Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes

Dongruo Zhou, Quanquan Gu, Csaba Szepesvári

Presenter: Hao Qin

Outline

Introduction to the linear mixture model setting

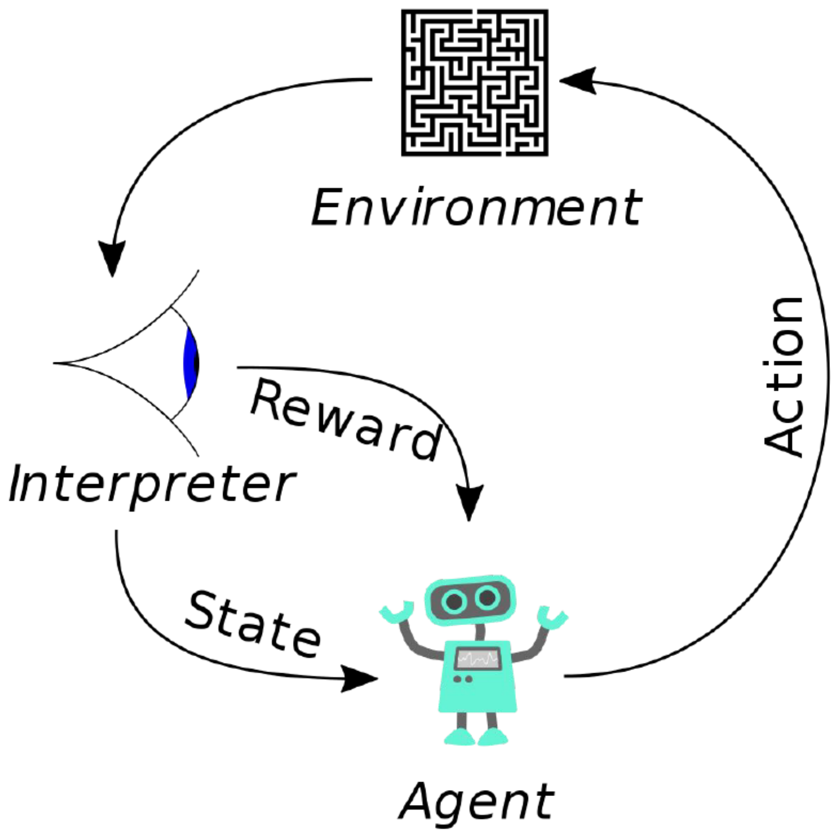
Value target regression and UCRL-VTR algorithm

Weighted regression model

Final regret analysis

Summary

Online Reinforcement Learning



Online Reinforcement Learning

Application: game, autonomous driving, dialogue system (Siri) ...

Online learning Time-inhomogeneous Episodic MDPs

$$M(\underbrace{s \in \mathcal{S}}_{\text{state space}}, \underbrace{a \in \mathcal{A}}_{\text{action space}}, \underbrace{H}_{\text{episode length}}, \underbrace{(r_h(s,a))_h}_{\text{reward functions}}, \underbrace{(\mathbb{P}_h(s'|s,a))_h}_{\text{transition dynamics}})$$

For k from 1 to K, starting from the initial state s_0^k,

Deterministic policy $\pi^k = \left(\pi_h^k\right)_{h=1}^H$

For h from 1 to H,

Select action $a_h^k \leftarrow \pi_h^k(s_h^k)$

Observe returned r_h^k reward and next-state s_{h+1}^k

Online learning Time-inhomogeneous Episodic MDPs

Objective

Minimize the regret

$$Regret(M,K) = \sum_{k=1}^{K} [V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)]$$

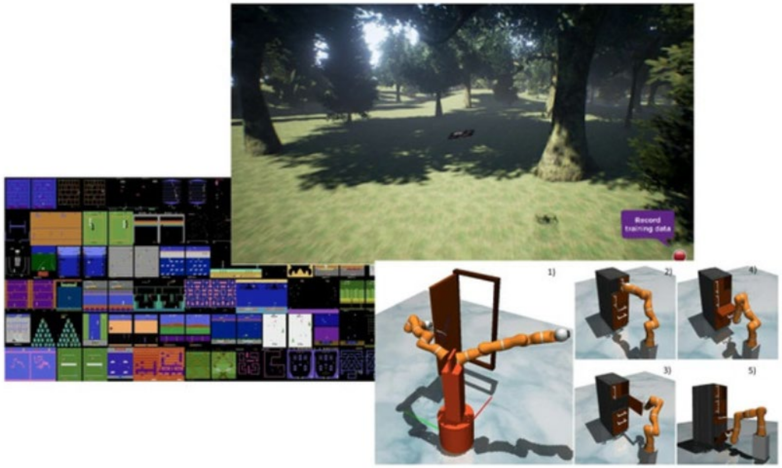
Where $V_h^\pi(s)$ and $V_h^*(s)$ have been defined as

