24. Featurized MDPs

In tabular (finite-horizon) MDPs $M = (\mathcal{S}, \mathcal{A}, P, r, \mu, H)$, roughly speaking, the learner has to learn about reward and transition probabilities for *all* states and actions in the worst-case. This is reflected in lower bounds on the regret that scale with $R_K \geq \Omega(H^{3/2}\sqrt{ASK})$ (in the time in-homogeneous case).

In many applications the state space can be huge, and reinforcement learning is often used together with function approximation. In such settings, we want to avoid bounds that scale directly with the number of states S. The simplest parametric models often rely on state-action features and linearly parametrized transition and reward functions. The goal is to obtain bounds that scale with the complexity of the function class (e.g. the feature dimension in linear models), and are *independent* of S and A.

Historically, many ideas for online learning in linear MDP models are borrowed from the linear bandit model. Beyond what is written here, you may find it helpful to read about stochstic linear bandits and LinUCB (see chapters 19 and 20 of the Bandit Book).

Linear Mixture MDPs

We focus on the episodic, finite-horzion MDPs $M = (\mathcal{S}, \mathcal{A}, P_h, r_h, \mu, H)$ with time in-homogenous reward r_h and transition matrix P_h. We let $\mathcal{S}$ be a finite but possibly very large state space, and $\mathcal{A}$ be a finite action space. With care, most of the analysis can be extended to infinite state and action spaces. As before, we assume that the reward function $r_h(s,a) \in [0,1]$ is known.

We now impose additional (linear) structure on the transition kernel P_h. For this we assume the learner has access to features $\phi(s,a,s') \in \mathbb{R}^d$ that satisfy $\|\phi(s,a,s')\|_2 \leq 1$. In time-inhomogeneous *linear mixture MDPs*, the transition kernel is of the form

$$P_{h,a}(s,s') = \langle \phi(s,a,s'), \theta_h^* \rangle$$

for some unkown parameter $\theta_h^* \in \mathbb{R}^d$ with $\|\theta_h^*\|_2 \leq 1$. We remark that tabular MDPs are recovered using $\phi(s,a,s') = e_{s,a,s'}$, where $e_{s,a,s'}$ are the unit vectors in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$.

For any function $V : \mathcal{S} \to \mathbb{R}$, we define

$$\phi_V(s,a) = \sum_{s'} \phi(s,a,s') V(s') \in \mathbb{R}^d$$

Note that $\langle \phi_V(s,a), \theta^* \rangle$ predicts the expected value of $V(s')$ when s' is sampled from $P_{h,a}(s)$:

$$P_{h,a}(s)V = \sum_{s'} P_{h,a}(s,s')V(s') = \sum_{s'} \langle \phi(s,a,s'), \theta_h^* \rangle V(s') = \langle \phi_V(s,a), \theta_h^* \rangle$$

Value Targeted Regression (VTR)

Now that we have specified the parametrized model, the next step is to construct an estimator of the unknown paramter. An estimator of θ^* allows us to predict the value of any policy. For the algorithm, we are particularly interested in constructing optimistic estimates of the value function. Hence we will also need a confidence set.

Let $(V_h^{(j)})_{h \le H}^{j<k}$ be a sequence of value functions constructed up to episode $k-1$. Let $\phi_{h,j} = \phi_{V_{h+1}^{(j)}}(S_h^{(j)}, A_h^{(j)})$ and $y_{h,j} = V_{h+1}^{(j)}(S_{h+1}^{(j)})$. By constrution, we have that $\mathbb{E}[y_{h,j}] = \langle \phi_{h,j}, \theta^* \rangle$ and $|y_{h,j}| \le H$. Define the *regularized least-squares estimator*

$$\hat{\theta}_{h,k} = \arg\min_{\theta} \sum_{j=0}^{k-1} \left(\langle \phi_{h,j}, \theta \rangle - y_{h,j} \right)^2 + \lambda \|\theta\|^2$$

Let $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ be the idenity matrix. We have the following closed form for $\hat{\theta}_{h,k}$:

$$\hat{\theta}_{h,k} = \Sigma_{h,k}^{-1} \sum_{j=0}^{k-1} \phi_{h,j} y_{h,j} \qquad \text{where} \quad \Sigma_{h,k} = \sum_{j=0}^{k-1} \phi_{h,j} \phi_{h,j}^\top + \lambda \mathbf{I}_d$$

The next step is to quantify the uncertainy in the estimation. Mirroring the steps in the tabular setting, we construct a confidence set for $\hat{\theta}_{h,k}$.

