Non-linear Contextual Bandit

August 7, 2022

1 Introduction

We focus on contextual bandit in this note.

2 Contextual Bandit

2.1 Problem Formulation

Definition 1 (Contextual bandit)**.** *A contextual bandit problem is a tuple* $(\mathcal{X}, \mathcal{A}, r)$*. Given context* $x \in \mathcal{X}$*, we take an action* $a \in \mathcal{A}$*, and observe a reward* $r \in \mathbb{R}$ *that can depend on* (x, a)*. The bandit game repeats as follows: at each time step* t*,*

- *We observe a context* $x_t \in \mathcal{X}$*;*
- *The player chooses one arm* $a_t \in \mathcal{A}$*;*
- *The reward* r_t *is revealed.*

The goal is to maximize the expected cumulative reward:

$$\sum_{t=1}^{T} \mathbb{E}_{a_t \sim \pi_t}[r_t(a_t)],$$

where $\pi_t : \mathcal{X} \to \Delta_{\mathcal{A}}$ is a policy. The reward generalization can be either adversarial or stochastic.

- Stochastic: there exists an unknown *value function*:

$$f_*(x, a) = \mathbb{E}[r|x, a], \qquad f_*(x) = \max_{a \in \mathcal{A}} f(x, a).$$

- Adversarial: we are given an arbitrary reward sequence $\{[r_t(a) : a \in \mathcal{A}] : t \in [T]\}$ before hand (also referred to as an oblivious adversary).

2.2 Preliminary

Lemma 1. *Given any function $U(w)$, we have*

$$\min_p \left[\mathbb{E}_{w\sim p}U(w) + \mathrm{KL}\left(p\|p_0\right)\right] = -\ln \mathbb{E}_{w\sim p_0} \exp(-U(w)),$$

where the minimum is achieved by the Gibbs distribution $q(w) \propto p_0(w)\exp(-U(w))$.

3 EXP4 for Adversarial Contextual Bandit

We consider the finite-arm setting where $\mathcal{A} = \{1, \cdots, K\}$. Assume that we are given a random policy class indexed by w:

$$\mathcal{G} = \{p(a|w, x) : w \in \Omega\},$$

where each $p(\cdot|w, x)$ is a conditional distribution over $\{1, \cdots, K\}$. The EXP4 algorithm maintains a distribution over the policy class, which induces a distribution over $\mathcal{A}$ by:

$$p_t(a) = (1-\gamma)\mathbb{E}_{w\sim p_{t-1}(w)}p(a|w, x_t) + \frac{\gamma}{K}, \tag{3.1}$$

where $\gamma > 0$ is a parameter controlling exploration. It remains to construct the posterior distribution $p_t(w)$ over $\mathcal{G}$.

We start with a prior $p_0(w)$. The posterior is constructed by standard online aggregation trick. For each time step t, we use the following reward estimators:

$$\hat{r}_t(w, x_t) = \frac{p(a_t|w, x_t)}{p_t(a_t)}(r_t(a_t) - b). \tag{3.2}$$

Then, the posterior is given by

$$p_t(w) = \frac{p_0(w)\exp\left(\eta \sum_{i=1}^t \hat{r}_t(w, x_t)\right)}{\mathbb{E}_{w\sim p_0(w)}p_0(w)\exp\left(\eta \sum_{i=1}^t \hat{r}_t(w, x_t)\right)}. \tag{3.3}$$

The estimator is unbiased for

$$r_t(w, x_t) - b = \sum_{a=1}^K p(a|w, x_t)(r_t(a) - b),$$

which relies on the full reward vector at time step t. The parameter b also controls exploration by put more penalty on the observed arm and thereby favors arms that are not observed.

We have the following theoretical guarantee.