$$V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)),\ Q_h^\pi(s,a) = \mathbb{E}_{\pi,h,s,a}\left[\sum_{h'=h}^{H} r_h(s_{h'}, a_{h'})\right],\ V_{H+1}^\pi(s) = 0,$$

$$V_h^*(s) = \sup_\pi V_h^\pi(s),\ Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a)$$

RL with Linear Function Approximation

In reality, the feature space ($space \times action$) can be quite large



Linear mixture MDPs

$$P_h(s'|s,a) = \sum_{j=1}^{d} \theta_{h,j} \phi_j(s'|s,a)$$

RL with Linear Function Approximation

Linear mixture MDPs model assumption

1. Linear Mixture model

$$P_h(s'|s,a) = \sum_{j=1}^{d} \theta_{h,j} \phi_j(s'|s,a)$$

2. Reward function $r_h(s,a)$ is known.

3. Known transition model family $\mathcal{P}$, where $\phi_1, \phi_2, \cdots, \phi_d \in \Phi$

RL with Linear Function Approximation

Linear mixture MDPs model assumption

1. Linear Mixture model

$$\mathbb{P}_h(s'|s,a) = \langle \phi(s'|s,a), \theta_h^\star \rangle \quad \|\theta_h^*\|_2 \le B$$

$$\phi_V(s,a) = \sum_{s' \in \mathcal{S}} \phi(s'|s,a) V(s'), \forall V: \mathcal{S} \to [0,1], (s,a) \in \mathcal{S} \times \mathcal{A}, \|\phi_V(s,a)\|_2 \le 1$$

$$[\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) = \left\langle \sum_{s' \in \mathcal{S}} \phi(s'|s_h^k, a_h^k) V_{k,h+1}(s'), \theta_h^\star \right\rangle = \left\langle \phi_{V_{k,h+1}}(s'|s,a), \theta_h^\star \right\rangle$$

2. Rewa

3. Known transition model family Φ, where $\varphi_1, \varphi_2, \cdots, \varphi_d \in \Phi$

Example:

Tabular MDPs $d = S^2 A, \phi(s'|s,a) = e_{s,a,s'} \in \mathbb{R}^d, \theta_h^* = [\mathbb{P}_h(s'|s,a)]_{s,a,s'} \in \mathbb{R}^d$

Outline

Introduction to the linear mixture model setting

Value target regression and UCRL-VTR algorithm

Weighted regression model

Final regret analysis

Summary

Upper Bound: From Linear Bandits to Linear Mixture MDPs

- For linear mixture MDPs, only need to estimate the unknown $\theta_h^\star$

- Key observation For function $V, \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} V(s') = \langle \theta_h^*, \phi_V(s,a) \rangle$

Value-targeted regression (VTR):

With enough number of pairs $(s_h^k, a_h^k, s_{h+1}^k, V_{k,h+1}), s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k, a_h^k)$

estimating θ_h^* is possible by doing regression over $(\phi_{V_{k,h+1}}(s_h^k, a_h^k), V_{k,h+1}(s_{h+1}^k))$

$[\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) = \left\langle \phi_{V_{k,h+1}}(s'|s,a), \theta_h^\star \right\rangle$

- Example: **UCRL-VTR** (Jia et al., 2020) : Value-target regression + OFUL
- Theorem (Jia et al., 2020): UCRL-VTR enjoys regret

$$\text{Regret}(M, K) = \widetilde{O}(dH^2\sqrt{K})$$

RL with Linear Function Approximation

Notation

$$\phi_{V_{k,h+1}}(s_h^k, a_h^k) = \phi_{k,h+1}$$

$$\phi_{V_{k,h+1}^2}(s_h^k, a_h^k) = \chi_{k,h+1}$$

For example

$$\mathbb{E}_{s' \sim \mathbb{P}_h\left(\cdot \middle| s_h^k, a_h^k\right)} V_{k,h+1}(s') = \left\langle \theta_h, \phi_{V_{k,h+1}}(s_h^k, a_h^k) \right\rangle = \langle \theta_h, \phi_{k,h+1} \rangle$$