For a positive (semi-)definite matrix $\Sigma \in \mathbb{R}^{d \times d}$ and vector $v \in \mathbb{R}^d$, define the (semi-)norm $\|a\|_\Sigma = \sqrt{\langle v, \Sigma v \rangle}$. We make use of the following elliptical confidence set for $\hat{\theta}_{h,k}$

$$C_{h,\delta}^{(k)} = \{ \theta : \|\theta - \hat{\theta}_{h,k}\|_{\Sigma_{h,k}}^2 \le \beta_{h,k,\delta} \}$$

where

$$\beta_{h,k,\delta}^{1/2} = H \sqrt{\log\det(\Sigma_{h,k}) - \log\det(\Sigma_{h,0}) + 2\log(1/\delta)} + \sqrt{\lambda}$$

The log determinant of $\Sigma_{h,k}$ can be computed online by the algorithm. For the analysis, it is useful to further upper bound $\beta_{h,k,\delta}$. It is possible to show the following upper bound on $\beta_{h,k,\delta}$ that holds independent of the data sequence:

$$\beta_{h,k,\delta}^{1/2} \le H \sqrt{d \log(1 + k/(d\lambda)) + 2\log(1/\delta)} + \sqrt{\lambda}$$

For a derivation of the above inequality see Lemma 19.4 of the Bandit Book. The next lemma formally specifies the confidence probabilty.

Lemma (Online Least-Squares Confidence) Fix some $0 \leq h < H$. Then

$$\mathbb{P}[\theta_h^* \in \cap_{k \geq 1} C_{h,k,\delta}] \geq 1 - \delta$$

Proof: The above result is presented as Theorem 2 in Abbasi-Yadkori et al (2011), where the proof can also be found. ■

The confidence set can be used to derive bounds on the estimation error with probability at least $1 - \delta$ as follows:

$$|\langle \phi_V(s,a), \hat{\theta}_{h,k} - \theta^* \rangle| \leq \|\phi_V(s,a)\|_{\Sigma_{h,k}^{-1}} \|\hat{\theta}_{h,k} - \theta^*\|_{\Sigma_{h,k}} \leq \beta_{h,k,\delta}^{1/2} \|\phi_V(s,a)\|_{\Sigma_{h,k}^{-1}}$$

The first inequality is by Cauchy-Schwarz and the second inequality uses the confidence bound from the previous lemma.

UCRL-VTR

Similar to the tabular UCRL and UCBVI algorithms, UCRL-VTR uses the estimates $\hat{\theta}_{h,k}$ to compute an optimistic policy. One way of obtaining an optimistic policy is from optimistic Q-estimates $Q_h^{(k)}(s,a)$ defined via backwards induction. Then UCRL-VTR follows the greedy policy w.r.t. the optimistic Q-values.

UCRL-VTR

In episodes $k = 1, \ldots, K$,

1 Set $V_H^{(k)}(s) = 0$. Compute $\hat{\theta}_{h,k}$ and $\Sigma_{h,k}$. Recursively define optimistic value functions

For $h = H-1, \ldots, 0$:

$$\hat{\theta}_{h,k} = \arg\min_{\theta} \sum_{j=1}^{k-1} \left(\langle \phi_{h,j}, \theta \rangle - y_{h,j}\right)^2 + \lambda \|\theta\|_2^2$$

$$\Sigma_{h,k} = \sum_{j=1}^{k} \phi_{h,j} \phi_{h,j}^\top + \lambda \mathbf{I}_d$$

$$Q_h^{(k)}(s,a) = \left(r_h(s,a) + \langle \phi_{V_{h+1}^{(k)}}(s,a), \theta_{h,k} \rangle + \beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(s,a)\|_{\Sigma_{h,k}^{-1}}\right) \wedge H$$