Theorem 1. *For any $K, T \geq 0$ and any $\gamma \in (0, 1], \eta > 0$ and $b \geq 0$. Consider any family of policies*

$\mathcal{G} = \{p(a|w,x) : w \in \Omega\}$ with prior $p_0(w)$. Then, we have

$$\begin{aligned}\mathbb{E}\sum_{t=1}^{T} r_t\left(a_t\right) \geq (1-\gamma)\max_q &\left[\mathbb{E}_{w\sim q}\sum_{t=1}^{T}\mathbb{E}_{a\sim p(\cdot|w,x_t)} r_t(a) - \frac{1}{\eta}\mathrm{KL}\left(q||p_0\right),\right] \\ &- c(\eta,b)\eta\sum_{t=1}^{T}\sum_{a=1}^{K}\left|r_t(a)-b\right|,\end{aligned} \tag{3.4}$$

where the expectation is w.r.t. the randomness of the algorithm,

$$c(\eta,b) = \psi\left(z_0\right)\max(b,1-b), \quad z_0 = \max(0,\eta(1-b)K/\gamma),$$

and $\psi(z) = (e^z - 1 - z)/z^2$.

We have the following corollary.

Corollary 2. *Let $\eta = \gamma/K$ and $b = 0$. Assumes that the uniform random policy belongs to $\mathcal{G}$ and $|\mathcal{G}| = N < \infty$. Let $p_0(w)$ be the uniform prior over Ω, then*

$$G_* - \mathbb{E}\sum_{t=1}^{T} r_t\left(a_t\right) \leq (e-1)\gamma G_* + \frac{K\ln N}{\gamma}, \tag{3.5}$$

where the expectation is with respect to the randomness of algorithm, and

$$G_* = \operatorname*{argmax}_w \sum_{t=1}^{T}\mathbb{E}_{a\sim p(\cdot|w,x_t)}[r_t(a)].$$

Remark 1. *For Hedge with full feedback, we do not have to explore in order to obtain rewards for different arms. This removes the K-dependency in the resulting bound.*

Remark 2. *EXP4 tries to find a best policy within a policy class, which can be regarded as a policy-based algorithm.*

3.1 Analysis

Theorem 1. We will first estimate the first-order moment and second-order moment of the reward estimator, respectively. Then, we will use the fact that $\psi(z) = (e^z - 1 - z)/z^2$ is increasing so we can bound e^z (which is the likelihood) by $1 + z + \psi(z_0)z^2$ (which have been estimated). We then use standard online aggregation analysis trick to finish the proof.

By Eqn. (3.1), we know that

$$\mathbb{E}_{w\sim p_{t-1}(w)} p(a_t|w,x_t) \leq p_t(a_t)/(1-\gamma). \tag{3.6}$$

This implies that

$$\begin{aligned}\mathbb{E}_{w\sim p_{t-1}(w)}\hat{r}_t(w,x_t) &= \mathbb{E}_{w\sim p_{t-1}(w)}p(a_t \mid w,x_t)\left[r_t(a_t)-b\right]/p_t(a_t)\\ &\le \frac{1}{1-\gamma}r_t(a_t)-q_t(a_t)\,b,\end{aligned} \tag{3.7}$$

where $q_t(a) = \mathbb{E}_{w\sim p_{t-1}(w)}p(a|w,x_t)/p_t(a)$. We also have

$$\begin{aligned}&\mathbb{E}_{w\sim p_{t-1}(w)}\hat{r}_t(w,x_t)^2\\ =&\mathbb{E}_{w\sim p_{t-1}(w)}p(a_t \mid w,x_t)^2\left((r_t(a_t)-b)/p_t(a_t)\right)^2\\ \le&\max(b,1-b)\mathbb{E}_{w\sim p_{t-1}(w)}p(a_t\mid w,x_t)\left(|r_t(a_t)-b|/p_t(a_t)^2\right)\\ \le&\frac{\max(b,1-b)}{1-\gamma}\left(|r_t(a_t)-b|/p_t(a_t)\right),\end{aligned} \tag{3.8}$$

where the first inequality uses $p(a_t|w,x_t)\le 1$ and $|r_t(a_t)-b|\le\max(b,1-b)$; the second inequality uses Eqn. (3.6). We still need a range estimation for $\eta\hat{r}_t(w,x_t)$:

$$\eta\hat{r}_t(w,x_t) = \eta\frac{p(a_t|w,x_t)}{p_t(a_t)}(r_t(a_t)-b)\le\max(0,\eta(1-b)K/\gamma),$$

as $p(a_t|w,x_t)/p_t(a_t)\ge\gamma/K$. We now define

$$W_t = \mathbb{E}_{w\sim p_0(w)}\exp(\eta\sum_{k=1}^{t}\hat{r}_k(w,x_k)).$$

It follows that

$$\begin{aligned}&\ln\frac{W_t}{W_{t-1}} = \ln\mathbb{E}_{w\sim p_0(w)}\frac{\exp(\eta\sum_{k=1}^{t}\hat{r}_k(w,x_k))}{W_{t-1}}\\ =&\ln\underbrace{\mathbb{E}_{w\sim p_0(w)}\frac{\exp(\eta\sum_{k=1}^{t-1}\hat{r}_k(w,x_k))}{\mathbb{E}_{w\sim p_0(w)}\exp(\eta\sum_{k=1}^{t-1}\hat{r}_k(w,x_k))}}_{\text{density of } p_{t-1}(w)}\exp(\eta\hat{r}_t(w,x_t))\\ =&\ln\mathbb{E}_{w\sim p_{t-1}(w)}\exp\left(\eta\hat{r}_t(w,x_t)\right)\\ \le&\ln\mathbb{E}_{w\sim p_{t-1}(w)}\left[1+(\eta\hat{r}_t(w,x_t))+\psi(z_0)(\eta\hat{r}_t(w,x_t))^2\right]\\ \le&\mathbb{E}_{w\sim p_{t-1}(w)}(\eta\hat{r}_t(w,x_t))+\psi(z_0)\mathbb{E}_{w\sim p_{t-1}(w)}(\eta\hat{r}_t(w,x_t))^2\\ \le&\frac{\eta}{1-\gamma}r_t(a_t)-\eta q_t(a_t)\,b+\frac{c(\eta,b)\eta^2}{(1-\gamma)}\frac{|r_t(a_t)-b|}{p_t(a_t)},\end{aligned}$$

where in the first inequality we uses $z=\eta\hat{r}_t(w,x_t)\le\max(0,\eta(1-b)K/\gamma)$; the second inequality uses $\ln(1+z)\le z$; and the last inequality uses Eqn. (3.7) and (3.8), with $c(\eta,b)=\psi(z_0)\max(b,1-b)$. Note $W_0=1$. We now sum over $t\in[T]$ to obtain that

$$\ln W_T = \ln\frac{W_T}{W_0}\le\frac{\eta}{1-\gamma}\sum_{t=1}^{T}r_t(a_t)-\eta b\sum_{t=1}^{T}q_t(a_t)+\frac{c(\eta,b)\eta^2}{(1-\gamma)}\sum_{t=1}^{T}\frac{|r_t(a_t)-b|}{p_t(a_t)}.$$

Taking expectation with respect to the randomness of the algorithm, we have

$$\begin{aligned}\mathbb{E}\ln W_T &= \mathbb{E}\ln \mathbb{E}_{w\sim p_0(w)} \exp\left(\eta\sum_{t=1}^{T}\hat{r}_t\left(w, x_t\right)\right)\\ &\le \frac{\eta}{1-\gamma}\mathbb{E}\sum_{t=1}^{T} r_t\left(a_t\right) - \eta T b + \frac{c(\eta, b)\eta^2}{(1-\gamma)}\mathbb{E}\sum_{t=1}^{T}\sum_{a=1}^{K}\left|r_t(a) - b\right|.\end{aligned}$$

We now invoke Lemma 1 to derive an lower bound of $\mathbb{E}\ln W_T$.