UCRL-VTR algorithm

At kth episode, the agent

- Receive the initial state. At hth stage,
- Solution to ridge regression over pairs:

$$\left(\phi_{k,h+1}, V_{k,h+1}(s_{h+1}^k)\right)_k \to \hat{\theta}_{k,h}$$

$$\hat{\theta}_{k,h} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, \lambda \|\theta\|_2^2 + \sum_{j=1}^{k-1} \left[\langle \phi_{j,h+1}, \theta \rangle - V_{j,h+1}\left(s_{h+1}^j\right)\right]^2$$

- Construct confidence sets

$$\hat{\mathcal{C}}_{k,h} = \left\{\theta : \left\| \hat{\Sigma}_{k,h}^{\frac{1}{2}} \left(\theta - \hat{\theta}_{k,h}\right) \right\|_2 \le \hat{\beta}_k \right\}, \hat{\Sigma}_{k,h} = \lambda \boldsymbol{I} + \sum_{j=1}^{k-1} \phi_{j,h+1} \phi_{j,h+1}^T$$

- Estimate previous value function by Bellman equation

$$Q_{k,h}(\cdot,\cdot) = \left[r_h(\cdot,\cdot) + \max_{\theta \in \hat{\mathcal{C}}_{k,h}} \left\langle \theta, \phi_{V_{k,h+1}}(\cdot,\cdot) \right\rangle \right]_{[0,H]}, V_{k,h}(\cdot) = \max_{\alpha \in \mathcal{A}} Q_{k,h}(\cdot, a)$$

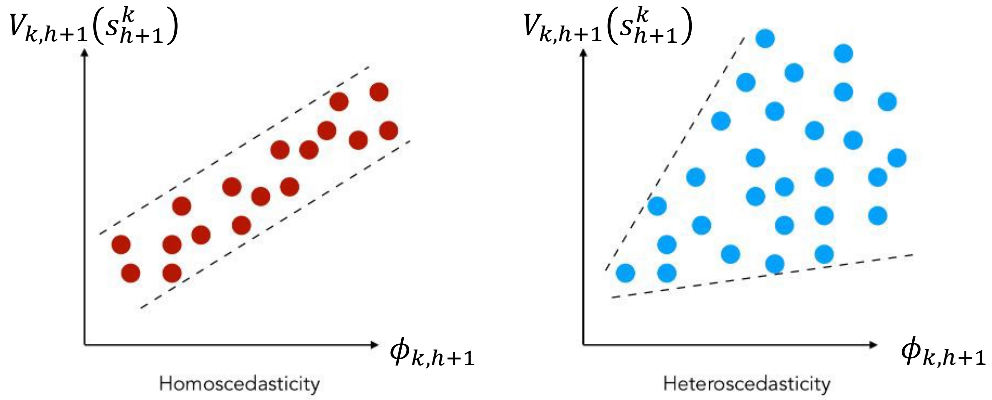
- Select action $a_h^k = \arg\max_{a \in \mathcal{A}} Q_{k,h}\left(s_h^k, a\right), h = 1, \cdots$

UCRL-VTR algorithm

Shortcomings of UCRL-VTR

$$\langle \theta_h^*, \phi_{k,h+1} \rangle + \varepsilon_k = V_{k,h+1}(s_{h+1}^k)$$

- Choose β_k(the radius of confidence set $\hat{\mathcal{C}}_{k,h}$) proportional to the magnitude of the value function $V_{k,h+1}(\cdot)$ rather than its variance $[\mathbb{V}_h V_{k,h+1}](\cdot,\cdot)$
- In the heteroscedastic model, value functions $[\mathbb{V}_h V_{k,h+1}](\cdot,\cdot)$ cannot be bounded uniformly at different stage, hence we need to make some adjustments such as using weighted least-square estimator.



