+1}) = \sum_{k'=1}^{k} \sum_{h=1}^{H} (\langle P(\cdot|s_h^{k'}, a_h^{k'}) - \hat{P}_{k+1}(\cdot|s_h^{k'}, a_h^{k'}), V_{h+1,k'} \rangle)^2$$

$$B_{k+1} = \{P' \in \mathcal{P} | L_{k+1}(P', \hat{P}_{k+1}) \leq \beta_{k+1}\}$$

$$\beta_k = 2H^2 \log(\frac{2\mathcal{N}_\alpha}{\delta}) + 2H(kH-1)\alpha(2 + \sqrt{\log(\frac{4kH(kH-1)}{\delta})}) \tag{2}$$

where $\mathcal{N}_\alpha$ is a function of α.

Denote V_P^* as the optimal value function under a model P. We can find the model $P \in B_k$ that maximizes the value $V_{P',1}^*(s_1^k)$.

$$P_k = \text{argmax}_{P' \in B_k} V_{P',1}^*(s_1^k)$$

Given P_k, an optimal policy is extracted from the model using standard dynamic programming.

$$\begin{aligned} Q_{H+1,k}(s,a) &= 0 \\ V_{h,k}(s) &= \max_{a \in \mathcal{A}} Q_{h,k}(s,a) \\ Q_{h,k}(s,a) &= r(s,a) + \langle P_k(\cdot|s,a), V_{h+1,k} \rangle \end{aligned} \tag{3}$$

2.2 Regret Bound

The regret bound of the UCRL-VTR algorithm is shown as follows:

Theorem 1 (Regret of Algorithm 1). *Let Assumption 1 hold and let $\alpha \in (0,1)$. For $k \in [K]$, let β_k be*

$$\beta_k = 2H^2 \log(\frac{2\mathcal{N}_\alpha}{\delta}) + 2H(kH-1)\alpha(2 + \sqrt{\log(\frac{4kH(kH-1)}{\delta})})$$

Then, with probability $1 - 2\delta$,

$$R(K) \leq \alpha + H(d \wedge K(H-1)) + 4\sqrt{\beta_K dK(H-1)} + H\sqrt{2K(H-1)}\log(\frac{1}{\delta})$$

where $d = \dim_{\mathcal{E}}(\mathcal{F}, \alpha)$ is the Eluder dimension of $\mathcal{F}$ at scale α

3 Regret Decomposition

We are looking at the pseudo regret which captures the expected loss due to an imperfect policy. The pseudo regret may be large during the initial stage of the algorithm when the estimation error is large, so the policy is very suboptimal. In the long run, we aim to get less pseudo regret in each episode and our policy can approximate the underlying unknown optimal policy when the latest regret approaches zero. We can show this by showing that the accumulated pseudo regret grows sublinearly with respect to the episode number K. Under this condition we can state that the per episode pseudo regret is approaching zero as the number of episodes played approaches infinity.

To get this conclusion, we first in this section focus on per-episode pseudo regret and its decomposition into some sequence with good analytic property. Then we sum it over all the episodes and give bounds over the series sums. For those series sums, we will show they can be bounded with high probability in the later sections.

$$R_k = V_1^* \left(s_0\right) - V_1^{\pi_k} \left(s_0\right)$$

(by optimism of the value function)

$$\begin{aligned}
&\leq \mathrm{V}_{1,\mathrm{k}} \left(s_0\right) - V_1^{\pi_k} \left(s_0\right) \\
&= \left(r\left(s_1^k, a_1^k\right) + \langle P^k_{a_1^k}\left(s_1^k\right), V_{2,k}\rangle\right) - \left(r\left(s_1^k, a_1^k\right) + \langle P_{a_1^k}\left(s_1^k\right), V_2^{\pi_k}\rangle\right) \\
&= \langle P^k_{a_1^k}\left(s_1^k\right), V_{2,k}\rangle - \langle P_{a_1^k}\left(s_1^k\right), V_2^{\pi_k}\rangle
\end{aligned}$$

(by adding and subtracting a term)

$$\begin{aligned}
=&\langle\left[P^k_{a_1^k}\left(s_1^k\right) - P_{a_1^k}\left(s_1^k\right)\right], V_{2,k}\rangle + \langle P_{a_1^k}\left(s_1^k\right), [V_{2,k} - V_2^{\pi_k}]\rangle \\
&- \left(V_{2,k}\left(s_2^k\right) - V_2^{\pi_k}\left(s_2^k\right)\right) + \left(V_{2,k}\left(s_2^k\right) - V_2^{\pi_k}\left(s_2^k\right)\right)
\end{aligned}$$