$$V_h^{(k)}(s) = \max_a Q_h^{(k)}(s,a)$$

2 Follow greedy policy w.r.t. $Q_h^{(k)}(s,a)$.

For $h = 0, \dots, H-1$:

$$A_h^{(k)} = \arg\max_{a \in \mathcal{A}} Q_h^{(k)}(S_h^{(k)}, a)$$

Let $\phi_{h,k} = \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})$ and $y_{h,k} = V_{h+1}^{(k)}(S_{h+1}^{(k)})$.

We are now in the position to state a regret bound for UCRL-VTR.

Theorem (UCRL-VTR Regret) The regret of UCRL-VTR satisfies with probability at least $1-2\delta$:

$$R_K \le \mathcal{O}\big(dH^2 \log(K) \sqrt{K \log(KH/\delta)}\big)$$

Note that the bound scales with the feature dimension d, but not the size of the state space or action space. The lower bound for this setting is $R_K \ge \Omega(dH^{3/2}\sqrt{K})$, therefore our upper bound is tight except for a factor $\sqrt{H}$.

Proof:

Our proof strategy follows the same steps as in the proof of UCRL.

Step 1 (Optimism):

Taking the union bound over $h = 0, \dots, H-1$, the previous lemma implies that with probability at least $1-\delta$, for all $h \in [H-1]$ and all $k \ge 0$, $\theta_h^* \in C_{h,\delta/H}^{(k)}$. In the following, we condition on this event. Using induction over $h = H, H-1, \dots, 0$, we can show that

$$V_0^*(S_h^{(k)}) \le V_0^{(k)}(S_h^{(k)})$$

Step 2 (Bellman recursion and estimation error):

For any $h = 0, \dots, H-1$, we find

$$\begin{aligned}
&V_h^{(k)}(S_h^{(k)}) - V_h^{\pi_k}(S_h^{(k)}) \\
&\le \langle \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)}), \hat{\theta}_{h,k} \rangle + \beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}} - P^*_{h,A_h^{(k)}}(S_h^{(k)}) V_{h+1}^{\pi_k} \\
&= \langle \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)}), \hat{\theta}_{h,k} - \theta^* \rangle + \beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}} + P^*_{h,A_h^{(k)}}(S_h^{(k)})(V_{h+1}^{(k)} - V_{h+1}^{\pi_k})
\end{aligned}$$

The inequality is by the definition of $V_h^{(k)}$ and dropping the truncation, and in the last line we add and subtract $P^*_{h,A_h^{(k)}}(S_h^{(k)}) V_{h+1}^{(k)} = \langle \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)}), \theta^* \rangle$. Further, by Cauchy-Schwarz on the event $\theta^* \in C_{k,\delta/H}$ we get

$$\langle \phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)}), \hat{\theta}_{h,k} - \theta^* \rangle \le \beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}}$$

Continuing the previous display, we find

$$\begin{aligned}
&V_h^{(k)}(S_h^{(k)}) - V_h^{\pi_k}(S_h^{(k)}) \\
&\le 2\beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}} + P^*_{h,A_h^{(k)}}(S_h^{(k)})(V_{h+1}^{(k)} - V_{h+1}^{\pi_k}) \\
&= 2\beta_{h,k,\delta/H}^{1/2} \|\phi_{V_{h+1}^{(k)}}(S_h^{(k)}, A_h^{(k)})\|_{\Sigma_{h,k}^{-1}} + V_{h+1}^{(k)}(S_{h+1}^{(k)}) - V_{h+1}^{\pi_k}(S_{h+1}^{(k)}) + \xi_{h,k}
\end{aligned}$$

where we defined

$$\xi_{h,k} = \left(P^*_{h,A_h^{(k)}}(S_h^{(k)})(V_{h+1}^{(k)} - V_{h+1}^{\pi_k})\right) - \left(V_{h+1}^{(k)}(S_{h+1}^{(k)}) - V_{h+1}^{\pi_k}(S_{h+1}^{(k)})\right)$$