$$\begin{aligned}&\mathbb{E}\ln \mathbb{E}_{w\sim p_0(w)} \exp\left(\eta\sum_{t=1}^{T}\hat{r}_t\left(w, x_t\right)\right)\\ =&\mathbb{E}\max_q\left[\mathbb{E}_{w\sim q}\eta\sum_{t=1}^{T}\hat{r}_t\left(w, x_t\right) - \mathrm{KL}\left(q\|p_0\right)\right]\\ \ge&\max_q\mathbb{E}\left[\mathbb{E}_{w\sim q}\eta\sum_{t=1}^{T}\hat{r}_t(w, x_t) - \mathrm{KL}\left(q\|p_0\right)\right]\\ =&\max_q\mathbb{E}\left[\mathbb{E}_{w\sim q}\eta\sum_{t=1}^{T}\left[r_t\left(w, x_t\right) - b\right] - \mathrm{KL}\left(q\|p_0\right)\right],\end{aligned}$$

where we use Lemma 1 in the first equality and $r_t(w, x_t) = \mathbb{E}_{a\sim p(\cdot|w,x_t)} r(a)$. The desired theorem then follows from rearranging terms. □

We now prove the corollary.

Proof of Corollary 2. With the specified choice of parameters, we now have $\eta\hat{r}_t(w, x_t) \le 1$ and $c(\eta, b) = e - 2$. Note that the uniform random policy belongs to Ω implies that

$$\frac{1}{K}\sum_{t=1}^{T}\sum_{a=1}^{K} r_t(a) \le G_*.$$

With $q(w) := I(w = w_*)$, where w_* achieves the maximum of G_*, from Theorem 1, we have

$$\mathbb{E}\sum_{t=1}^{T} r_t\left(a_t\right) \ge (1-\gamma)\left[G_* - \frac{K}{\gamma}\ln N\right] - (e-2)\gamma G_*.$$

□

4 LinUCB for Stochastic Contextual Bandit

We consider the stochastic contextual bandit with linear function approximation.

Definition 2 (Stochastic Linear Contextual Bandit)**.** *The reward at each time step is given by*

$$r_t(a) = r_t(x_t, a) = w_*^\top\psi(x_t, a) + \epsilon_t(x_t, a),$$

where $\psi(\cdot,\cdot) : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a known feature map and $\epsilon_t(x,a)$ is a zero-mean random variable.

In this setting, the number of arms can be either infinite or finite. We remark that the condition that $|\mathcal{A}|$ is finite can be used to achieve sharper regret bound, known as the finite-action case (Chu et al., 2011). Here we focus the UCB-type algorithm presented in Abbasi-Yadkori et al. (2011).

4.1 Optimism in Face of Uncertainty

The core design of a UCB-type algorithm is to determine the confidence set such that:

- Optimism is achieved: the optimal target lies in the confidence set;
- The confidence set is as sharp as possible.

In this case, we start with $A_0 = \lambda I, w_0 = b_0 = 0$. At each iteration step t, after observe context x_t,

- We choose $a_t \in \mathrm{argmax}_a[w_{t-1}^\top \psi(x_t, a) + \beta_{t-1} \|\psi(x_t, a)\|_{A_{t-1}^{-1}}]$;
- $b_t = b_{t-1} + r_t(x_t, a_t)\psi(x_t, a_t)$;
- $A_t = A_{t-1} + \psi(x_t, a_t)\psi(x_t, a_t)^\top$;
- $w_t = A_t^{-1} b_t$.

The proof employs the following famous self-normalized process concentration bound, which holds for all arms (possible infinitely many).

