Contextual Bandits

Mathematical Analysis of Machine Learning Algorithms
(Chapter 17)

Contextual Bandits

Definition 1 (Contextual Bandit Problem)

In contextual bandit, we consider a context space $\mathcal{X}$ and an action space $\mathcal{A}$. 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 contextual bandit problem is a repeated game: at each time step t:

- The player observes a sample $x_t \in \mathcal{X}$
- The player chooses precisely one action (or arm) $a_t \in \mathcal{A}$
- The reward r_t is revealed.

Policy for Contextual Bandit

Definition 2

A policy π for contextual bandit is a map $\mathcal{X} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes probability measures over $\mathcal{A}$ with an appropriately defined σ-algebra. One may also write it as a conditional distribution $\pi(a|x)$, and the policy draws $a \sim \pi(\cdot|x)$ when it observes context x. A contextual bandit algorithm $\hat{q}$ maps historic observations

$$\mathcal{S}_{t-1} = \{(x_1, a_1, r_1), \ldots, (x_{t-1}, a_{t-1}, r_{t-1})\}$$

to a policy $\pi_t = \hat{q}(\cdot|\mathcal{S}_{t-1})$ at each time step t, and pulls an arm $a_t \sim \pi_t(\cdot|x_t)$ based on the observation x_t. In this chapter, we will also write the history dependent policy as $a_t \sim \hat{q}(a_t|x_t, \mathcal{S}_{t-1})$.

Adversarial Setting

Similar to the case of multi-armed bandit problem, we may also consider the adversarial setting with an oblivious adversary as follows. At each time step t, we have the information of all rewards $[r_t(a) : a \in \mathcal{A}]$, but only reveals the value of $r_t(a_t)$ for the chosen arm a_t. The goal is to maximize the expected cumulative reward

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

If we are given a policy class Π, then regret of a contextual bandit algorithm with respect to Π can be written as follows.

Regret

$$\mathrm{REG}_T = \sup_{\pi \in \Pi} \sum_{t=1}^{T} \mathbb{E}_{a_t \sim \pi}[r_t(a_t)] - \sum_{t=1}^{T} \mathbb{E}_{a_t \sim \pi_t}[r_t(a_t)]. \tag{1}$$

Stochastic Setting

If we consider the stochastic contextual bandit setting with unknown *value functions*

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

that do not change over time, then the goal becomes to maximize the expected reward

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

Regret

The regret of a bandit algorithm that produces policy sequence $\{\pi_t\}$ is:

$$\mathrm{REG}_T = \sum_{t=1}^{T} \mathbb{E}_{a_t \sim \pi_t}[f_*(x_t) - f_*(x_t, a_t)]. \tag{2}$$

EXP4: Policy based Algorithm

EXP4 is a generalization of the EXP3. It can be regarded as a policy based method for adversarial contextual bandits.

Assume that we have an expert class indexed by w:

$$\mathcal{G} = \{[\hat{q}_t(\cdot|w, x_t)]_{t=1,2,\ldots} : w \in \Omega\}.$$

Given any context $x_t \in \mathcal{X}$, an expert w returns a probability distribution $\hat{q}_t(\cdot|w, x_t)$ on $a_t \in \{1, \ldots, K\}$.

Let $p_0(w)$ be a prior on Ω, then the EXP4 algorithm, has a regret bound that is logarithmic in $|\mathcal{G}|$ for finite $\mathcal{G}$, if the regret is to compete with the best expert in $\mathcal{G}$.

Example

Example 3

Any stationary policy can be regarded as an expert. As an example, we may consider experts of logistic policies (parametrized by w) defined as

$$\hat{q}_t(a|w,x) = \hat{q}(a|w,x) = \frac{\exp(w^\top \psi(x,a))}{\sum_{\ell=1}^{K} \exp(w^\top \psi(x,\ell))},$$

with Gaussian prior $p_0(w)$:

$$w \sim N(0, \sigma^2).$$

EXP4 Algorithm

Algorithm 1: EXP4

Input: K, T, $\mathcal{G}$, $p_0(\cdot)$, $\gamma \in (0, 1]$, $\eta > 0$, $b \geq 0$

Let $u_0(w) = 1$

for $t = 1, 2, \ldots, T$ **do**

  Observe x_t

  for $a = 1, \ldots, K$ **do**

    Let $\hat{\pi}_t(a) = (1 - \gamma)\mathbb{E}_{w \sim p_{t-1}(w)} \hat{q}_t(a|w, x_t) + \gamma/K$