Bernstein-Type Self Normalized Inequality

- **Theorem 2 (Bernstein inequality for vector-valued martingales)**

If there is stochastic process $\{x_t, \eta_t\}_{t\geq 1}$ *and a linear relationship that* $y_t = \langle \mu^*, x_t \rangle + \eta_t$. *Also,* η_t, x_t *satisfy that*

$$|\eta_t| \leq R, \quad \mathbb{E}[\eta|\mathcal{G}_t] = 0, \quad \|x_t\|_2 \leq X \quad \text{and} \quad \mathbb{E}\left[\eta_t^2 \middle| \mathcal{G}_t\right] \leq \sigma^2$$

We have

$$\forall T, \quad \|\mu^* - \hat{\mu}_t\|_{A_t} \leq \hat{\beta}_t + \sqrt{\lambda}\|\mu^*\|_2, \qquad \hat{\beta}_t = \tilde{O}\left(\sigma\sqrt{d} + R\right)$$

Where $A_t = \lambda \boldsymbol{I} + \sum_{i=1}^{t} x_i x_i^T$, $\mu^t = A_t^{-1}(\sum_{i=1}^{t} y_i x_i)$ *and* $\lambda > 0$.

$$\hat{\beta}_t = \tilde{O}\left(R\sqrt{d}\right) \rightarrow \tilde{O}\left(\sigma\sqrt{d} + R\right)$$

- Can be extended to sub-exponential random variable case
- Strict improvement from Abbasi-Yadkori et al. (2011): $\hat{\beta}_t = \tilde{O}(R\sqrt{d})$
- Following induction proof by Dani et al. (2008)

Outline

Introduction to the linear mixture model setting

Value target regression and UCRL-VTR algorithm

Weighted regression model

Final regret analysis

Summary

Warmup: Heteroscedastic Linear Bandits

For t from 1 to T,

Receive the decision set: $\mathcal{D}_t$

Select an action $\alpha_t \in \mathcal{D}_t$ based on past observation

Observe reward r_t and variance σ_t

$$r_t = \langle \boldsymbol{\mu}^*, \mathbf{a}_t \rangle + \epsilon_t,\ |\epsilon_t| \leq R,\ \mathbb{E}[\epsilon_t | \mathbf{a}_{1:t}, \epsilon_{1:t-1}] = 0,\ \mathbb{E}[\epsilon_t^2 | \mathbf{a}_{1:t}, \epsilon_{1:t-1}] \leq \sigma_t^2$$

Objective: Minimize the regret

$$R(T) = \sum_{t=1}^{T} \sup_{\mathbf{a} \in \mathcal{D}_t} \langle \mathbf{a}, \boldsymbol{\mu}^* \rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^* \rangle$$

OFUL and weighted OFUL methods

OFUL (Abbasi-Yadkori et al., 2011): ‘Optimistic’ estimation of reward

Compute the ridge regression over predictor-response pairs:

$$\hat{\mu}_t \leftarrow arg \min_{\mu \in \mathbb{R}^d} \lambda \|\mu\|_2^2 + \sum_{i=1}^{t} [\langle \mu, \alpha_i \rangle - r_i]^2$$

Weighted OFUL:

Solution to ridge regression over **weighted** predictor-response pairs $\left(\frac{\alpha_t}{\bar{\sigma}_t}, \frac{r_t}{\bar{\sigma}_t}\right) \rightarrow \hat{\mu}_t$

$$\hat{\mu}_t \leftarrow arg \min_{\mu \in \mathbb{R}^d} \lambda \|\mu\|_2^2 + \sum_{i=1}^{t} \frac{[\langle \mu, \alpha_i \rangle - r_i]^2}{\bar{\sigma}_i^2}$$

$\bar{\sigma}_t = \max\{\alpha, \sigma_t\} \neq 0$

OFUL and weighted OFUL methods

OFUL (Abbasi-Yadkori et al., 2011):

Theorem : *the regret of OFUL is*

$$\widetilde{O}(dR\sqrt{T})$$

Weighted OFUL:

Theorem 3 *Weighted OFUL enjoys the regret*

$$R(T) = \widetilde{O}\left(R\sqrt{dT} + d\sqrt{\sum_{t=1}^{T}\sigma_t^2}\right)$$

'Nearly' independent of T when $\sum_{t=1}^{T}\sigma_t^2 \geq \frac{TR^2}{d}$

Strictly improves $\tilde{O}(dR\sqrt{T})$.