Lemma 2 (Self-normalized process concentration inequality)**.** *Let $\{(X_t, \epsilon_t)\}$ be a sequence in $\mathbb{R}^d \times \mathbb{R}$ w.r.t. a filtration $\{\mathcal{F}_t\}$ so that*

$$\mathbb{E}[\epsilon_t | X_t, \mathcal{F}_{t-1}] = 0, \qquad \mathrm{var}[\epsilon_t | X_t, \mathcal{F}_{t-1}] \leq \sigma^2$$

. Assume also that $|\epsilon_t| \leq M$. Let Λ_0 be a positive definite matrix, and

$$\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} X_s X_s^\top.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$:

$$\left\| \sum_{t=1}^{t} \epsilon_s X_s \right\|_{\Lambda_t^{-1}}^2 \leq 1.3\sigma^2 \ln |\Lambda_0^{-1} \Lambda_t| + 4M^2 \ln(2/\delta).$$

Lemma 3 (Concentration and Optimism)**.** *Assume that $r_t(x_t, a_t) \in [0, 1]$ and*

$$\mathrm{var}_{r_t | x_t, a_t}(r_t(x_t, a_t)) \leq \sigma^2.$$

Assume further that $\|w_*\|_2 \leq B$ for some constant B. Then, with probability at least $1 - \delta$, for all $t \geq 0$, and $u \in \mathbb{R}^d$, we have :

$$|u^\top (w_t - w_*)| \leq \beta_t \sqrt{u^\top A_t^{-1} u},$$

where $\beta_t = \sqrt{\lambda}B + 1.3\sigma\sqrt{\ln|A_t/\lambda|} + 4\sqrt{\ln(2/\delta)}$.

Proof. We apply the self-normalized process concentration inequality to obtain that

$$\left\|\sum_{s=1}^{t} \epsilon_s(x_s, a_s)\psi(x_s, a_s)\right\|_{A_t^{-1}} \leq 1.3\sigma\sqrt{\ln|A_0^{-1}A_t|} + 4\sqrt{\ln(2/\delta)}, \qquad \forall t.$$

Then, we can add and subtract $u^\top A_t^{-1}\sum_{s=1}^t w_*^\top \psi(x_s, a_s)\psi(x_s, a_s)$ to obtain that

$$\begin{aligned}
u^\top(w_t - w_*) &= u^\top A_t^{-1}\sum_{s=1}^{t} r_s(x_s, a_s)\,\psi(x_s, a_s) - u^\top w_* \\
&= u^\top A_t^{-1}\sum_{s=1}^{t} \epsilon_s(x_s, a_s)\,\psi(x_s, a_s) - \lambda u^\top A_t^{-1} w_* \\
&\leq \|u\|_{A_t^{-1}}\left\|\sum_{s=1}^{t} \epsilon_s(x_s, a_s)\,\psi(x_s, a_s)\right\|_{A_t^{-1}} + \lambda\|u\|_{A_t^{-1}}\|w_*\|_{A_t^{-1}} \\
&\leq \|u\|_{A_t^{-1}}\left(1.3\sigma\sqrt{\ln\left(\left|A_0^{-1}A_t\right|\right)} + 4\sqrt{\ln(2/\delta)}\right) + \sqrt{\lambda}\|u\|_{A_t^{-1}}\|w_*\|_2\,.
\end{aligned}$$

□

We have the following theoretical result.

Theorem 3. *Assume that* $r_t(x_t, a_t) \in [0, 1]$ *and*

$$\mathrm{var}_{r_t|x_t,a_t} r_t(x_t, a_t) \leq \sigma^2, \qquad \|w_*\| \leq B.$$

Let $\mu_t(x, a) = \mathbb{E}_{\epsilon_t(x,a)} r_t(x, a) = w_*^\top \psi(x, a)$ *and* $a_*(x) \in \operatorname{argmax}_a \mu_t(x, a)$. *Then, with probability at least* $1 - \delta$, *for any* $t \geq 0$, *and* $u \in \mathbb{R}^d$, *LinUCB satisfies*

$$\mathbb{E}\sum_{t=1}^{T}\left[\mu_t(x_t, a_*(x_t)) - \mu_t(x_t, a_t)\right] \leq 2.5\sqrt{\ln|A_T/\lambda|\sum_{t=1}^{T}\beta_t^2},$$

where $\beta_t = \sqrt{\lambda}B + 1.3\sigma\sqrt{\ln|A_t/\lambda|} + 4\sqrt{\ln(2/\delta)}$.