  Sample a_t according to $\hat{\pi}_t(\cdot)$

  Pull arm a_t and observe reward $r_t(a_t) \in [0, 1]$

  Let $\hat{r}_t(w, x_t, a_t) = \hat{q}_t(a_t|w, x_t)(r_t(a_t) - b)/\hat{\pi}_t(a_t)$

  Let $u_t(w) = u_{t-1}(w) \exp(\eta \hat{r}_t(w, x_t, a_t))$

  Let $p_t(w) = p_0(w)u_t(w)/\mathbb{E}_{w \sim p_0(w)} u_t(w)$

Some Intuitions

Note that conditioned on the history, the estimator $\hat{r}_t(w, x_t, a_t)$ is a random estimator that depends on the partial reward $r_t(a_t)$ received for $a_t \sim \hat{\pi}_t(a)$.

Moreover, it is an unbiased estimator of the following shifted reward of w, according to policy $\hat{q}_t(\cdot|w, x_t)$:

$$\mathbb{E}_{a_t \sim \hat{\pi}_t} \hat{r}_t(w, x_t, a_t) = \mathbb{E}_{a \sim \hat{q}_t(a|w, x_t)}(r_t(a) - b). \tag{3}$$

which relies on the full reward vector $[r_t(a)]$ at time step t over all arms a.

Theorem 4 (EXP4 Regret Bound, Thm 17.4)

For any K, $T \geq 0$, and any $\gamma \in (0, 1]$, $\eta > 0$ and $b \geq 0$. Consider any expert class $\mathcal{G} = \{[\hat{q}_t(\cdot|w, x)]_{t=1,2,\ldots} : w \in \Omega\}$ with prior $p_0(w)$. Let

$$R_T(w) = \mathbb{E} \sum_{t=1}^{T} \mathbb{E}_{a \sim \hat{q}_t(\cdot|w, x_t)} r_t(a)$$

be the reward of expert w. Then the expected reward of EXP4 satisfies:

$$\mathbb{E} \sum_{t=1}^{T} r_t(a_t) \geq (1 - \gamma) \max_q \left[\mathbb{E}_{w \sim q} R_T(w) - \frac{1}{\eta} \mathrm{KL}(q||p_0), \right] - c(\eta, b)\eta \sum_{t=1}^{T} \sum_{a=1}^{K} |r_t(a) - b|,$$

where the expectation is with respect to the randomization of the algorithm,

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

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

Finite Policy

The following result is a direct consequence of Theorem 4. It can be regarded as a direct generalization of that of EXP3.

Corollary 5 (Cor 17.5)

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

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

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

$$G_* = \arg\max_w \sum_{t=1}^{T} \mathbb{E}\, \mathbb{E}_{a\sim\hat{q}_t(\cdot|w,x_t)}[r_t(a)].$$

Proof of Corollary 5

We have $\eta \hat{r}_t(w, x_t, a_t) \leq 1$, and thus $c(\eta, b) = e - 2$. Note that the uniform random policy belongs to Ω implies that

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

We consider Theorem 4, with q defined as $q(w) = \mathbb{1}(w = w_*)$, where w_* achieves the maximum of G_*. This implies

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

This implies the bound.

Example 6 (Negative Bias)

We can take $\gamma = 0$ and $b = 1$ in Algorithm 1. By noting that $\phi(z)$ is increasing in z and $\eta(1-b) \leq 0$, we may take $c(\eta, b) = 0.5$. Theorem 4 implies that

$$\mathbb{E} \sum_{t=1}^{T} r_t(a_t) \geq \max_q \left[R_T(w) - \frac{1}{\eta} \mathrm{KL}(q||p_0) \right] - 0.5\eta KT.$$

In the finite policy case $|\Omega| = N$ with uniform prior:

$$G_* - \mathbb{E} \sum_{t=1}^{T} r_t(a_t) \leq \frac{\ln N}{\eta} + 0.5\eta KT.$$

By choosing $\eta = \sqrt{\ln N/(KT)}$, we obtain

$$G_* - \mathbb{E} \sum_{t=1}^{T} r_t(a_t) \leq 2\sqrt{KT \ln N}.$$

Partial versus Full Information

We may compare EXP4 to its full information counterpart Hedge in Algorithm 14.4. We have the following Hedge online regret bound (in the full information case) from Theorem 14.15:

$$\mathbb{E}\sum_{t=1}^{T} r_t(a_t) \geq \max_q \left[R_T(w) - \frac{1}{\eta}\mathrm{KL}(q||p_0)\right] - \eta T/8.$$

Assume $|\Omega| = N < \infty$. Let $p_0(w)$ be the uniform prior over Ω, then we obtain the following online regret bound for Hedge (full information):

$$G_* - \mathbb{E}\sum_{t=1}^{T} r_t(a_t) \leq \frac{\ln N}{\eta} + \frac{\eta T}{8}.$$

With $\eta = \sqrt{\ln N/T}$, we obtain the full information regret bound of

$$G_* - \mathbb{E}\sum_{t=1}^{T} r_t(a_t) \leq 2\sqrt{T \ln N},$$

which is better than the EXP4 result in Example 6 by a factor of $\sqrt{K}$.