Recursively appliying the previous inequality and summing over all episodes yields

$$\sum_{k=1}^{K} V_0^{(k)}(S_0^{(k)}) - V_0^{\pi_k}(S_0^{(k)}) \le \sum_{k=1}^{K} \sum_{h=0}^{H-1} 2\beta_{h,k,\delta}^{1/2} \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}} + \xi_{h,k}$$

Note that $\xi_{h,k}$ is a martingale difference sequence, hence by Azuma-Hoeffdings inequality we have with probability at least $1-\delta$,

$$\sum_{k=1}^{K} \sum_{h=0}^{H-1} \xi_{h,k} \le H\sqrt{\frac{HK}{2}\log(1/\delta)}$$

Step 3 (Cauchy-Schwarz):

Note that $\beta_{h,k,\delta}$ is non-decreasing in both h and k. Very little is lost by bounding $\beta_{h,k,\delta} \le \beta_{H,K,\delta}$. From the previous step, we are left to bound the sum over uncertainties $\|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}}$. We start with an application of the Cauchy-Schwarz inequality. Applied to

sequences $(a_i)_{i=1}^n$, $(b_i)_{i=1}^n$, we have that $|\sum_{i=1}^n a_i b_i| \leq \sqrt{\sum_{i=1}^n a_i^2 \sum_{j=1}^n b_i^2}$. Applied to the regret, we get:

$$\sum_{k=1}^{K} \sum_{h=0}^{H-1} 2\beta_{h,k,\delta}^{1/2} \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}} \leq \sum_{h=0}^{H-1} 2\beta_{h,K,\delta}^{1/2} \sqrt{K \sum_{k=1}^{K} \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}}^2}$$

Step 4 (Elliptic potential lemma):

The penultima step is to control the sum over squared uncertainties $\|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}}^2$. This classical result is sometimes refered to as the elliptic potential lemma:

$$\sum_{k=1}^{K} \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}}^2 \leq \mathcal{O}(d \log(K))$$

The proof, as mentioned earlier, can be found as Lemma 19.4 in the Bandit Book.

Step 5 (Summing up):

It remains to chain the previous steps and take the union bound over the event where the confidence set contains the true parameter and the application of Azuma-Hoeffdings.

$$\begin{aligned} R_K &= \sum_{k=1}^{K} V_0^{(k)}(S_0^{(k)}) - V_0^{\pi_k}(S_0^{(k)}) \\ &\leq \sum_{k=1}^{K} \sum_{h=0}^{H-1} \left(2\beta_{h,k,\delta}^{1/2} \|\phi_{h,k}\|_{\Sigma_{h,k}^{-1}} + \xi_{h,k}\right) \\ &\leq C \cdot H\beta_{H,K,\delta}^{1/2} \sqrt{d \log(K) K} + H^{3/2} \sqrt{2K \log(1/\delta)} \end{aligned}$$

For some universal constant C. This completes the proof. ■

Linear MDPs

So far we have seen the linear mixture MDP model. This is not the only way one can parameterize the transition matrix. An alternative is the *linear MDP* model, defined as follows for features $\phi(s,a) \in \mathbb{R}^d$ and parameters $\psi_h^* \in \mathbb{R}^{d \times S}$ and $\theta_h^* \in \mathbb{R}^d$:

$$\begin{aligned} P_h^*(s,s') &= \langle \phi(s,a), \psi_h^*(s') \rangle \\ r_h(s,a) &= \langle \phi(s,a), \theta_h^* \rangle \end{aligned}$$

Note that tabular MDPs are recovered using $\phi(s,a) = e_{s,a}$, where $e_{s,a}$ are the unit vectors in $\mathbb{R}^{\mathcal{S} \times \mathcal{A}}$.

Compared to the linear mixture model, an immediate observation is that the dependence on the the next state s' is pushed into the parameter $\psi_h(s') \in \mathbb{R}^d$. Consequently, the dimension of the parameter space scales with the number of states, and it is not immediately clear how we can avoid the S dependence in the regret bounds.

Another consequence of this model is that the Q-function for *any* policy is linear in the features $\phi(s,a)$.