Proof. For $t \geq 1$, with probability at least $1 - \delta$, we have

$$\begin{aligned}
&w_*^\top \psi(x_t, a_*(x_t)) \\
\leq &w_{t-1}^\top \psi(x_t, a_*(x_t)) + \beta_{t-1}\sqrt{\psi(x_t, a_*(x_t))^\top A_{t-1}^{-1}\psi(x_t, a_*(x_t))} \\
\leq &w_{t-1}^\top \psi(x_t, a_t) + \beta_{t-1}\sqrt{\psi(x_t, a_t)^\top A_{t-1}^{-1}\psi(x_t, a_t)} \qquad \text{By optimism.} \\
\leq &w_*^\top \psi(x_t, a_t) + 2\beta_{t-1}\sqrt{\psi(x_t, a_t)^\top A_{t-1}^{-1}\psi(x_t, a_t)}.
\end{aligned}$$

The result then follows a careful analysis of the self-normalized process. Since $w_*^\top\psi(x_t, a) \in [0, 1]$, we can refined the regret bound by

$$w_*^\top \psi\left(x_t, a_*\left(x_t\right)\right) - w_*^\top \psi\left(x_t, a_t\right) \leq 2\beta_{t-1}\sqrt{\min\left(\psi\left(x_t, a_t\right)^\top A_{t-1}^{-1}\psi\left(x_t, a_t\right), 0.25\right)}.$$

By summing over $t = 1$ to $t = T$, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\left[\mu_t\left(x_t, a_*\left(x_t\right)\right) - \mu_t\left(x_t, a_t\right)\right] \\
\leq &2\sum_{t=1}^{T}\beta_{t-1}\sqrt{\min\left(\psi\left(x_t, a_t\right)^\top A_{t-1}^{-1}\psi\left(x_t, a_t\right), 0.25\right)} \\
\leq &2\sqrt{\sum_{t=1}^{T}\beta_t^2}\sqrt{\sum_{t=1}^{T}\min\left(\psi\left(x_t, a_t\right)^\top A_{t-1}^{-1}\psi\left(x_t, a_t\right), 0.25\right)} \\
\leq &2\sqrt{\sum_{t=1}^{T}\beta_t^2}\sqrt{1.25\sum_{t=1}^{T}\frac{\psi\left(x_t, a_t\right)^\top A_{t-1}^{-1}\psi\left(x_t, a_t\right)}{1 + \psi\left(x_t, a_t\right)^\top A_{t-1}^{-1}\psi\left(x_t, a_t\right)}}.
\end{aligned}$$

The proof is completed with the following lemma. □

Lemma 4. *Let Σ_0 be a $d \times d$ symmetric positive definite matrix and $\{\psi(X_t)\}$ be a sequence of vectors in $\mathbb{R}^d$. Let $\Sigma_t = \Sigma_0 + \sum_{s=1}^{t}\psi(X_s)\psi(X_s)^\top$, then*

$$\sum_{s=1}^{t}\frac{\psi\left(X_s\right)^\top \Sigma_{s-1}^{-1}\psi\left(X_s\right)}{1 + \psi\left(X_s\right)^\top \Sigma_{s-1}^{-1}\psi\left(X_s\right)} \leq \ln\left|\Sigma_0^{-1}\Sigma_t\right|.$$

5 Weakly Nonlinear UCB with Eluder Coefficient

Definition 3. *Stochastic nonlinear contextual bandit is a contextual bandit problem, where the reward at each time step t is given by*

$$r_t(a) = r_t(x_t, a) = f_*(x_t, a) + \epsilon_t(x_t, a),$$

where $\epsilon_t(x, a)$ is a zero-mean random variable. We assume that $f_*(x, a) \in \mathcal{F}$ for a known function class $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. We also define

$$f(x) = \max_{a \in \mathcal{A}} f(x, a).$$

5.1 Non-linear UCB

In this section, we still consider a UCB-type algorithm where we maintain a confidence set, also referred to as a version space, $\mathcal{F}_t$, such that $f_* \in \mathcal{F}_t$ with high probability. Then, given x_t, the algorithm chooses f_t by

$$f_t = \operatorname*{argmax}_{f \in \mathcal{F}_{t-1}} f(x_t), \qquad a_t \in \operatorname*{argmax}_{a} f_t(x_t, a).$$