UCRL-VTR+ algorithm

At kth episode, the agent

- Receive the initial state. At hth stage,
- Solution to ridge regression over weighted pairs:

$$\left(\frac{\phi_{k,h+1}}{\bar{\sigma}_{k,h}}, \frac{V_{k,h+1}}{\bar{\sigma}_{k,h}}\right)_k \rightarrow \hat{\theta}_{k,h}$$

$$\hat{\theta}_{k,h} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \lambda \|\theta\|_2^2 + \sum_{j=1}^{k-1} \frac{\left[\langle \phi_{j,h+1}, \theta \rangle - V_{j,h+1}\left(s_{h+1}^j\right)\right]^2}{\bar{\sigma}_{k,h}^2}$$

- $\boxed{\text{Construct confidence sets } \hat{\mathcal{C}}_{k,h} \text{ with diameter } \hat{\beta}_k}$

- Estimate previous value function by Bellman equation

$$Q_{k,h}(\cdot,\cdot) = \left[r_h(\cdot,\cdot) + \max_{\theta \in \hat{\mathcal{C}}_{k,h}} \left\langle \theta, \phi_{V_{k,h+1}}(\cdot,\cdot) \right\rangle\right]_{[0,H]}, V_{k,h}(\cdot) = \max_{a \in \mathcal{A}} Q_{k,h}(\cdot, a)$$

- Select action $a_h^k = \arg\max_{a \in \mathcal{A}} Q_{k,h}\left(s_h^k, a\right), h = 1, \cdots$

UCRL-VTR+ algorithm

Three major problems

- Q1: How to calculate the empirical variance $[\widehat{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k)$?

The variance $[\mathbb{V}_{k,h}V_{k,h+1}](s_h^k, a_h^k)$ has been defined as

$$\left[\mathbb{V}_{k,h}V_{k,h+1}\right]\left(s_h^k, a_h^k\right) \coloneqq \mathbb{E}_{s' \sim P\left(\cdot \middle| s_h^k, a_h^k\right)}\left[V_{k,h+1}(s') - (\mathbb{P}V_{k,h+1})(s_h^k . a_h^k)\right]^2$$

- Q2 How to select $E_{k,h}$ to guarantee that $\left[\widehat{\mathbb{V}}_{k,h}V_{k,h+1}\right]\left(s_h^k, a_h^k\right) + E_{k,h}$ upper bound $\sigma_{k,h}^\star$ w.h.p.?

- Q3 : How to choose an appropriate $\hat{\beta}_k$, such that $\hat{\mathcal{C}}_{k,h}$ contains $\theta_h^\star$ w.h.p.?

UCRL-VTR+ algorithm

A1: Variance Estimator

- Unlike bandit setting, we need to estimate the variance

$$
\begin{aligned}
[\mathbb{V}V](s,a) &= \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}V^2(s') - \left[\mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}V(s')\right]^2 \\
&= \underbrace{\langle\theta_h^*,\chi\rangle}_{\langle\tilde{\theta}_h,\chi\rangle} - \underbrace{[\langle\theta_h^*,\phi\rangle}_{\langle\hat{\theta}_h,\phi\rangle}]^2
\end{aligned}
$$

- Regression over predictor-response $\left(\chi_{k,h+1}, V_{k,h+1}^2(s_{h+1}^k)\right)$

$$
\tilde{\theta}_{k,h} = arg\min_{\theta\in\mathbb{R}^d} \lambda\|\theta\|_2^2 + \sum_{j=1}^{k-1}\left[\langle\chi_{j,h+1},\theta\rangle - V_{j,h+1}^2\left(s_{h+1}^j\right)\right]^2
$$

- Variance estimator with clip

$$
\bar{\mathbb{V}}_{k,h}V_{k,h+1}(s_h^k,a_h^k) = \left[\langle\chi_{k,h+1},\tilde{\theta}_{k,h+1}\rangle\right]_{[0,H^2]} - \left[\left[\langle\phi_{k,h+1},\hat{\theta}_{k,h+1}\rangle\right]^2\right]_{[0,H^2]}
$$