Linear Contextual Bandit

The EXP4 algorithm tries to find the best policy in a policy class. We can also design an algorithm that finds the best value function from a value function class.

Definition 7

Stochastic linear contextual bandit (or stochastic contextual bandit with linear payoff) is a contextual bandit problem, where the reward at each time step t is given by

$$r_t(a) = r_t(x_t, a) = w_*^\top \psi(x_t, a) + \epsilon_t(x_t, a), \tag{4}$$

where $\epsilon_t(x, a)$ is a zero-mean random variable. We assume that $\mathcal{H}$ is a known inner product space, $w_* \in \mathcal{H}$ is the unknown model parameter, and the feature vector $\psi(x, a) \in \mathcal{H}$ is known.

Linear UCB

Algorithm 2: Linear UCB Algorithm

Input: λ, T, $\{\beta_t\}$

Let $A_0 = \lambda I$

Let $w_0 = 0$

Let $b_0 = 0$

for $t = 1, 2, \ldots, T$ **do**

  Observe x_t

  Let $a_t \in \arg\max_a \left[w_{t-1}^\top \psi(x_t, a) + \beta_{t-1} \sqrt{\psi(x_t, a)^\top A_{t-1}^{-1} \psi(x_t, a)} \right]$

  Pull arm a_t and observe reward $r_t(x_t, a_t)$

  Let $b_t = b_{t-1} + r_t(x_t, a_t)\psi(x_t, a_t)$

  Let $A_t = A_{t-1} + \psi(x_t, a_t)\psi(x_t, a_t)^\top$

  Let $w_t = A_t^{-1} b_t$

Confidence Interval Bound

Lemma 8 (Lem 17.8)

Assume that in the stochastic linear bandit model, $\|w_*\|_{\mathcal{H}} \leq B$ *for some constant B, and in Algorithm 2, assume that* $\{\beta_t\}$ *is any sequence so that*

$$\Pr\left[\forall 0 \leq t \leq T : \beta_t \geq \sqrt{\lambda}B + \left\|\sum_{s=1}^{t} \epsilon_s(x_s, a_s)\psi(x_s, a_s)\right\|_{A_t^{-1}}\right] \geq 1 - \delta. \tag{5}$$

Then with probability at least $1 - \delta$*, for all* $t = 0, \ldots, T$ *and* $u \in \mathcal{H}$*:*

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

Proof of Lemma 8

We have

$$
\begin{aligned}
u^\top (w_t - w_*) =& u^\top A_t^{-1} \sum_{s=1}^{t} r_s(x_s, a_s)\psi(x_s, a_s) - u^\top w_* \\
=& u^\top A_t^{-1} \sum_{s=1}^{t} \epsilon_s(x_s, a_s)\psi(x_s, a_s) - \lambda u^\top A_t^{-1} w_* \\
\leq& \|u\|_{A_t^{-1}} \left\| \sum_{s=1}^{t} \epsilon_s(x_s, a_s)\psi(x_s, a_s) \right\|_{A_t^{-1}} + \lambda \|u\|_{A_t^{-1}} \|w_*\|_{A_t^{-1}} \\
\leq& \|u\|_{A_t^{-1}} \left[\left\| \sum_{s=1}^{t} \epsilon_s(x_s, a_s)\psi(x_s, a_s) \right\|_{A_t^{-1}} + \sqrt{\lambda} B \right] \leq \beta_t \|u\|_{A_t^{-1}}.
\end{aligned}
$$

The second equality used (4). The first inequality used the Cauchy-Schwartz inequality. The last inequality used the definition of β_t. This implies the desired bound.

Example (Sub-Gaussian Noise)

Example 9

Assume that noise in (4) satisfies the sub-Gaussian conditions of Theorem 13.7, and assume that $d = \dim(\mathcal{H})$ is finite dimensional, with $B' = \sup_{x,a} \|\psi(x,a)\|_{\mathcal{H}}$. Then in Lemma 8 we can set

$$\beta_t = \sqrt{\lambda}B + \sigma\sqrt{2\ln(1/\delta) + d\ln(1 + T(B')^2/d\lambda)}$$

so that (5) holds. Note that Proposition 15.8 is used to obtain a bound on the log determinant function in Theorem 13.7.