As a special case, we consider the linear setting where $\mathcal{F} = \{f_w(x,a) = w^\top \psi(x,a) : w \in \mathbb{R}^d\}$. Let

$$\mathcal{F}_t = \{f_w(\cdot) : \sum_{s=1}^{t} (w^\top \psi(x_s, a_s) - r_s(x_s, a_s))^2 + \lambda \left\|w\right\|_2^2 \leq \inf_{w_0} \sum_{s=1}^{t} (w_0^\top \psi(x_s, a_s) - r_s(x_s, a_s))^2 + \lambda \left\|w_0\right\|_2^2 + \beta_t^2\}.$$

Then, we have

$$\mathcal{F}_{t-1} = \{f_w(x,a) : \left\|w - w_{t-1}\right\|_{A_{t-1}} \leq \beta_{t-1},$$

and

$$\max_{f \in \mathcal{F}_{t-1}} f\left(x_t, a\right) = w_{t-1}^\top \psi\left(x_t, a\right) + \beta_{t-1} \left\|\psi\left(x_t, a\right)\right\|_{A_{t-1}^{-1}},$$

where $w_{t-1} = \operatorname{argmax}_{w'} \sum_{s=1}^{t-1} ((w')^\top \psi(x_s, a_s) - r_s(x_s, a_s))^2 + \lambda \left\|w'\right\|_2^2$.

Intuitively, the version space $\mathcal{F}_t$ contains functions that fit well on the historical dataset $\mathcal{S}_t = \{(x_s, a_s, r_s)\}_{s=1}^t$ and we expect that they perform well on the unseen sample at iteration $t+1$, which corresponds to the out-of-sample error. To analyze the algorithm, we need some structural information to ensure certain good generalization property.

Definition 4 (Eluder Coefficient)**.** *Given a function class $\mathcal{F}$, its Eluder coefficient* $\mathrm{EC}(\epsilon, \mathcal{F}, T)$ *is defined to be the smallest number d so that for any sequence $\{(x_t, a_t)\}_{t=1}^T$ and $\{f_t\}_{t=1}^T \in \mathcal{F}$:*

$$\sum_{t=2}^{T} [f_t(x_t, a_t) - f_*(x_t, a_t)] \leq \sqrt{d \sum_{t=2}^{T} \left(\epsilon + \sum_{s=1}^{t-1} |f_t(x_s, a_s) - f_*(x_s, a_s)|^2\right)}.$$

Theorem 4. *Assume that ϵ_t is conditioned zero-mean sub-Gaussian noise: for all $\lambda \in \mathbb{R}$,*

$$\ln \mathbb{E}[e^{\lambda \epsilon_t} | x_t, \mathcal{F}_{t-1}] \leq \frac{\lambda^2}{2} \sigma^2.$$

If we define

$$\hat{f}_t = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{s=1}^{t} \left(f(x_s, a_s) - r_s\right)^2,$$

and

$$\mathcal{F}_t = \left\{f \in \mathcal{F} : \sum_{s=1}^{t} \left(f(x_s, a_s) - \hat{f}(x_s, a_s)\right)^2 \leq \beta_t^2\right\},$$

where

$$\beta_t^2 = \inf_{\epsilon > 0} [9\epsilon t(\sigma + 2\epsilon) + 12\sigma^2 \ln\left(2N(\epsilon, \mathcal{F}, \left\|\cdot\right\|_\infty\right) / \delta].$$

