Reinforcement Learning with Linear Function Approximation

Lizhang Miao

August 20, 2020
Outline

Linear Bandits
  UCB
  Techniques from Linear Bandits

Episodic RL with Linear approximation
  Settings
  Proofs
Linear Bandits
Formulation

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

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

- Contextual bandits: contextual information $c_t$ and action $a_t \in [K]$, reward $r_t = T(c_t, a_t) + \eta_t$ → 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;→ Linear bandits: action vector $a_t \in \mathbb{R}^d$, observed reward
  \[ r_t = \langle a_t, \theta^* \rangle + \eta_t, \eta_t \text{ is zero-mean noise} \]

- Contextual bandits: contextual information $c_t$ and action $a_t \in [K]$, reward
  \[ r_t = T(c_t, a_t) + \eta_t \rightarrow \text{Contextual linear bandits: feature map } \psi : C \times [K] \rightarrow \mathbb{R}^d, \]
  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 \( r_t = \langle a_t, \theta^* \rangle + \eta_t \), \( \eta_t \) 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: \( R_n = \mathbb{E}\left[\sum_{t=1}^{n} r_t^* - r_t\right] \)

- Regret bound: linear bandits \( \tilde{O}(d\sqrt{n}) \); contextual linear bandits \( \tilde{O}(\sqrt{dn}) \)

- Linear approximation and exploration in RL? Transition model?
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 $C_t$ based on collected data $(a_1, r_1, \ldots, a_{t-1}, r_{t-1})$ that contains unknown parameter $\theta^*$ with high probability.
LinUCB

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

- Construct confidence set $C_t$ based on collected data $(a_1, r_1, \cdots, a_{t-1}, r_{t-1})$ that contains unknown parameter $\theta^*$ with high probability
- For a fixed $\hat{\theta}$, the set can be constructed as $C_t = \{ \theta \in \mathbb{R}^d \| \theta - \hat{\theta} \|_V^2 \leq \beta \}$ where $V$ is positive definite and $\| x \|_V = \sqrt{x^T V x}$
- With a unit ball $B_2 = \{ x \in \mathbb{R}^d \| x \|_2 \leq 1 \}$, $C_t = \hat{\theta} + \beta^{1/2} V^{-1/2} B_2$
- $\bar{r}_t(a) = \langle a_t, \hat{\theta} \rangle + \beta^{1/2} \| a \|_V^{-1} \geq r_t^*(a)$ with selection of $\beta$
Construct confidence set $C_t$ based on collected data $(a_1, r_1, \cdots, a_{t-1}, r_{t-1})$ that contains unknown parameter $\theta^*$ with high probability.

For a fixed $\hat{\theta}$, the set can be constructed as $C_t = \{ \theta \in \mathbb{R}^d | \| \theta - \hat{\theta} \|^2_V \leq \beta \}$ where $V$ is positive definite and $\| x \|_V = \sqrt{x^T V x}$.

With a unit ball $B_2 = \{ x \in \mathbb{R}^d | \| x \|_2 \leq 1 \}$, $C_t = \hat{\theta} + \beta^{1/2} V^{-1/2} B_2$.

$\bar{r}_t(a) = \langle a_t, \hat{\theta} \rangle + \beta^{1/2} \| a \|_V^{-1} \geq r_t^*(a)$ with selection of $\beta$.

Exploitation: least square value iteration for $\hat{\theta}$.

Exploration: parameter space confidence ball $\rightarrow$ adding exploration bonus.

Regret $\leq \mathbb{E} \left[ \sum_{t=1}^n \bar{r}_t - r_t \right]$. 

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

Let $\{F_t\}_{t=0}^{\infty}$ be a filtration. Let $\{x_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that $x_t|F_{t-1}$ is $\sigma$-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 \( x_t \mid \mathcal{F}_{t-1} \) is \( \sigma \)-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 \( X_1, X_2, \ldots, X_t \in \mathbb{R}^d \) 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: $M = (S, A, p, r, H)$ with state space $S$, action space $A$, transition kernel $p_t$, reward function $r$ and horizon length $H$.

- $V$-value: $V^\pi_t : S \to \mathbb{R}$ is the expected value of cumulative rewards received under policy $\pi$ when starting from an arbitrary state at the $h$th step

  $$V^\pi_t(x) = \mathbb{E} \left[ \sum_{t' = t}^H r_t(s_t', \pi_t(s_t')) | x_t = x \right], \quad \forall s \in S, t \in [H].$$

  Optimal value $V^*_t(s) = \sup_{\pi} V^\pi_t(s)$ for all $s \in S$ and $t \in [H]$.

