Reinforcement Learning with Linear Function Approximation

Lizhang Miao

August 20, 2020

Review of "Zanette A, Lazaric A, Kochenderfer M, et al. Learning Near Optimal Policies with Low Inherent Bellman Error, 2020."

Outline

Linear Bandits
- UCB
- Techniques from Linear Bandits

Episodic RL with Linear approximation
- Settings
- Proofs

Linear Bandits

Formulation

- Bandits: K-arms; $\rightarrow$ Linear bandits: action vector $a_t \in \mathbb{R}^d$, observed reward $r_t = \langle a_t, \theta^* \rangle + \eta_t$, η_t is zero-mean noise

Formulation

- Bandits: K-arms; $\rightarrow$ Linear bandits: action vector $a_t \in \mathbb{R}^d$, observed reward $r_t = \langle a_t, \theta^* \rangle + \eta_t$, η_t is zero-mean noise
- Contextual bandits: contextual information c_t and action $a_t \in [K]$, reward $r_t = T(c_t, a_t) + \eta_t \rightarrow$ Contextual linear bandits: feature map $\psi : C \times [K] \rightarrow \mathbb{R}^d$, reward $r_t(c_t, a_t) = \langle \psi(c_t, a_t), \theta^* \rangle + \eta_t$

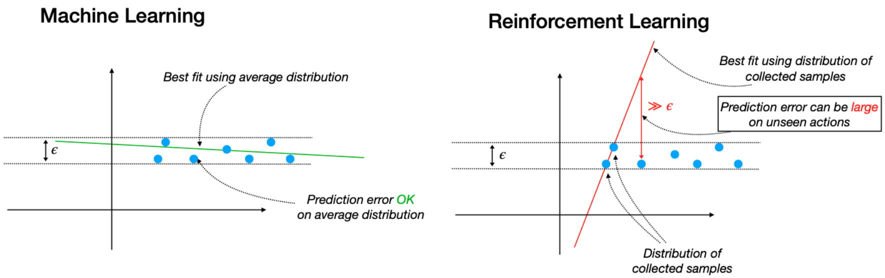
Formulation

- Bandits: K-arms; $\rightarrow$ Linear bandits: action vector $a_t \in \mathbb{R}^d$, observed reward $r_t = \langle a_t, \theta^* \rangle + \eta_t$, η_t is zero-mean noise
- Contextual bandits: contextual information c_t and action $a_t \in [K]$, reward $r_t = T(c_t, a_t) + \eta_t \rightarrow$ Contextual linear bandits: feature map $\psi : C \times [K] \rightarrow \mathbb{R}^d$, reward $r_t(c_t, a_t) = \langle \psi(c_t, a_t), \theta^* \rangle + \eta_t$
- Regret: $R_n = \mathbb{E}[\sum_{t=1}^n r_t^* - r_t]$
- Regret bound: linear bandits $\tilde{O}(d\sqrt{n})$; contextual linear bandits $\tilde{O}(\sqrt{dn})$

Formulation

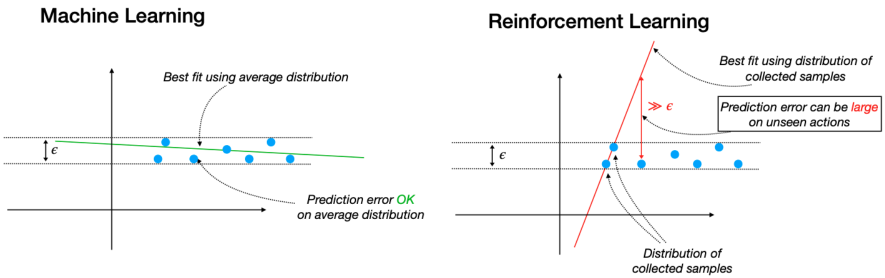
- Bandits: K-arms; $\rightarrow$ Linear bandits: action vector $a_t \in \mathbb{R}^d$, observed reward $r_t = \langle a_t, \theta^* \rangle + \eta_t$, η_t is zero-mean noise
- Contextual bandits: contextual information c_t and action $a_t \in [K]$, reward $r_t = T(c_t, a_t) + \eta_t$ $\rightarrow$ Contextual linear bandits: feature map $\psi : C \times [K] \rightarrow \mathbb{R}^d$, reward $r_t(c_t, a_t) = \langle \psi(c_t, a_t), \theta^* \rangle + \eta_t$
- Regret: $R_n = \mathbb{E}[\sum_{t=1}^n r_t^* - r_t]$
- Regret bound: linear bandits $\tilde{O}(d\sqrt{n})$; contextual linear bandits $\tilde{O}(\sqrt{dn})$
- Linear approximation and exploration in RL? Transition model?