Then with probability at least $1-\delta$*:*

$$\sum_{t=2}^{T}\left[f_*\left(x_t\right)-f_*\left(x_t,a_t\right)\right]\le\sqrt{\mathrm{EC}(\epsilon,\mathcal{F},T)\left(\epsilon T+4\sum_{t=2}^{T}\beta_{t-1}^2\right)}.$$

Proof. We have

$$\begin{aligned}&f_*\left(x_t\right)-f_*\left(x_t,a_t\right)\\=&f_*\left(x_t\right)-f_t\left(x_t\right)+f_t\left(x_t,a_t\right)-f_*\left(x_t,a_t\right)\\\le&f_t\left(x_t,a_t\right)-f_*\left(x_t,a_t\right),\end{aligned}$$

where we use $f_t(x_t)=f_t(x_t,a_t)$ as a_t is greedy with respect to f_t and the inequality is due to optimism of f_t. It follows that

$$\begin{aligned}&\sum_{t=2}^{T}\left[f_*\left(x_t\right)-f_*\left(x_t,a_t\right)\right]\\\le&\sum_{t=2}^{T}\left[f_t\left(x_t,a_t\right)-f_*\left(x_t,a_t\right)\right]\\\le&\sqrt{\mathrm{EC}(\epsilon,\mathcal{F},T)\sum_{t=2}^{T}\left(\epsilon+\sum_{s=1}^{t-1}\left|f_t\left(x_s,a_s\right)-f_*\left(x_s,a_s\right)\right|^2\right)}\\\le&\sqrt{\mathrm{EC}(\epsilon,\mathcal{F},T)\left(\epsilon T+4\sum_{t=2}^{T}\beta_{t-1}^2\right)},\end{aligned}$$

where the last inequality follows from

$$\begin{aligned}&\sum_{s=1}^{t-1}\left|f_t\left(x_s,a_s\right)-f_*\left(x_s,a_s\right)\right|^2\\\le&4\sum_{s=1}^{t-1}\left[\left|f_t\left(x_s,a_s\right)-\hat{f}_{t-1}\left(x_s,a_s\right)\right|^2+\left|f_*\left(x_s,a_s\right)-\hat{f}_{t-1}\left(x_s,a_s\right)\right|^2\right]\le4\beta_{t-1}^2.\end{aligned}$$

as $f_t,f_*\in\mathcal{F}_t$. It remains to determine the value of β_t^2 and to show that the sequence ensures optimism. This follows from standard ridge regression analysis and we omit it here. □

5.2 Estimating Eluder Coefficient

Lemma 5. *Consider a RKHS* $\mathcal{H}$ *with feature representation* $f(x,a)=w\cdot\psi(x,a)$ *for all* $f\in\mathcal{H}$ *and* $\|f\|_{\mathcal{H}}=\|w\|_2$*. Assume that* $\|f-f_*\|_{\mathcal{H}}\le B$ *for all* $f\in\mathcal{F}\subset\mathcal{H}$ *and* $\psi(x,a)=[\psi_j(x,a)]_{j=1}^{\infty}$*. Given any* $\epsilon'>0$*, we also denote*

$$d(\epsilon')=\min\left\{|S|:\sup_{x,a}\sum_{j\notin S}(\psi_j(x,a))^2\le\epsilon'\right\},$$

and $\|\psi(x,a)\|_2 \leq B'$*. If* $|f - f_*| \leq M$ *for all* $f \in \mathcal{F}$*, then we have*

$$\mathrm{EC}(\epsilon, \mathcal{F}, T) \leq (1+\epsilon^{-1}) d(\epsilon B^{-2}) \ln\left(1 + \frac{T(BB')^2}{d(\epsilon B^{-2})\epsilon}\right). \tag{5.1}$$

In particular, if $\mathcal{H}$ *is* d*-dimensional for a finite* d*, then we have*

$$\mathrm{EC}(M^2, \mathcal{F}, T) \leq 2d \ln\left(1 + 4T(BB'/M)^2/d\right).$$

References

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Chu, W., Li, L., Reyzin, L., & Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (pp. 208–214).: JMLR Workshop and Conference Proceedings.