(by the definition of $\xi_{k,2}$)

$$=\langle\left[P^k_{a_1^k}\left(s_1^k\right) - P_{a_1^k}\left(s_1^k\right)\right], V_{2,k}\rangle + \xi_2 + \left(V_{2,k}\left(s_2^k\right) - V_2^{\pi_k}\left(s_2^k\right)\right)$$

(by iterated expansion)

$$\begin{aligned}
&= \sum_{h=1}^{H-1} \langle\left[P^k_{a_h^k}\left(s_h^k\right) - P_{a_h^k}\left(s_h^k\right)\right], V_{h+1,k}\rangle + \sum_{h=2}^{H} \xi_{k,h} \\
&\leq \sup_{\hat{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle\left[\hat{P}_{a_h^k}\left(s_h^k\right) - P_{a_h^k}\left(s_h^k\right)\right], V_{h+1,k}\rangle + \sum_{h=2}^{H} \xi_{k,h} \\
&= W_k + \sum_{h=2}^{H} \xi_{k,h}
\end{aligned}$$

In the equation we define two new terms W_k and $\xi_{k,h}$ as follows:

$$\begin{aligned}
W_k &= \sup_{\hat{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle\left[\hat{P}_{a_h^k}\left(s_h^k\right) - P_{a_h^k}\left(s_h^k\right)\right], V_{h+1,k}\rangle \\
\xi_{k,h} &= \langle P_{a_{h-1}^k}\left(s_{h-1}^k\right), [V_{h,k} - V_h^{\pi_k}]\rangle - \left(V_{h,k}\left(s_h^k\right) - V_h^{\pi_k}\left(s_h^k\right)\right)
\end{aligned}$$

The term W_k is strongly related to the construction of $\mathcal{B}_k$. The term $\xi_{k,h}$ is the difference between the value function between steps which has zero mean on the filtration on past history. It forms a martingale which will be bounded later.

The overall regret overtime can be therefore bounded by

$$R(K) = \sum_{k=1}^{K} R_k \leq \sum_{k=1}^{K} W_k + \sum_{k=1}^{K}\sum_{h=2}^{H} \xi_{k,h}$$

4 Confidence Set & Proof of Upper bound

For the W_k, assume $P \in B_k$, we have,

$$\begin{aligned} W_k &= \sup_{\tilde{P} \in B_k} \sum_{h=1}^{H-1} (\tilde{P}(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k))^T V_{h,k} \\ &\leq \sum_{h=1}^{H-1} \sup_{\tilde{P} \in B_k} (\tilde{P}(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k))^T V_{h,k} \end{aligned}$$

Since the two transition matrices $\tilde{P}$ and P are both in the confidence set B_k, we need to construct a suitable confidence set to derive the upper bound of W_k.

4.1 Construction of the Confidence Set

The requirement of constructing the confidence set is to ensure the establishment of the assumption, i.e., $P \in B_k$ with high probability.

For the algorithm, before the start of the k th episode, we have collected a series of state, action and value pairs $(s_h^{k'}, a_h^{k'}, V_{h+1,k'})$ and their observation to the next state $s_{h+1}^{k'}$ for $h = 1, 2, \ldots, H-1$ and $k' = 1, 2, \ldots, k-1$. With this data, we can derive an estimator $\hat{P}_k$ of P,

$$\hat{P}_k = argmin_{\tilde{P} \in \mathcal{P}} \sum_{k'=1}^{k-1} \sum_{h=1}^{H-1} (\tilde{P}(\cdot|s_h^{k'}, a_h^{k'})^T V_{h+1,k'} - V_{h+1,k'}(s_{h+1}^{k'}))^2$$

It is reasonable to say that P is in the neighborhood of $\hat{P}_k$. Therefore, the confidence set should be like,

$$B_{k+1} = \{\tilde{P} \in \mathcal{P} | \sum_{k'=1}^{k} \sum_{h=1}^{H-1} ((\tilde{P}(\cdot|s_h^{k'}, a_h^{k'}) - \hat{P}_k(\cdot|s_h^{k'}, a_h^{k'}))^T V_{h+1,k'})^2 \leq \beta_k\}$$

where β_k to be determine.

Here we define a set of measurable functions $\mathcal{F}$ on $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}$,

$$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} : \exists \tilde{P} \in \mathcal{P} s.t. f(s, a, V) = \tilde{P}(\cdot|s, a)^T V\}$$

It is obvious that different $\tilde{P}$ corresponds to different f. Therefore, we denote the f corresponding to $\tilde{P}$ by $f_{\tilde{P}}$.