Exploration



- Core problem: Exploration-Exploitation trade-off, especially model misspecification

Exploration



- Core problem: Exploration-Exploitation trade-off, especially model misspecification
- Exploitation: fit collected data
- Explore with a confidence ball: Upper Confidence Bound algorithm which is near-minmax optimal in bandits

LinUCB

- Construct confidence set $\mathcal{C}_t$ based on collected data $(a_1, r_1, \cdots, a_{t-1}, r_{t-1})$ that contains unknown parameter θ^* with high probability

LinUCB

- Construct confidence set $\mathcal{C}_t$ based on collected data $(a_1, r_1, \cdots, a_{t-1}, r_{t-1})$ that contains unknown parameter θ^* with high probability
- For a fixed $\hat{\theta}$, the set can be constructed as $\mathcal{C}_t = \{\theta \in \mathbb{R}^d | \|\theta - \hat{\theta}\|_V^2 \leq \beta\}$ where V is positive definite and $\|x\|_V = \sqrt{x^T V x}$

LinUCB

- Construct confidence set $\mathcal{C}_t$ based on collected data $(a_1, r_1, \cdots, a_{t-1}, r_{t-1})$ that contains unknown parameter θ^* with high probability
- For a fixed $\hat{\theta}$, the set can be constructed as $\mathcal{C}_t = \{\theta \in \mathbb{R}^d | \|\theta - \hat{\theta}\|_V^2 \leq \beta\}$ where V is positive definite and $\|x\|_V = \sqrt{x^T V x}$
- With a unit ball $\mathcal{B}_2 = \{x \in \mathbb{R}^d | \|x\|_2 \leq 1\}$, $\mathcal{C}_t = \hat{\theta} + \beta^{1/2} V^{-1/2} \mathcal{B}_2$
- $\bar{r}_t(a) = \langle a_t, \hat{\theta} \rangle + \beta^{1/2} \|a\|_{V^{-1}} \geq r_t^*(a)$ with selection of β

LinUCB

- Construct confidence set $\mathcal{C}_t$ based on collected data $(a_1, r_1, \cdots, a_{t-1}, r_{t-1})$ that contains unknown parameter θ^* with high probability
- For a fixed $\hat{\theta}$, the set can be constructed as $\mathcal{C}_t = \{\theta \in \mathbb{R}^d | \|\theta - \hat{\theta}\|_V^2 \leq \beta\}$ where V is positive definite and $\|x\|_V = \sqrt{x^T V x}$
- With a unit ball $\mathcal{B}_2 = \{x \in \mathbb{R}^d | \|x\|_2 \leq 1\}$, $\mathcal{C}_t = \hat{\theta} + \beta^{1/2} V^{-1/2} \mathcal{B}_2$
- $\bar{r}_t(a) = \langle a_t, \hat{\theta} \rangle + \beta^{1/2} \|a\|_{V^{-1}} \geq r_t^*(a)$ with selection of β
- Exploitation: least square value iteration for $\hat{\theta}$
- Exploration: parameter space confidence ball $\rightarrow$ adding exploration bonus
- *Regret* $\leq \mathbb{E}[\sum_{t=1}^n \bar{r}_t - r_t]$

Technical lemmas

Lemma (Self-normalized bound for vector-valued martingales)

Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{x_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that $x_t|\mathcal{F}_{t-1}$ is σ-subGaussian. Assume V_0 is a $d \times d$ positive definite matrix, and let $V_t = V_0 + \sum_{s=1}^{t} \phi_s \phi_s^T$. Then with probability at least $1-\delta$, we have

$$\left\| \sum_{s=1}^{t} \phi_s x_s \right\|_{V_t^{-1}}^2 \leq 2\sigma^2 \log[\det(V_t)^{1/2} \det(V_0)^{-1/2}/\delta].$$

Technical lemmas

Lemma (Self-normalized bound for vector-valued martingales)

Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{x_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that $x_t|\mathcal{F}_{t-1}$ is σ-subGaussian. Assume V_0 is a $d \times d$ positive definite matrix, and let $V_t = V_0 + \sum_{s=1}^{t} \phi_s \phi_s^T$. Then with probability at least $1-\delta$, we have

$$\left\| \sum_{s=1}^{t} \phi_s x_s \right\|_{V_t^{-1}}^2 \leq 2\sigma^2 \log[\det(V_t)^{1/2} \det(V_0)^{-1/2}/\delta].$$

Lemma (Determinant-Trace Inequality)

Suppose $X_1, X_2, \cdots, X_t \in \mathbb{R}^d$ and for any $1 < s < t$, $\|X_s\|_2 \leq L$. Let $V_t = \lambda I + \sum_{s=1}^{t} X_s X_s^T$ for some $\lambda > 0$. Then,

$$det(V_t) \leq (\lambda + tL^2/d)^d$$

.

Episodic RL with Linear approximation

Episodic RL Notations

- Undiscounted finite-horizon MDP: $M = (\mathcal{S}, \mathcal{A}, p, r, H)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel p_t, reward function r and horizon length H.
- V-value: $V_t^\pi : \mathcal{S} \to \mathbb{R}$ is the expected value of cumulative rewards received under policy π when starting from an arbitrary state at the hth step

$$V_t^\pi(x) = \mathbb{E}\left[\sum_{t'=t}^{H} r_{t'}(s_{t'}, \pi_{t'}(s_{t'}))|x_t = x\right], \qquad \forall s \in \mathcal{S}, t \in [H].$$

  Optimal value $V_t^*(s) = \sup_\pi V_t^\pi(s)$ for all $s \in \mathcal{S}$ and $t \in [H]$.
- Q-value: $Q_t^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ gives the expected value of cumulative rewards when the agent starts from an arbitrary state-action pair at the tth step and follows policy π afterwards

$$Q_t^\pi(x, a) = r_t(x, a) + \mathbb{E}\left[\sum_{l=t+1}^{H} r_l(s_l, \pi_l(s_l))|s_l = s, a_l = a\right], \forall (s, a) \in \mathcal{S}\times\mathcal{A}, t \in [H].$$

Notations (Cont.)

- Bellman equation associated with a policy π becomes:

$$V_t^\pi(s) = Q_t^\pi(s, \pi_t(s)),$$
$$Q_t^\pi(s, a) = (r_t + \mathbb{P}_t V_{t+1}^\pi)(s, a).$$

Notations (Cont.)

- Bellman equation associated with a policy π becomes:

$$\begin{aligned} V_t^\pi(s) &= Q_t^\pi(s, \pi_t(s)), \\ Q_t^\pi(s,a) &= (r_t + \mathbb{P}_t V_{t+1}^\pi)(s,a). \end{aligned}$$

- Bellman optimality equation

$$\begin{aligned} V_t^*(s) &= \max_{a \in \mathcal{A}} Q_t^*(s,a), \\ Q_t^*(s,a) &= (r_t + \mathbb{P}_h V_{t+1}^*)(x,a). \end{aligned}$$

Notations (Cont.)