Lemma:

Under the linear MDP assumption, for any policy π the Q-function $Q_h^\pi(s,a)$ is linear in the features $\phi(s,a)$. That is, there exist parameters $w_h^\pi \in \mathbb{R}^d$ such that

$$Q_h^\pi(s,a) = \langle \phi(s,a), w_h^\pi \rangle$$

Proof: The claim follows directly from the definition of Q_h^π and the assumptions on $r_h(s,a)$ and $P_{h,a}(s)$.

$$\begin{aligned} Q_h^\pi(s,a) &= r_h(s,a) + P_{h,a}(s) V_{h+1}^\pi \\ &= \langle \phi(s,a), \theta_h^* \rangle + \sum_{s'} V_{h+1}^\pi(s') \langle \phi(s,a), \psi_h^*(s') \rangle \\ &= \langle \phi(s,a), w_h^\pi \rangle \end{aligned}$$

where we defined $w_h^\pi = \theta_h^* + \sum_{s'} \psi_h^*(s') V_{h+1}^\pi(s')$ for the last equation. ■

In light of this lemma, our goal is to estimate $w_h^{\pi^*}$. This can be done using *least-squares value iteration* (LSVI). Let $\{S_1^{(j)}, A_1^{(j)}, \dots, S_{H-1}^{(j)}, A_{H-1}^{(j)}, S_H^{(j)}\}_{j=1}^{k-1}$ be the data available at the beginning of episode k. Denote $\phi_{h,j} = \phi(S_h^{(j)}, A_h^{(j)})$ and define targets $y_{h,j} = r_h(S_h^{(j)}, A_h^{(j)}) + \max_{a \in \mathcal{A}} Q_{h+1}^{(j)}(S_h^{(j)}, a)$ based on $Q_{h+1}^{(j)}(s,a)$ estimates obtained in episodes $j = 1, \dots, k-1$.

Least-squares value iteration solves the following problem:

$$\hat{w}_{h,k} = \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left(\langle \phi_{j,h}, w \rangle - y_{j,h} \right)^2 + \lambda \|w\|_2^2$$

The closed form solution is $w_{h,k} = \Sigma_{h,k}^{-1} \sum_{j=1}^{k-1} \phi_{h,j} y_{h,j}$ where $\Sigma_{h,k} = \sum_{j=1}^{k-1} \phi_{j,h} \phi_{j,h}^\top + \lambda \mathbf{I}_d$.

Based on the estimate $\hat{w}_{h,k}$, we can define optimistic Q- and V-estimates:

$$Q_h^{(k)}(s,a) = (\langle \phi(s,a), \hat{w}_{h,k} \rangle + \tilde{\beta}_{k,\delta}^{1/2} \|\phi(s,a)\|_{\Sigma_{h,k}^{-1}}) \wedge H$$
$$V_h^{(k)}(s) = \max_{a \in \mathcal{A}} Q_h^{(k)}$$

Assuming that the features satisfy $\|\phi(s,a)\|_2 \leq 1$ and the true parameters satisfy $\|\theta_h^*\|_2 \leq 1$ and $\|\psi_h^* v\|_2 \leq \sqrt{d}$ for all $v \in \mathbb{R}^S$ with $\|v\|_\infty \leq 1$, one can choose the confidence parameter as follows:

$$\tilde{\beta}_{k,h,\delta} = \mathcal{O}\left(d^2 \log(\frac{HK}{\delta}) \right)$$

This result is the key to unlock a regret bound that is independent of the size of the state space S. The proof requires a delicate covering argument. For details refer to chapter 8 of the RL Theory Book