We use $p = (k'-1)(H-1) + h$ to represent the step of the algorithm in the whole range of time. Given a state action pair (s_p, a_p), we can query a s_{p+1} from $P(\cdot|s_p, a_p)$. With the given value V_p, we have,

$$\mathbb{E}(V_p(s_{p+1})|s_p, a_p) = P(\cdot|s_p, a_p)^T V_p$$

Therefore, if we define $X_p = (s_p, a_p, V_p)$ and $Y_p = V_t(s_{p+1})$, then before the k th episode, we have a sequence of $(X_p, Y_p)_{p=1,2,\ldots,t}$, where $t = (k-1)(H-1)$. If we let $\hat{f}_t$ to represent the f

corresponding to $\hat{P}_k$, then we have,

$$\hat{f}_t = argmin_{f\in\mathcal{F}}\sum_{p=1}^{t}(f(X_p)-Y_p)^2$$

If let ϕ denote the mapping from $\mathcal{P}$ to $\mathcal{F}$, then the confidence set could be written as

$$B_{k+1} = \{\phi^{-1}(f) : f\in\mathcal{F} : \sum_{p=1}^{t}(f(X_p)-\hat{f}_t(X_p))^2 \le \beta_k\} = \phi^{-1}(\mathcal{F}_t(\beta_k))$$

Theorem 2 shows the relation between the ground truth f and the confidence set $\mathcal{F}_t(\beta)$.

Theorem 2 *Assume that the functions in $\mathcal{F}$ are bounded by the constant $C > 0$, and the sequence of $(X_p, Y_p)_{p=1,2,\ldots,t}$ defined above satisfies that for each p, $Z_p = Y_p - f(X_p)$ is conditionally σ-subgaussian. Then for any $\alpha > 0$, w.p. $1-\delta$, $f\in\mathcal{F}_t(\beta_t(\delta,\alpha))$ for $\mathcal{F}_t(\beta)$ defined above, where,*

$$\beta_t(\delta,\alpha) = 8\sigma^2\log(\frac{2N_\alpha}{\delta}) + 4t\alpha(C+\sqrt{\sigma^2\log(\frac{4t(t+1)}{\delta})})$$

where N_α is the $\|\cdot\|_\infty$-covering number of $\mathcal{F}$ at scale α.

In the finite MDPs, the value of a state wouldn't be larger than H. Therefore, $C = H$ and Y_p, $f(X_p)\in[0,H]$. What's more, since $\mathbb{E}(Y_p|X_p) = f(X_p)$, then $Z_p = Y_p - f(X_p)$ is conditionally $\frac{H}{2}$-subgaussian. Therefore, with $C = H$, $\sigma = \frac{H}{2}$ and $t = k(H-1)$, the β_k of the confidence set is given by,

$$\beta_k = 2H^2\log(\frac{2\mathcal{N}_\alpha}{\delta}) + 2Hk(H-1)\alpha(2+\sqrt{\log(\frac{4k(H-1)(k(H-1)+1)}{\delta}}))$$

With a relaxation, we can set,

$$\beta_k = 2H^2\log(\frac{2\mathcal{N}_\alpha}{\delta}) + 2H(kH-1)\alpha(2+\sqrt{\log(\frac{4kH(kH-1)}{\delta}}))$$

4.2 Upper bound of $\sum_{k=1}^{K} W_k$

Previously, we got,

$$W_k \le \sum_{h=1}^{H-1}\sup_{\tilde{P}\in B_k}(\tilde{P}(\cdot|s_h^k,a_h^k)-P(\cdot|s_h^k,a_h^k))^T V_{h,k}$$

We can use the f functions in $\mathcal{F}$ to rewrite W_k and relax it more,

$$\begin{aligned} W_k &\le \sum_{h=1}^{H-1}\sup_{\tilde{P},P\in B_k}(\tilde{P}(\cdot|s_h^k,a_h^k)-P(\cdot|s_h^k,a_h^k))^T V_{h,k} \\ &= \sum_{p=(k-1)(H-1)+1}^{k(H-1)}\sup_{f_1,f_2\in\mathcal{F}_p}|f_1(X_p)-f_2(X_p)| \end{aligned}$$

where $\mathcal{F}_t$ represents the confidence set.

For a sequence $(x_p)_{p=1,2,\ldots,t}$, let $f|_{x_{1:t}} = (f(x_1), f(x_2), \ldots, f(x_t))^T$. Define $diam(\mathcal{F}|x_{1:t}) := \sup_{f_1,f_2\in\mathcal{F}} ||f_1|_{x_{1:t}} - f_2|_{x_{1:t}}||_2^2$ be the diameter of $\mathcal{F}$ on $x_{1:t}$. Therefore, we can derive that,

$$\sum_{k=1}^{K} W_k \leq \sum_{p=1}^{K(H-1)} diam(\mathcal{F}_p|X_p)$$

where X_p is in the sequence $(X_p, Y_p)_{p=1,2,\ldots,K(H-1)}$.