- Bellman equation associated with a policy π becomes:

$$\begin{aligned} V_t^\pi(s) &= Q_t^\pi(s, \pi_t(s)), \\ Q_t^\pi(s, a) &= (r_t + \mathbb{P}_t V_{t+1}^\pi)(s, a). \end{aligned}$$

- Bellman optimality equation

$$\begin{aligned} V_t^*(s) &= \max_{a \in \mathcal{A}} Q_t^*(s, a), \\ Q_t^*(s, a) &= (r_t + \mathbb{P}_h V_{t+1}^*)(x, a). \end{aligned}$$

- Bellman operator $\mathcal{T}$ applied to Q_{t+1} is defined as

$$\mathcal{T}_t(Q_{t+1})(s, a) = r_t(s, a) + \mathbb{E}_{s' \sim p_t(s,a) \max_{a'} Q_{t+1}(s', a')}$$

Linear Value Function

- Feature map: $\phi_t : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^{d_t}$
- $Q_t(s, a) = \phi_t(s, a)^T \theta_t$

Linear Value Function

- Feature map: $\phi_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_t}$
- $Q_t(s, a) = \phi_t(s, a)^T \theta_t$
- Define space of parameters inducing uniformly bounded action-value functions

$$\mathcal{B}_t = \{\theta_t \in \mathbb{R}^{d_t} || \phi_t(s, a)^T \theta_t| \leq D, \forall (s, a)\}$$

- Each parameter θ identifies an (action) value function

$$Q_t(\theta_t)(s, a) = \phi_t(s, a)^T \theta_t, \qquad V_t(\theta_t) = \max_a \phi_t(s, a)^T \theta_t$$

Linear Value Function

- Feature map: $\phi_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_t}$
- $Q_t(s, a) = \phi_t(s, a)^T \theta_t$
- Define space of parameters inducing uniformly bounded action-value functions

$$\mathcal{B}_t = \{\theta_t \in \mathbb{R}^{d_t} | |\phi_t(s, a)^T \theta_t| \leq D, \forall (s, a)\}$$

- Each parameter θ identifies an (action) value function

$$Q_t(\theta_t)(s, a) = \phi_t(s, a)^T \theta_t, \qquad V_t(\theta_t) = \max_a \phi_t(s, a)^T \theta_t$$

- So consider function classes

$$\mathcal{Q}_t = \{Q_t(\theta_t) | \theta_t \in \mathcal{B}_t\}, \mathcal{V}_t = \{V_t(\theta_t) | \theta_t \in \mathcal{B}_t\}$$

Inherent Bellman Error

- Inherent Bellman error of an MDP with a linear feature representation ϕ is

$$I = \sup_{\theta_{t+1} \in \mathcal{B}_{t+1}} \inf_{\theta_t \in \mathcal{B}_t} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi_t(s,a)^T \theta_t - (\mathcal{T}_t Q_{t+1}(\theta_{t+1}))(s,a)|$$

Inherent Bellman Error

- Inherent Bellman error of an MDP with a linear feature representation ϕ is

$$I = \sup_{\theta_{t+1} \in \mathcal{B}_{t+1}} \inf_{\theta_t \in \mathcal{B}_t} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |\phi_t(s,a)^T \theta_t - (\mathcal{T}_t Q_{t+1}(\theta_{t+1}))(s,a)|$$

- $\forall Q_{t+1} \in \mathcal{Q}_{t+1} \qquad (\mathcal{T}_t Q_{t+1}) \in \mathcal{Q}_t$
- If $\forall Q_{t+1} \in \mathcal{Q}_{t+1} \qquad (\mathcal{T}_t Q_{t+1}) \notin \mathcal{Q}_t \qquad (\Pi \mathcal{T}_t Q_{t+1}) \in \mathcal{Q}_t$
- Projection is done by least square; inherent Bellman error is the projection error.