- $Q$-value: $Q^\pi_t : S \times A \to \mathbb{R}$ gives the expected value of cumulative rewards when the agent starts from an arbitrary state-action pair at the $t$th step and follows policy $\pi$ afterwards

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

- Bellman equation associated with a policy $\pi$ becomes:
  
  \[
  V_t^{\pi}(s) = Q_t^{\pi}(s, \pi_t(s)), \\
  Q_t^{\pi}(s, a) = (r_t + P_t V_{t+1}^{\pi})(s, a).
  \]
Bellman equation associated with a policy $\pi$ becomes:

$$V^\pi_t(s) = Q^\pi_t(s, \pi_t(s)),$$

$$Q^\pi_t(s, a) = (r_t + \mathbb{P}_t V^\pi_{t+1})(s, a).$$

Bellman optimality equation

$$V^*_t(s) = \max_{a \in A} Q^*_t(s, a),$$

$$Q^*_t(s, a) = (r_t + \mathbb{P}_h V^*_t(s, a)).$$
Bellman equation associated with a policy $\pi$ becomes:

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

Bellman optimality equation

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

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

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

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

- Each parameter $\theta$ identifies an (action) value function

$$Q_t(\theta_t)(s, a) = \phi_t(s, a)^T \theta_t, \quad V_t(\theta_t) = \max_a \phi_t(s, a)^T \theta_t$$
Linear Value Function

- Feature map: \( \phi_t : S \times A \rightarrow \mathbb{R}^{d_t} \)
- \( Q_t(s, a) = \phi_t(s, a)^T \theta_t \)
- 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 \( \theta \) identifies an (action) value function
  \[ Q_t(\theta_t)(s, a) = \phi_t(s, a)^T \theta_t, \quad V_t(\theta_t) = \max_a \phi_t(s, a)^T \theta_t \]
- So consider function classes
  \[ Q_t = \{ Q_t(\theta_t) | \theta_t \in \mathcal{B}_t \}, \quad V_t = \{ V_t(\theta_t) | \theta_t \in \mathcal{B}_t \} \]
Inherent Bellman Error

- Inherent Bellman error of an MDP with a linear feature representation $\phi$ is

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

- MDP is low rank indicates $I = 0$; the converse does not hold.
Inherent Bellman Error

- Inherent Bellman error of an MDP with a linear feature representation $\phi$ is

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

- $\forall Q_{t+1} \in Q_{t+1}, (\mathcal{T}_t Q_{t+1}) \in Q_t$

- If $\forall Q_{t+1} \in Q_{t+1}, (\mathcal{T}_t Q_{t+1}) \notin Q_t, (\prod \mathcal{T}_t Q_{t+1}) \in Q_t$

- 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 $\phi$ is

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

- $\forall Q_{t+1} \in Q_{t+1}$, $(T_t Q_{t+1}) \in Q_t$
- If $\forall Q_{t+1} \in Q_{t+1}$, $(T_t Q_{t+1}) \notin Q_t$, $(\prod T_t Q_{t+1}) \in Q_t$
- 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

- $|Q^\pi_t(s, a)| \leq 1, \quad \forall \pi, \forall (s, a, t)$

- $\|\phi_t(s, a)\|_2 \leq L_\phi \leq 1, \quad \forall (s, a, t)$

- For any $Q_t \in Q_t$ and any $(s, a, t) \in S \times A \times [H]$ define the random variable $X = R_t(s, a) + \max_{a'} Q_{t+1}(s', a')$. Then the noise $\eta = X - \mathbb{E} X$ is 1-subgaussian.

- $\forall t \in [H], \forall \theta_t \in B_t$, it holds that $\|\theta_t\| \leq R_t \leq \sqrt{d_t}$, and $B_t$ is compact.
Algorithm

- Regularized least square

\[
\sum_{i=1}^{k-1} \left( \phi_{ti}^T \theta - r_{ti} - V_{t+1}(\theta_{t+1})(s_{t+1,i}) \right)^2 + \lambda \| \theta \|^2
\]
Algorithm

- Regularized least square

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

- Global optimistics LSVI

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

with \( \Sigma_{tk} = \sum_{i=1}^{k-1} \phi_i \phi_i^T + \lambda I \)
Compare to LSVI-UCB

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

\[
\tilde{\theta} = \sum_{t=1}^{k-1} \phi_{ti} [r_{ti} + \max_a (\phi_{t+1}(s'_{t+1}, a)^T \tilde{\theta} + \sqrt{\beta \| \phi_{t+1}(s'_{t+1}, a) \|_{\Sigma^{-1}_{tk}}})] 
\]
Compare to LSVI-UCB

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

\[
\tilde{\theta} = \sum_{t=1}^{k-1} \sum_{i=1}^{\Sigma_{tk}} \phi_{ti} \left[ r_{ti} + \max_a (\phi_{t+1}(s_{t+1}^{'} , a)^T \tilde{\theta} + \sqrt{\beta} \| \phi_{t+1}(s_{t+1}^{'} , a) \| \Sigma_{t}^{-1}) \right]
\]