UCRL-VTR+ algorithm

A2: Bonus for variance estimation

- Bonus $E_{k,h}$ has been defined as follows:

$$E_{k,h} = \min\left\{H^2, 2H\check{\beta}_k \left\|\hat{\Sigma}_{k,h}^{-\frac{1}{2}}\phi_{k,h+1}\right\|_2\right\} + \min\left\{H^2, \tilde{\beta}_k \left\|\hat{\Sigma}_{k,h}^{-\frac{1}{2}}\chi_{k,h+1}\right\|_2\right\}$$

$$\hat{\beta}_k = \tilde{O}\left(\sqrt{d}\right), \check{\beta}_k = \tilde{O}(d),\ \tilde{\beta}_k = \tilde{O}(H^2\sqrt{d})$$

Final variance upper bound: $\bar{\sigma}_{k,h} \leftarrow \sqrt{\max\{\frac{H^2}{d}, \bar{\mathbb{V}}_{k,h}V_{k,h+1} + E_{k,h}\}}$

Then w.h.p., for all $\mathrm{k} \in [\mathrm{K}]$ and $\mathrm{h} \in [\mathrm{H}]$

$$\left|\bar{\mathbb{V}}_{k,h}V_{k,h+1} - \mathbb{V}_{k,h}V_{k,h+1}\right| \leq E_{k,h}, \sum_{k=1}^{K}\sum_{h=1}^{H} E_{k,h} = \tilde{O}(d^{\frac{3}{2}}H^3\sqrt{K})$$

UCRL-VTR+ algorithm

A3: Confidence set

- Lemma 5

The confidence set has been defined as

$$\hat{\mathcal{C}}_{k,h} = \left\{\theta : \left\| \hat{\Sigma}_{k,h}^{\frac{1}{2}} \left(\theta - \hat{\theta}_{k,h}\right) \right\|_2 \leq \hat{\beta}_k \right\}, \hat{\Sigma}_{k,h} = \lambda \boldsymbol{I} + \sum_{i=1}^{k-1} \frac{\phi_{j,h+1} \phi_{j,h+1}^T}{\bar{\sigma}_{j,h}^2}$$

Where $\hat{\beta}_k = \tilde{O}(\sqrt{d})$

UCRL-VTR+ algorithm

At kth episode, the agent

- Receive the initial state. At hth stage,
- Solution to ridge regression over weighted pairs:

$$\left(\frac{\phi_{k,h+1}}{\bar{\sigma}_{k,h}}, \frac{V_{k,h+1}\left(s_{h+1}^{k}\right)}{\bar{\sigma}_{k,h}}\right)_{k} \to \hat{\theta}_{k,h}$$

$$\hat{\theta}_{k,h} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}}\, \lambda \|\theta\|_2^2 + \sum_{j=1}^{k-1} \left[\langle \phi_{j,h+1}, \theta \rangle - V_{j,h+1}\left(s_{h+1}^{j}\right)\right]^2 / \boxed{\bar{\sigma}_{j,h}^2}$$

- Construct confidence sets

$$\hat{\mathcal{C}}_{k,h} = \left\{\theta : \left\|\hat{\Sigma}_{k,h}^{\frac{1}{2}}\left(\theta - \hat{\theta}_{k,h}\right)\right\|_2 \leq \hat{\beta}_k\right\}, \hat{\Sigma}_{k,h} = \lambda \boldsymbol{I} + \sum_{j=1}^{k-1} \phi_{j,h+1} \phi_{j,h+1}^{T} / \boxed{\bar{\sigma}_{j,h}^2}$$

- Estimate previous value function by Bellman equation

$$Q_{k,h}(\cdot,\cdot) = \left[r_h(\cdot,\cdot) + \max_{\theta \in \hat{\mathcal{C}}_{k,h}} \left\langle \theta, \phi_{V_{k,h+1}}(\cdot,\cdot)\right\rangle\right]_{[0,H]}, V_{k,h}(\cdot) = \max_{\alpha \in \mathcal{A}} Q_{k,h}(\cdot, a)$$

- Select action $a_h^k = \arg\max_{a \in \mathcal{A}} Q_{k,h}\left(s_h^k, a\right), h = 1, \cdots$