Inherent Bellman Error

- Inherent Bellman error of an MDP with a linear feature representation ϕ is

$$I = \sup_{\theta_{t+1} \in \mathcal{B}_{t+1}} \inf_{\theta_t \in \mathcal{B}_t} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| \phi_t(s,a)^T \theta_t - (\mathcal{T}_t Q_{t+1}(\theta_{t+1}))(s,a) \right|$$

- $\forall Q_{t+1} \in \mathcal{Q}_{t+1} \qquad (\mathcal{T}_t Q_{t+1}) \in \mathcal{Q}_t$
- If $\forall Q_{t+1} \in \mathcal{Q}_{t+1} \qquad (\mathcal{T}_t Q_{t+1}) \notin \mathcal{Q}_t \qquad (\Pi \mathcal{T}_t Q_{t+1}) \in \mathcal{Q}_t$
- Projection is done by least square; inherent Bellman error is the projection error.
- MDP is low rank indicates $I = 0$; the converse does not hold.

Assumption

- $|Q_t^\pi(s,a)| \leqslant 1, \quad \forall \pi, \forall (s,a,t)$
- $\|\phi_t(s,a)\|_2 \leqslant L_\phi \leqslant 1, \quad \forall (s,a,t)$
- *For any* $Q_t \in \mathcal{Q}_t$ *and any* $(s,a,t) \in \mathcal{S} \times \mathcal{A} \times [H]$ *define the random variable*[5] $X = R_t(s,a) + \max_{a'} Q_{t+1}(s',a')$. *Then the noise* $\eta = X - \mathbb{E}X$ *is* 1*-subgaussian*
- $\forall t \in [H], \forall \theta_t \in \mathcal{B}_t$, *it holds that* $\|\theta_t\| \leqslant \mathcal{R}_t \leqslant \sqrt{d_t}$, *and* $\mathcal{B}_t$ *is compact*

Algorithm

- Regularized least square

$$\sum_{i=1}^{k-1}(\phi_{ti}^T\theta - r_{ti} - V_{t+1}(\theta_{t+1})(s_{t+1,i}))^2 + \lambda\|\theta\|_2^2$$

Algorithm

- Regularized least square

$$\sum_{i=1}^{k-1}(\phi_{ti}^T\theta - r_{ti} - V_{t+1}(\theta_{t+1})(s_{t+1,i}))^2 + \lambda\|\theta\|_2^2$$

- Global optimistics LSVI

$$\begin{aligned}
\max_{\xi_1,\cdots,\xi_H} \max_a & \phi_1(s_{1k},a)^T\bar{\theta}_1 \\
s.t. & \|\xi_t\|_{\Sigma_{tk}} \leq \sqrt{\alpha_{tk}} \\
& \bar{\theta}_t = \Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}[r_{ti} + \max_a \phi_{t+1}(s'_{t+1},a)^T\bar{\theta}] + \xi_t \\
& \bar{\theta}_t \in \mathcal{B}_t, for\ t = H,\cdots,1
\end{aligned}$$

with $\Sigma_{tk} = \sum_{i=1}^{k-1}\phi_{ti}\phi_{ti}^T + \lambda I$

Compare to LSVI-UCB

- Approximated with closed form by adding exploration bonus, similar to linear bandits

$$\bar{\theta} = \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}[r_{ti} + \max_{a}(\phi_{t+1}(s'_{t+1}, a)^T \bar{\theta} + \sqrt{\beta}\|\phi_{t+1}(s'_{t+1}, a)\|_{\Sigma_{tk}^{-1}})]$$

Compare to LSVI-UCB

- Approximated with closed form by adding exploration bonus, similar to linear bandits

$$\bar{\theta} = \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}[r_{ti} + \max_a (\phi_{t+1}(s'_{t+1}, a)^T \bar{\theta} + \sqrt{\beta} \|\phi_{t+1}(s'_{t+1}, a)\|_{\Sigma_{tk}^{-1}})]$$

- LSVI-UCB solve local optimism state by state

Compare to LSVI-UCB

- Approximated with closed form by adding exploration bonus, similar to linear bandits

$$\bar{\theta} = \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}[r_{ti} + \max_a(\phi_{t+1}(s'_{t+1}, a)^T \bar{\theta} + \sqrt{\beta}\|\phi_{t+1}(s'_{t+1}, a)\|_{\Sigma_{tk}^{-1}})]$$

- LSVI-UCB solve local optimism state by state
- Destroys linear structure and increase complexity