Regret Bound for Algorithm 2

Theorem 10 (Thm 17.11)

Assume that in the stochastic linear bandit model, $r_t(x_t, a_t) \in [0, 1]$ *and* $\|w_*\|_2 \leq B$ *for some constant* B*. Let* $\mu_t(x, a) = \mathbb{E}_{\epsilon_t(x,a)} r_t(x, a) = w_*^\top \psi(x, a)$*. Let* $a_*(x) \in \arg\max_a \mu_t(x, a)$ *be the optimal arm for each context* x*. Then in Algorithm 2, with probability at least* $1 - \delta$*,*

$$\mathbb{E} \sum_{t=1}^{T} [\mu_t(x_t, a_*(x_t)) - \mu_t(x_t, a_t)] \leq 3 \sqrt{\ln |A_T/\lambda| \sum_{t=1}^{T} \beta_{t-1}^2} + 2 \ln |A_T/\lambda|,$$

where $\{\beta_t\}$ *is any sequence that satisfies*(5)*.*

Result used in the Proof of Theorem 10

Lemma 11 (Lem 13.9)

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

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

Proof of Theorem 10 (I/II)

We have for $t \geq 1$:

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

where the first and the third inequalities used Lemma 8 . The second inequality is due to the UCB choice of a_t in Algorithm 2.
Let E_t be the event of $\|\psi(x_t, a_t)\|_{A_{t-1}} \leq 1$. Since $w_*^\top \psi(x_t, a) \in [0, 1]$, we have

$$
w_*^\top \psi(x_t, a_*(x_t)) - w_*^\top \psi(x_t, a_t) \leq 2\beta_{t-1} \|\psi(x_t, a_t)\|_{A_{t-1}^{-1}} \mathbb{1}(E_t) + \mathbb{1}(E_t^c).
$$

Proof of Theorem 10 (II/II)

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

$$\sum_{t=1}^{T} [\mu_t(x_t, a_*(x_t)) - \mu_t(x_t, a_t)]$$

$$\leq 2 \sum_{t=1}^{T} \beta_{t-1} \|\psi(x_t, a_t)\|_{A_{t-1}^{-1}} \mathbb{1}(E_t) + \sum_{t=1}^{T} \mathbb{1}(E_t^c)$$

$$\leq 2 \sum_{t=1}^{T} \beta_{t-1} \sqrt{\frac{2\|\psi(x_t, a_t)\|_{A_{t-1}^{-1}}^2}{1 + \|\psi(x_t, a_t)\|_{A_{t-1}^{-1}}^2}} + 2 \sum_{t=1}^{T} \frac{\psi(x_t, a_t)^\top A_{t-1}^{-1} \psi(x_t, a_t)}{1 + \psi(x_t, a_t)^\top A_{t-1}^{-1} \psi(x_t, a_t)}$$

$$\leq 3 \sqrt{\sum_{t=1}^{T} \beta_{t-1}^2} \sqrt{\sum_{t=1}^{T} \frac{\psi(x_t, a_t)^\top A_{t-1}^{-1} \psi(x_t, a_t)}{1 + \psi(x_t, a_t)^\top A_{t-1}^{-1} \psi(x_t, a_t)}} + 2 \ln |A_T / \lambda|.$$

The second inequality used simple algebraic inequalities under E_t and E_t^c. The third inequality used the Cauchy-Schwartz inequality and Lemma 11. We can now apply Lemma 11 again to obtain the desired bound.

Interpretation of Theorem 10

We may consider the noise assumption and the choice of β_t in Example 9 with $\sigma = O(1)$. It implies a bound

$$\mathbb{E} \sum_{t=1}^{T} [\mu_t(x_t, a_*(x_t)) - \mu_t(x_t, a_t)] = \tilde{O}\left(\sqrt{\lambda d T} B + d\sqrt{T}\right),$$

where $\tilde{O}$ hides logarithmic factors. Proposition 15.8 is used to obtain a bound on the log determinant function in Theorem 10.
Set λ to a small number, we obtain

$$\mathbb{E}\, \mathrm{REG}_T = \mathbb{E} \sum_{t=1}^{T} [\mu_t(x_t, a_*(x_t)) - \mu_t(x_t, a_t)] = \tilde{O}(d\sqrt{T}).$$