Given $\mathcal{F}_t = \{f|\sum_{p=1}^{t}(f(X_p)-\hat{f}_t(X_p))^2 \leq \beta\}$, we can easily derive that $diam(\mathcal{F}_t|x_{1:t}) = 2\sqrt{\beta}$.

Theorem 3 shows the relations between $\sum_{p=1}^{t} diam(\mathcal{F}_p|X_p)$ and $diam(\mathcal{F}_t|x_{1:t})$.

Theorem 3 *Assume that the functions in $\mathcal{F}$ are bounded by the constant $C > 0$. If the sequences $(\mathcal{F}_p)_{p\geq 1}$ and $(X_p)_{p\geq 1}$ satisfy that $\mathcal{F}_p \subset \mathcal{F}$ and $X_p \in \mathcal{X}$ hold for each $p \geq 1$, then for any $T \geq 1$ and $\alpha > 0$, it holds that,*

$$\sum_{p=1}^{T} diam(\mathcal{F}_p|X_p) \leq \alpha + C(d \wedge T) + 2\delta_T\sqrt{dT}$$

where $\delta_T == \max_{1\leq t\leq T} diam(\mathcal{F}_t|x_{1:t})$ and d is the Eluder dimension of $\mathcal{F}$ at scale α. Here $\wedge$ is the operator which selects the smaller element between the two elements.

Since the β of $\mathcal{F}_t$ increases when t gets larger, then $\delta_T = diam(\mathcal{F}_T|x_{1:T}) = 2\sqrt{\beta_K}$ when $T = K(H-1)$.

Therefore, let $C = H$ and $T = K(H-1)$, we derive an upper bound of $\sum_{k=1}^{K} W_k$.

For $\alpha > 0$, w.p. $1-\delta$, we have,

$$\sum_{k=1}^{K} W_k \leq \alpha + H(d \wedge K(H-1)) + 4\sqrt{\beta_K dK(H-1)}$$

where

$$\beta_K = 2H^2\log(\frac{2\mathcal{N}_\alpha}{\delta}) + 2H(KH-1)\alpha(2 + \sqrt{\log(\frac{4KH(KH-1)}{\delta})})$$

4.3 Upper bound of $\sum_{k=1}^{K}\sum_{h=2}^{H}\xi_{h,k}$

Previously, we have that,

$$\begin{aligned}\xi_{h+1,k} &= \langle P_{a_h^k}(s_h^k), V_{h+1,k} - V_{h+1}^{\pi_k}\rangle - (V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k)) \\ &= P(\cdot|s_h^k, a_h^k)^T V_{h+1,k} - V_{h+1,k}(s_{h+1}^k) - (P(\cdot|s_h^k, a_h^k)^T V_{h+1}^{\pi_k} - V_{h+1}^{\pi_k}(s_{h+1}^k))\end{aligned}$$

Notice that $s_{h+1}^k \sim P(\cdot|s_h^k, a_h^k)$, we have,

$$\mathbb{E}(\xi_{h+1,k}) = 0$$

Therefore, the sequence of $(\xi_{2,1}, \xi_{3,1}, \dots, \xi_{H,1}, \dots, \xi_{2,2}, \xi_{3,2}, \dots, \xi_{H,2}, \xi_{2,3}, \dots)$ is a sequence of martingale differences.

Since the finite MDPs and $\xi_{h+1,k} \in [-H, H]$, we have for each pair of h and k, $\xi_{h+1,k}$ is H-subgaussian. Therefore, we can derive an Azuma Hoeffding bound for $\sum_{k=1}^{K} \sum_{h=1}^{H-1} \xi_{h+1,k}$. We have w.p. $1 - \delta$,

$$\sum_{k=1}^{K} \sum_{h=2}^{H} \xi_{h,k} = \sum_{k=1}^{K} \sum_{h=1}^{H-1} \xi_{h+1,k} \leq \sqrt{2 \sum_{i=1}^{K(H-1)} H^2 \log(\frac{1}{\delta})} = H\sqrt{2K(H-1)} \log(\frac{1}{\delta})$$

4.4 Upper bound of pseudo regret

Since,

$$R(K) \leq \sum_{k=1}^{K} W_k + \sum_{k=1}^{K} \sum_{h=2}^{H} \xi_{h,k}$$

we use intersection of events to derive the upper bound of the pseudo regret. Therefore, we have for any $\alpha > 0$, w.p. $1 - 2\delta$,

$$R(K) \leq \alpha + H(d \wedge K(H-1)) + 4\sqrt{\beta_K d K(H-1)} + H\sqrt{2K(H-1)} \log(\frac{1}{\delta})$$

where β_K has been given before.