Sketch proof

- There exists a parameter $\dot{\theta}$ depending on $\bar{Q}_{t+1}$, such that $\Delta_t(\bar{Q}_{t+1})(s,a) = (\mathcal{T}_t\bar{Q}_{t+1})(s,a) - \phi_t(s,a)^T\dot{\theta}_t(\bar{Q}_{t+1})$ with $\|\Delta_t(\bar{Q}_{t+1})\|_\infty \leq I$
- Sample noise $\eta_{ti}(\bar{V}_{t+1}) = r_{ti} - r_t(s_{ti},a_{ti}) + \bar{V}_{t+1}(s_{t+1},i) - \mathbb{E}_{s'\sim p_t(s_{ti},a_{ti})}\bar{V}_{t+1}(s')$
- $\phi_t(s,a)^T\hat{\theta}_{tk}$ becomes

$$
\begin{aligned}
&\phi_t(s,a)^\top \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}\left(\mathcal{T}_t\overline{Q}_{t+1}(s_{ti},a_{ti}) + \eta_{ti}(\overline{V}_{t+1})\right) \\
&= \phi_t(s,a)^\top \Big[\mathring{\theta}_t(\overline{Q}_{t+1}) + \\
&+ \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}\left(\mathring{\Delta}_{ti} + \eta_{ti}\right)(\overline{Q}_{t+1})\Big] \\
&\overset{eq.\ (6)}{=} \mathcal{T}_t(\overline{Q}_{t+1})(s,a) + \mathring{\Delta}_t(\overline{Q}_{t+1})(s,a) + \\
&+ \phi_t(s,a)^\top \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}\left(\mathring{\Delta}_{ti} + \eta_{ti}\right)(\overline{Q}_{t+1})\,.
\end{aligned}
$$

Sketch proof

- Inherent Bellman error

$$|\phi_t(s,a)^\top \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \mathring{\Delta}_{ti}(\overline{Q}_{t+1})| \leqslant \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \sqrt{k}\mathcal{I}.$$

- Recall Σ_{tk}^{-1}-norm of feature is about $\sqrt{d_t/k}$

Sketch proof

- Inherent Bellman error

$$|\phi_t(s,a)^\top \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \mathring{\Delta}_{ti}(\overline{Q}_{t+1})| \leqslant \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \sqrt{k}\mathcal{I}.$$

- Recall Σ_{tk}^{-1}-norm of feature is about $\sqrt{d_t/k}$
- Noise error

$$\begin{aligned}
&|\phi_t(s,a)^\top \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \eta_{ti}(\overline{V}_{t+1})| \\
&\leqslant \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \| \sum_{i=1}^{k-1} \phi_{ti} \eta_{ti}(\overline{V}_{t+1})\|_{\Sigma_{tk}^{-1}} \\
&\stackrel{def}{\leqslant} \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \sqrt{\beta_{tk}}
\end{aligned}$$

Sketch proof

Lemma 3 (**Transition Noise High Probability Bound**). *If $\lambda = 1$, with probability at least $1 - \delta'$ for all $V_{t+1} \in \mathcal{V}_{t+1}$ it holds that*

$$\left\| \sum_{i=1}^{k-1} \phi_{ti} \left(r_{ti} - r_t(s_{ti}, a_{ti}) + V_{t+1}(s_{t+1,i}) - \mathbb{E}_{s' \sim p_t(s_{ti}, a_{ti})} V_{t+1}(s') \right) \right\|_{\Sigma_{tk}^{-1}} \leqslant \sqrt{\beta_{tk}} \tag{41}$$

where:

$$\sqrt{\beta_{tk}} \stackrel{def}{=} \sqrt{d_t \ln\left(1 + L_\phi^2 k / d_t\right) + 2d_{t+1} \ln(1 + 4\mathcal{R}_t L_\phi \sqrt{k}) + \ln\left(\frac{1}{\delta'}\right)} + 1. \tag{42}$$

- Using ϵ-covering to have a uniform bound for value function class; $\sqrt{\beta_{tk}} = \tilde{O}(\sqrt{d_t})$