Outline

Introduction to the linear mixture model setting

Value target regression and UCRL-VTR algorithm

Weighted regression model

Final regret analysis

Summary

UCRL-VTR+ algorithm

Upper Bound: Regret Decomposition

$$V_{k,h}(s_h^k) - V_h^{\pi^k}(s_h^k)$$

$$\leq \langle \hat{\theta}_{k,h}, \phi_{k,h+1} \rangle + \hat{\beta}_k \left\| \hat{\Sigma}_{k,h}^{-\frac{1}{2}} \phi_{k,h+1} \right\|_2 - [\mathbb{P}_h V_{k.h+1}^{\pi^k}](s_h^k, a_h^k)$$

$$\leq [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - \left[\mathbb{P}_h V_{k.h+1}^{\pi^k}\right](s_h^k, a_h^k) + 2\hat{\beta}_k \left\| \hat{\Sigma}_{k,h}^{-\frac{1}{2}} \phi_{k,h+1} \right\|_2$$

$$\leq [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - \left[\mathbb{P}_h V_{k.h+1}^{\pi^k}\right](s_h^k, a_h^k) + 2\hat{\beta}_k \bar{\sigma}_{k,h} \left\| \hat{\Sigma}_{k,h}^{-\frac{1}{2}} \frac{\phi_{k,h+1}}{\bar{\sigma}_{k,h}} \right\|_2$$

Summing over $h \in]H]$ and $k \in [K]$

$$\sum_{k=1}^{K} V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)$$

$$\leq \sum_{k=1}^{K} \sum_{h=1}^{H} 2\hat{\beta}_k \bar{\sigma}_{k,h} \left\| \hat{\Sigma}_{k,h}^{-\frac{1}{2}} \frac{\phi_{V_{k,h+1}}}{\bar{\sigma}_{k,h}} \right\|_2 + \tilde{O}\left(H^2\sqrt{dK}\right)$$

$$\leq 2\hat{\beta}_K \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \bar{\sigma}_{k,h}^2} \cdot \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \left\| \hat{\Sigma}_{k,h}^{-\frac{1}{2}} \frac{\phi_{V_{k,h+1}}}{\bar{\sigma}_{k,h}} \right\|_2^2} + \tilde{O}(H^2\sqrt{dK})$$

UCRL-VTR+ algorithm

Naïve approach:

- Bound variance by H^2

$$Regret(M,K) = \tilde{O}\left(H^2\sqrt{dK} + d\sqrt{H}\left[\sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{\mathbb{V}}\hat{V}_{h+1}^{\pi^k}\left(s_h^k, a_h^k\right)\right]^{\frac{1}{2}}\right)$$

$$\leq \tilde{O}\left(H^2\sqrt{dK} + \mathrm{d}\sqrt{H}\cdot\left(HK\cdot H^2\right)^{\frac{1}{2}}\right)$$

Law of total variance

- Weighted OFUL suggests a regret of UCRL-VTR+

$$Regret(M,K) = \tilde{O}\left(H^2\sqrt{dK} + d\sqrt{H}\left[\sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{\mathbb{V}}\hat{V}_{h+1}^{\pi^k}\left(s_h^k, a_h^k\right)\right]^{\frac{1}{2}}\right)$$

$$\leq \tilde{O}\left(H^2\sqrt{dK} + \mathrm{d}\sqrt{H}\cdot\left(H^2K\right)^{\frac{1}{2}}\right)$$

$$\leq \tilde{O}\left(\sqrt{d^2H^3 + dH^4}\sqrt{K}\right)$$

- Use Law of total variance (Lattimore and Hutter, 2012; Azar et al., 2017)

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}V_{h+1}^{\pi^k}(s') = \tilde{O}\left(H^2K\right)$$

UCRL-VTR+ algorithm

Regret of UCRL-VTR+

Theorem 6 (Upper bound) *w.h.p., UCRL-VTR+ attains regret*

$$\text{Regret}(M, K) = \widetilde{O}\Big(\sqrt{d^2H^3 + dH^4}\sqrt{K} + d^2H^3 + d^3H^2\Big)$$

Theorem 8 (Lower bound) *Let B>1, for any RL algorithm, there exists a time-inhomogeneous episodic B-linear mixture MDP M, where*

$$\mathbb{E}[\text{Regret}(M, K)] \geq \Omega(dH^{3/2}\sqrt{K})$$

The regret of UCRL-VTR+ matches the lower bound when $d \geq H,\ K \geq d^4H + d^3H^2$

Outline

Introduction to the linear mixture model setting

Value target regression and UCRL-VTR algorithm

Weighted regression model

Final regret analysis

Summary

Summary

- Nearly minimax optimal RL utilize a variance-dependent weighted linear regression
- Optimistic estimation of value function
- Combining law of total variance to reach a lower bound

Thank you