LSVI-UCB

Algorithm: LSVI-UCB

In episodes $k = 1, \ldots, K$,

1. Initialize $V_H^{(j)}(s) = 0$ for $j = 1, \ldots, k-1$.

   For $h = H-1, \ldots, 0$, compute optimistic Q estimates:

$$y_{h,j} = r_h(S_h^{(j)}, A_h^{(j)}) + V_{h+1}^{(j)}(S_h^{(j)}) \quad \forall\, j = 1, \ldots, k-1$$
$$\phi_{h,j} = \phi(S_h^{(j)}, A_h^{(j)}) \quad \forall\, j = 1, \ldots, k-1$$
$$\hat{w}_{h,k} = \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{k-1} \left(\langle \phi_{j,h}, w \rangle - y_{j,h} \right)^2 + \lambda \|w\|_2^2$$
$$\Sigma_{h,k} = \sum_{j=1}^{k-1} \phi_{j,h} \phi_{j,h}^\top + \lambda \mathbf{I}_d$$
$$Q_h^{(k)}(s,a) = (\langle \phi(s,a), \hat{w}_{h,k} \rangle + \tilde{\beta}_{k,\delta}^{1/2} \|\phi(s,a)\|_{\Sigma_{h,k}^{-1}}) \wedge H$$

2. For $h = 0, \ldots, H-1$, follow greedy policy

$$A_h^{(k)} = \arg\max_{a \in \mathcal{A}} Q_h^{(k)}(S_h^{(k)}, a)$$

Note that computing the optimistic policy in episode k can be done in time $\mathcal{O}(Hd^2 + HAd)$ by incrementally updating the least-square estimates $\hat{w}_{h,k}$ using the

Sherman-Morrison formula. Compared to UCRL-VTR, this avoids iteration over the state space S, which is a big advantage!

Theorem (LSVI-UCB Regret)

The regret of LSVI-UCB is bounded up to logarihmic factors and with probability at least $1-\delta$ as follows:

$$R_K \leq \tilde{\mathcal{O}}(d^{3/2}H^2\sqrt{K})$$

Proof: The proof idea follows a similar strategy as the proof we presented for UCRL-VTR. As mentioned before, the crux is to show a confidence bound for LSVI that is indepenent of the size of the state space. For details, we again refer you to chapter 8 of the RL Theory Book. ■

Notes

Bernstein-type bounds for VTR (UCRL-VTR$^+$)

The UCRL-VTR$^+$ algorithm is computationally efficient and able to obtain a regret upper bound of $\mathcal{O}(dH\sqrt{K})$, and $\mathcal{O}(d\sqrt{T}(1-\gamma)^{-1.5})$ in the episodic and discounted, infinite horizon setting respectively. These results rely on using bernstein-type bounds.

Better regret bounds for Linear MDPs (Eleanor)?

A careful reader might have noticed that the regret bound for LSVI-UCB, $\tilde{\mathcal{O}}(d^{3/2}H^2\sqrt{K})$, is not tight with the tabular lower bound, $\Omega(d\sqrt{K})$. The difference is in a factor of $\sqrt{d}$. The Eleanor algorithm (Algorithm 1 in Zanette et al (2020)) is able to shave of the factor of $\sqrt{d}$, obtaining a regret upper bound of $\tilde{\mathcal{O}}(dH^2\sqrt{K})$. However, it is not currently known if the alogrithm can be implemented in a computationally efficient way. The Eleanor algorithm operates under the assumption of low inherent Bellman error (Definition 1 in Zanette et al (2020)), which means the function class is approximately closed under the Bellman optimality operator. It is interesting to note that this assumption is more general than the Linear MDP, thus Eleanor is also able to operate under the Linear MDP assumption.

References

The UCRL-VTR paper.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., & Yang, L. (2020, November). Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning (pp. 463-474). PMLR. [link]

The UCRL-VTR$^+$ paper. It also shows the regret lower bound for linear mixture MDPs $\Omega(dH^{3/2}\sqrt{K})$.

Zhou, D., Gu, Q., & Szepesvari, C. (2021, July). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Conference on Learning Theory (pp. 4532-4576). PMLR. [link]

The LSVI-UCB paper.

Jin, C., Yang, Z., Wang, Z., & Jordan, M. I. (2020, July). Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory (pp. 2137-2143). PMLR. [link]

The Eleanor paper.

Zanette, A., Lazaric, A., Kochenderfer, M., & Brunskill, E. (2020, November). Learning near optimal policies with low inherent bellman error. In International Conference on Machine Learning (pp. 10978-10989). PMLR. Link