Sketch proof

Lemma 3 (Transition Noise High Probability Bound). *If $\lambda = 1$, with probability at least $1 - \delta'$ for all $V_{t+1} \in \mathcal{V}_{t+1}$ it holds that*

$$\left\| \sum_{i=1}^{k-1} \phi_{ti} \left(r_{ti} - r_t(s_{ti}, a_{ti}) + V_{t+1}(s_{t+1,i}) - \mathbb{E}_{s' \sim p_t(s_{ti}, a_{ti})} V_{t+1}(s') \right) \right\|_{\Sigma_{tk}^{-1}} \leqslant \sqrt{\beta_{tk}} \tag{41}$$

where:

$$\sqrt{\beta_{tk}} \stackrel{\mathrm{def}}{=} \sqrt{d_t \ln\left(1 + L_\phi^2 k / d_t\right) + 2 d_{t+1} \ln(1 + 4\mathcal{R}_t L_\phi \sqrt{k}) + \ln\left(\frac{1}{\delta'}\right)} + 1. \tag{42}$$

- Using ϵ-covering to have a uniform bound for value function class; $\sqrt{\beta_{tk}} = \tilde{O}(\sqrt{d_t})$
- The function class is essentially linear, which is simpler compared to LSVI-UCB who uses quadratic exploration bonus, therefore save a $\sqrt{d}$ factor in regret bound

Sketch proof

- Add $\phi_t(s,a)^T \bar{x} i_t$

$$
\begin{aligned}
&| \left(\overline{Q}_t - \mathcal{T}_t \overline{Q}_{t+1}\right)(s,a)| = \\
&\qquad \leqslant \underbrace{\mathcal{I}}_{\text{misspecification}} + \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \times \\
&\qquad \left(\underbrace{\sqrt{k}\mathcal{I}}_{\text{misspecification}} + \underbrace{\sqrt{\alpha_{tk}}}_{\text{exploration}} + \underbrace{\sqrt{\beta_{tk}}}_{\text{noise}} \right).
\end{aligned}
$$

- It remains to define α_{tk}
- Now setting

$$
\overline{\xi}_t = -\Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left(\mathring{\Delta}_{ti} + \eta_{ti} \right) (Q_{t+1}(\theta^\star_{t+1}))
$$

Sketch proof

- So $\bar{Q}_t$ becomes

$$\phi_t(s,a)^\top \overline{\theta}_t$$
$$= \mathcal{T}_t(Q_{t+1}(\theta^\star_{t+1}))(s,a) + \mathring{\Delta}_t(Q_{t+1}(\theta^\star_{t+1}))(s,a).$$

- Thus the approximator satisfies

$$\bar{V}_1(s_{1k}) \geq V_1^*(s_{1k}) - HI$$

- $\bar{\xi}_t$ is bounded by inherent Bellman error and noise error, which satisfies constraints
- Finally we are ready to have regret bound

$$Regret(K) = \sum_{k=1}^{K}(V_1^* - \bar{V}_{1k} + \bar{V}_{1k} - V_1^{\pi_k})(s_{1k}) \leq \tilde{O}(\sum_{t=1}^{H} d_t\sqrt{K} + \sum_{t=1}^{H}\sqrt{d_t}KI)$$

Reference

- Chu W, Li L, Reyzin L, et al. Contextual bandits with linear payoff functions[C]//Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. 2011: 208-214.
- Abbasi-Yadkori Y, Pl D, Szepesvri C. Improved algorithms for linear stochastic bandits[C]//Advances in Neural Information Processing Systems. 2011: 2312-2320.
- Jin C, Yang Z, Wang Z, et al. Provably efficient reinforcement learning with linear function approximation[C]//Conference on Learning Theory. 2020: 2137-2143.
- Zanette A, Lazaric A, Kochenderfer M, et al. Learning Near Optimal Policies with Low Inherent Bellman Error[J]. arXiv preprint arXiv:2003.00153, 2020.
- https://banditalgs.com/2016/10/19/stochastic-linear-bandits/