Matching Lower Bound

Theorem 12 (Thm 17.13)

Given any integer $d \geq 1$ and $T \geq 1$, there exists a (noncontextual) stochastic linear bandit problem with 2^d arms corresponding to feature vectors $\{\pm 1\}^d$, and reward $r \in [-0.5, 0.5]$. So that regret of any bandit algorithm is at least

$$\min\left(0.05T, 0.12d\sqrt{T}\right).$$

Proof of Theorem 12 (I/III)

Consider 2^d arms, represented by feature vectors $\psi(a) = a \in \{-1, 1\}^d$. The reward r of pulling arm a (without context) is in $\{-0.5, 0.5\}$, and

$$\mathbb{E}[r|a] = w^\top a$$

for some $w \in \{-\epsilon, \epsilon\}^d$, where $\epsilon \in (0, 0.5/d]$ will be specified later.

Using notations of Theorem 13.24 with τ changed to w and $m = 2$, $\mathcal{P}_\mathcal{Z} = \{q_w(r|a)\}$, where each $q_w(r|a)$ is a $\{-0.5, 0.5\}$ valued binary random variable $\mathrm{Bernoulli}(0.5 + w^\top a) - 0.5$. A policy π is a probability distribution on $\mathcal{A}$, and we can define $q_w(r|\pi) = \mathbb{E}_{a\sim\pi} q_w(r|a)$.

Let θ indicate an arbitrary arm returned by a learning algorithm, represented by its feature vector $\theta \in \{\pm 1\}^d$. It follows that the regret of pulling arm θ is

$$Q(\theta, w) = \sum_{j=1}^{d} Q_j(\theta, w), \quad Q_j(\theta, w) = \epsilon - w_j\theta_j.$$

Proof of Theorem 12 (II/III)

This means that for $w \sim_j w'$ and $w \neq w'$ ($w' \sim_j w$ means that w' and w are identical except at the j-th component):

$$Q_j(\theta, q_w) + Q_j(\theta, q_{w'}) \geq 2\epsilon.$$

Let $w^{(j)} = w - w_j e_j$ be the vector with value zero at the j-th component but the same value as that of w elsewhere.
Now we can let $\epsilon = \min\left(0.1/d, 0.24\sqrt{1/T}\right)$. Given any learning algorithm $\hat{q}$, for all w, time step t, and a_t represented by feature representation in $\{-1, 1\}^d$:

$$\begin{aligned}
&\frac{1}{2}\sum_{w' \sim_j w} \mathrm{KL}(q_{w^{(j)}}(\cdot|\hat{q}(\mathcal{S}_{t-1}), \mathcal{S}_{t-1})||q_{w'}(\cdot|\hat{q}(\mathcal{S}_{t-1}), \mathcal{S}_{t-1})) \\
\leq &\frac{1}{2}\sup_{a_t}\sum_{w' \sim_j w} \mathrm{KL}(0.5 + (w^{(j)})^\top a_t||0.5 + (w')^\top a_t) \\
\leq &2.1\epsilon^2.
\end{aligned}$$

Proof of Theorem 12 (III/III)

We can now take $\beta_{j,t}^2 = 2.1\epsilon^2$ and apply Theorem 13.24. For $n \leq T$,

$$\begin{aligned}\frac{1}{2^d}\sum_w \mathbb{E}_{\theta,\mathcal{S}_n\sim p(\cdot|\hat{q},q_w)} Q(\theta, q_w) \geq & d\epsilon\left(1 - \sqrt{2 \times 2.1 n\epsilon^2}\right) \\ \geq & 0.5d\epsilon = 0.5d \min\left(0.1/d, 0.24\sqrt{1/T}\right).\end{aligned}$$

Since this holds for all $n \leq T$, we obtain the bound.

Suboptimality for Finite Arm Problems

Example 13

The stochastic multi-armed bandit with K arms and rewards in $[0, 1]$ can be considered as a stochastic linear bandit, where we take $w_* = [\mu(1), \ldots, \mu(K)]$, and $\psi(a, x) = e_a$ for $a \in \{1, \ldots, K\}$. Therefore we may chose $B = \sqrt{K}$, so that $\|w_*\|_2 \leq B$. We can also choose $\lambda = 1$ and $M = 1$. Theorem 10 implies

$$\mathbb{E}\mathrm{REG}_T = \mathbb{E}\sum_{t=1}^{T}[\mu_t(x_t, a_*(x_t)) - \mu_t(x_t, a_t)] = \tilde{O}(K\sqrt{T}),$$

which is suboptimal by a $\sqrt{K}$ factor (ignoring log factors). In comparison, for noncontextual stochastic linear bandit, Algorithm 16.2 achieves a better regret of

$$\tilde{O}(\sqrt{KT})$$

according to Theorem 16.14 and Example 16.15.