- LSVI-UCB solve local optimism state by state
Compare to LSVI-UCB

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

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

- 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
  \[
  \Delta_t(\bar{Q}_{t+1}) (s, a) = (T_t \bar{Q}_{t+1})(s, a) - \phi_t(s, a)^T \hat{\theta}_t(\bar{Q}_{t+1}) \quad \text{with} \quad \| \Delta_t(\bar{Q}_{t+1}) \|_\infty \leq I
  \]

- 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

\[
\phi_t(s, a)^T \sum_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left( T_t \bar{Q}_{t+1} (s_{ti}, a_{ti}) + \eta_{ti} (\bar{V}_{t+1}) \right)
\]

\[
= \phi_t(s, a)^T \left[ \hat{\theta}_t(\bar{Q}_{t+1}) + \sum_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left( \Delta_{ti} + \eta_{ti} \right) (\bar{Q}_{t+1}) \right]
\]

\[
eq (6) \quad T_t(\bar{Q}_{t+1})(s, a) + \Delta_t(\bar{Q}_{t+1})(s, a) + \phi_t(s, a)^T \sum_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left( \Delta_{ti} + \eta_{ti} \right) (\bar{Q}_{t+1}).
\]
Sketch proof

- Inherent Bellman error

\[ |\phi_t(s, a)^\top \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_i \hat{\Delta}_i (\overline{Q}_{t+1})| \leq \|\phi_t(s, a)\|_{\Sigma_{tk}^{-1}} \sqrt{k} \mathcal{I}. \]

- Recall \(\Sigma_{tk}^{-1}\)-norm of feature is about \(\sqrt{d_t/k}\)
Sketch proof

- **Inherent Bellman error**

\[ |\phi_t(s, a)^T \Sigma_{t_k}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \hat{\Delta}_{ti}(\overline{Q}_{t+1})| \leq \|\phi_t(s, a)\|_{\Sigma_{t_k}^{-1}} \sqrt{k} \mathcal{I}. \]

- Recall \( \Sigma_{t_k}^{-1} \)-norm of feature is about \( \sqrt{d_t/k} \)

- **Noise error**

\[ |\phi_t(s, a)^T \Sigma_{t_k}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \eta_{ti}(\overline{V}_{t+1})| \]

\[ \leq \|\phi_t(s, a)\|_{\Sigma_{t_k}^{-1}} \sum_{i=1}^{k-1} \phi_{ti} \eta_{ti}(\overline{V}_{t+1})\|_{\Sigma_{t_k}^{-1}} \]

\[ \xrightarrow{def} \leq \|\phi_t(s, a)\|_{\Sigma_{t_k}^{-1}} \sqrt{\beta_{t_k}} \]
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

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

where:

$$\sqrt{\beta_{tk}} \overset{\text{def}}{=} \sqrt{d_t \ln \left( 1 + \frac{L_\phi^2 k}{d_t} \right) + 2d_{t+1} \ln(1 + 4R_t L_\phi \sqrt{k}) + \ln \left( \frac{1}{\delta'} \right)} + 1.$$  

Using $\epsilon$-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

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

(41)

where:

$$\sqrt{\beta_{tk}} \overset{\text{def}}{=} \sqrt{d_t \ln \left(1 + \frac{L_{\phi}^2 k}{d_t}\right) + 2d_{t+1} \ln(1 + 4R_t L_{\phi} \sqrt{k}) + \ln \left(\frac{1}{\delta'}\right) + 1.}$$

(42)

- Using $\epsilon$-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 $\phi_t(s, a)^T \bar{x}_t$

$$| (\overline{Q}_t - T_t \overline{Q}_{t+1}) (s, a) | =$$

$$\leq \underbrace{\mathcal{I}}_{\text{misspecification}} + \| \phi_t(s, a) \| \Sigma_{t,k}^{-1} \times$$

$$\left( \underbrace{\sqrt{k: \mathcal{I}}}_{\text{misspecification}} + \sqrt{\alpha_{tk}} + \sqrt{\beta_{tk}} \right).$$

- It remains to define $\alpha_{tk}$

- Now setting

$$\bar{\xi}_t = -\Sigma_{t,k}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left( \hat{\Delta}_{ti} + \eta_{ti} \right) (Q_{t+1}(\theta^*_{t+1}))$$
Sketch proof

- So $\bar{Q}_t$ becomes

$$\phi_t(s, a)^\top \bar{\theta}_t$$

$$= \mathcal{T}_t(Q_{t+1}(\theta_{t+1}^*)) (s, a) + \hat{\Delta}_t(Q_{t+1}(\theta_{t+1}^*)) (s, a).$$

- Thus the approximator satisfies

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

- $\bar{\xi}_t$ is bounded by inherent Bellman error and noise error, which satisfies constraints.

- Finally we are ready to have regret bound

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

- https://banditalgs.com/2016/10/19/stochastic-linear-bandits/