Nonlinear Stochastic Bandits

Definition 14

The 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, where we assume that $f_*(x, a) \in \mathcal{F}$ for a known function class $\mathcal{F} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$. Given any $f(x, a) \in \mathcal{F}$, we also define

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

and the greedy policy of f as:

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

Nonlinear UCB

Algorithm 3: Version Space UCB Algorithm

Input: λ, T, $f_0 \in \mathcal{F}$

Let $\mathcal{F}_0 = \{f_0\}$

for $t = 1, 2, \ldots, T$ **do**

  Observe x_t

  Let $f_t \in \arg\max_{f \in \mathcal{F}_{t-1}} f(x_t)$

  Let $a_t \in \arg\max_a f_t(x_t, a)$

  Pull arm a_t and observe reward $r_t(x_t, a_t) \in [0, 1]$

  Let $\mathcal{F}_t$ be an appropriate version space based on $\mathcal{S}_t = \{(x_s, a_s)\}_{s=1}^t$.

In general, we say $\mathcal{F}_t$ is a version space if $f_* \in \mathcal{F}_t$ with high probability. Choosing the optimal f_t in a properly defined version space is a natural generalization of upper confidence bound.

Nonlinear UCB Generalizes Linear UCB

Proposition 15

Assume that $\mathcal{F} = \{f(w, x, a) = w^\top \psi(x, a) : w \in \mathbb{R}^d\}$. *Let*

$$\mathcal{F}_t = \left\{f(w, \cdot) : \phi_t(w) \le \phi_t(w_t) + \beta_t^2\right\},$$

where $w_t = \arg\min_w \phi_t(w)$, *and*

$$\phi_t(w) = \sum_{s=1}^{t} (w^\top \psi(x_s, a_s) - r_s(x_s, a_s))^2 + \lambda \|w\|_2^2.$$

Then Algorithm 3 is equivalent to Algorithm 2. In particular, we have

$$\mathcal{F}_{t-1} = \{f(w, x, a) : \|w - w_{t-1}\|_{A_{t-1}} \le \beta_{t-1}\},$$

and

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

Analysis of Nonlinear UCB: Eluder Coefficient

Definition 16

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

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

Property of Eluder Coefficient

Proposition 17 (Prop 17.20)

Assume that $\mathcal{F} \subset \mathcal{H}$, where $\mathcal{H}$ is a RKHS which does not need to be known to the learning algorithm. For all $f \in \mathcal{H}$, we have the feature representation $f(x,a) = \langle w(f), \psi(x,a) \rangle$. Assume $\|w(f) - w(f_*)\|_{\mathcal{H}} \leq B$ for all $f \in \mathcal{F}$ and $f - f_* \in [-1,1]$ for all $f \in \mathcal{F}$. Then

$$\mathrm{EC}(1, \mathcal{F}, T) \leq 2\mathrm{entro}(1/(B^2 T), \psi(\mathcal{X} \times \mathcal{A})), \tag{6}$$

where $\mathrm{entro}(\cdot)$ *is defined as*

$$\mathrm{entro}(\lambda, \psi(\mathcal{X} \times \mathcal{A})) = \sup_{\mathcal{D}} \ln \left| I + \frac{1}{\lambda} \mathbb{E}_{(x,a) \sim \mathcal{D}} \psi(x,a) \psi(x,a)^\top \right|.$$

Analysis of Nonlinear UCB

Lemma 18 (Lem 17.18)

In Algorithm 3, assume that $f_* \in \mathcal{F}_{t-1}$ for all $t \leq T$, and there exists $\hat{f}_t$ and $\beta_t > 0$ such that

$$\sup_{f \in \mathcal{F}_t} \sum_{s=2}^{t} |f(x_s, a_s) - \hat{f}_t(x_s, a_s)|^2 \leq \beta_t^2.$$

Then we have the following regret bound:

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

Proof of Lemma 18

We have

$$\begin{aligned} & f_*(x_t) - f_*(x_t, a_t) \\ =& f_*(x_t) - f_t(x_t) + f_t(x_t, a_t) - f_*(x_t, a_t) \\ \leq& f_t(x_t, a_t) - f_*(x_t, a_t). \end{aligned}$$

The first equality used $f_t(x_t, a_t) = f_t(x_t)$. The inequality used the fact that $f_* \in \mathcal{F}_{t-1}$ and thus $f_t(x_t) = \max_{f \in \mathcal{F}_{t-1}} f(x_t) \geq f_*(x_t)$.

We can now obtain

$$\sum_{t=2}^{T} [f_*(x_t) - f_*(x_t, a_t)] \leq \sum_{t=2}^{T} [f_t(x_t, a_t) - f_*(x_t, a_t)]$$

$$\leq \sqrt{\mathrm{EC}(\epsilon, \mathcal{F}, T) \sum_{t=2}^{T} \left(\epsilon + \sum_{s=1}^{t-1} |f_t(x_s, a_s) - f_*(x_s, a_s)|^2 \right)}$$

$$\leq \sqrt{\mathrm{EC}(\epsilon, \mathcal{F}, T) \left(\epsilon T + 4 \sum_{t=2}^{T} \beta_{t-1}^2 \right)}.$$

Regret Bound for Nonlinear UCB

Theorem 19 (Thm 17.19)

Assume that $r_t = f_*(x_t, a_t) + \epsilon_t$*, where* ϵ_t *is conditional zero-mean sub-Gaussian noise: for all* $\lambda \in \mathbb{R}$, $\ln \mathbb{E}[e^{\lambda \epsilon_t} | X_t, \mathcal{F}_{t-1}] \leq \frac{\lambda^2}{2}\sigma^2$*. In Algorithm 3, we define*

$$\hat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t} (f(x_s, a_s) - r_s)^2,$$

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

where $\beta_t^2 \geq \inf_{\epsilon > 0} \left[8\epsilon t(\sigma + 2\epsilon) + 12\sigma^2 \ln(2N(\epsilon, \mathcal{F}, \|\cdot\|_\infty)/\delta) \right]$*. Then with probability at least* $1 - \delta$*:*

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

Example

Example 20

If $\mathcal{F} \subset \mathcal{H} = \{w^\top \psi(x, a) : w \in \mathbb{R}^d\}$ can be embedded into a d dimensional linear function class for a finite d, then we have the following bound from Proposition 17 and Proposition 15.8:

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

Since Theorem 5.3 implies that the covering number of $\mathcal{F}$ is also $\tilde{O}(d)$, we can obtain the following regret bound from Theorem 19:

$$\sum_{t=2}^{T} [f_*(x_t) - f_*(x_t, a_t)] = \tilde{O}(d\sqrt{T}),$$

which is consistent with that of Theorem 10. One may also use Proposition 15.8 to obtain bounds for nonparametric models such as RKHS induced by RBF kernels (also see Example 15.10).

General Nonlinear Bandits with Finite Arms

Consider the Nonlinear UCB algorithm with function class $\mathcal{F}$. The regret bound in Theorem 19 requires that the eluder coefficient of the function class is bounded.

Assumption 21

Assume that the action space $\mathcal{A}$ is finite: $|\mathcal{A}| = K$, and $\mathcal{F}$ is also finite: $|\mathcal{F}| = N$. Moreover, assume realizability: $f_* \in \mathcal{F}$.

We show that under Assumption 21, it is possible to design a bandit algorithm with regret bound which does not depend on the eluder coefficient of the function class.

Feel Good Thompson Sampling

Algorithm 4: Feel-Good Thompson Sampling for Contextual Bandits

Input: p_0, T

for $t = 1, 2, \ldots, T$ **do**

Observe $x_t \in \mathcal{X}$

Draw $f_t \sim p(f|S_{t-1})$ according to

$$p(f|S_{t-1}) \propto p_0(f) \exp\left(-\sum_{s=1}^{t-1} L(f, x_s, a_s, r_s)\right), \tag{7}$$

where $L(f, x, a, r) = -\lambda f(x) + \eta(f(x, a) - r)^2$ and p_0 is a prior on $\mathcal{F}$

Pull arm $a_t = \pi_{f_t}(x_t) \in \arg\max_a f_t(x_t)$

Observe reward $r_t \in [0, 1]$

Regret Bound

Theorem 22 (Simplification of Thm 17.30)

Under Assumption 21 ($|\mathcal{F}| = N$, $|\mathcal{A}| = K$, and $f_* \in \mathcal{F}$). Assume that $f \in [0, 1]$ for all $f \in \mathcal{F}$, and $\eta \leq 0.5$, then the following bound holds for Algorithm 4:

$$\mathbb{E}\,\mathrm{REG}_T \leq \frac{\ln N}{\lambda} + \frac{\lambda T}{4} + \frac{\lambda K}{\eta} T.$$

Note: Theorem 17.30 can also handle infinite action and infinite function class. Taking $\eta = 0.5$ and $\lambda = \sqrt{\ln N/(KT)}$, we obtain

$$\mathbb{E}\,\mathrm{REG}_T = O(\sqrt{KT \ln N}).$$

It doesn't depend on the eluder coefficient of $\mathcal{F}$, and matches the lower bound up to log factors.

Lower Bound: Nonlinear Bandit with Finite Arms

Theorem 23 (Thm 17.33)

Consider $K \geq 2$ and $d \geq 1$. There exists a bandit problem with $\{0, 1\}$ valued rewards, such that the realizable condition holds with $|\mathcal{A}| = K$, $|\mathcal{F}| = K^d$, so that the expected regret of any bandit algorithm $\hat{q}$ is at least

$$\min\left(0.02T, 0.06\sqrt{KdT}\right).$$

Pure Exploration Problem

Consider a stochastic contextual bandit problem, in which the context comes from a fixed distribution: $x \sim \mathcal{D}$.
Let π_E be a bandit policy, referred to as an *exploration policy*.

Definition 24

Given any integer T, the goal of the pure exploration problem in contextual bandit is to design an exploration policy π_E, and draw T samples

$$x_t \sim \mathcal{D}, \quad a_t \sim \pi_E(\cdot|x_t), \quad (t = 1, \ldots, T),$$

so that one can learn a bandit policy $\hat{\pi}$ from the samples with small regret defined below:

$$\mathrm{REG}(\hat{\pi}) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{a\sim\hat{\pi}(\cdot|x)}[f_*(x) - f_*(x,a)],$$

where f_* is the true value function.

Regret Bound for Pure Exploration

Proposition 25 (Simplified from Prop 17.37)

Under Assumption 21 ($|\mathcal{F}| = N$, $|\mathcal{A}| = K$ and $f_* \in \mathcal{F}$). Assume that the context x are drawn from a fixed distribution $\mathcal{D}$ on $\mathcal{X}$. Let π_E be the exploration policy that draws action uniformly at random, then for all $f \in \mathcal{F}$:

$$\mathrm{REG}(\pi_f) \leq +2\sqrt{K\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{a\sim\pi_E(\cdot|x)}(f(x,a) - f_*(x,a))^2}.$$

where π_f is the greedy policy of f: $\pi_f(x) = \arg\max_a f(x,a)$.

The result reduces pure exploration contextual bandit to supervised least squares regression. Since Gibbs algorithm output $\hat{f}$ satisfies

$$\mathbb{E}\,\mathbb{E}_{a\sim\pi_E(\cdot|x)}(\hat{f}(x,a) - f_*(x,a))^2 = O(\ln N/T),$$

we obtain the following result for Gibbs algorithm:

$$\mathbb{E}\,\mathrm{REG}(\hat{\pi}) = O(\sqrt{K \ln N/T}),$$

which matches the bound for Feel Good Thompson Sampling.

Proof of Proposition 25

We note that

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[f_*(x) - f_*(x,\pi_f(x))\right] \\
\leq&\mathbb{E}_{x\sim\mathcal{D}}\left[f_*(x,\pi_{f_*}(x)) - f(x,\pi_{f_*}(x)) + f(x,\pi_f(x)) - f_*(x,\pi_f(x))\right] \\
\leq&2\mathbb{E}_{x\sim\mathcal{D}}\sup_{a\in\mathcal{A}}\left|f(x,a) - f_*(x,a)\right| \\
\leq&2\sqrt{\mathbb{E}_{x\sim\mathcal{D}}\sup_{a\in\mathcal{A}}\left(f(x,a) - f_*(x,a)\right)^2} \\
\leq&2\sqrt{K\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{a\sim\pi_E(\cdot|x)}\left(f(x,a) - f_*(x,a)\right)^2}.
\end{aligned}
$$

The first inequality used the fact that $0 \leq -f(x,\pi_{f_*}(x)) + f(x,\pi_f(x))$, which follows from the definition of greedy policy π_f. The third inequality used Jensen's inequality and the concavity of $\sqrt{\cdot}$. This implies the desired bound.

Summary (Chapter 17)

- Contextual bandit
  - adversarial and stochastic
- Policy based Method
- EXP4
  - control exploration with randomization or reward bias
- Linear Contextual Bandit
  - value based
- Linear UCB
  - uniform confidence interval
  - handles infinite many arms but can be suboptimal for finite arm case
- Nonlinear UCB
- Thompson Sampling and Pure Exploration