CSC 588: Machine learning theory **Spring 2021**

Lecture 27: Stochastic linear contextual bandits:LinUCB/OFUL algorithm

Lecturer: Chicheng Zhang *Scribe: Dan Li*

1 Stochastic linear contextual bandits

Stochastic linear contextual bandits setup:

For all $t = 1, 2, \ldots, T$:

- Observe context x_t in $\mathcal{X}$ (context space)
- Take action a_t in $\mathcal{A} = \{1, ...K\}$ (action space)
- Receive reward $r_t = f(x_t, a_t) + \epsilon_t$

Here ϵ_t's are independent Gaussian with mean 0 and variance 1. $f(x_t, a_t) = \langle \theta^*, \phi(x, a) \rangle$, where θ^* is unknown and ϕ is known. Our goal is to maximize the expectation of reward:

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} f(x_t, a_t)\right]$$

Performance measure of the pseudo-regret is as following

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{A}} f(x_t, a) - \sum_{t=1}^{T} f(x_t, a_t)\right]$$

Assume :$||\theta^*||_2 \leq 1$, for every x and a, $||\phi(x, a)||_2 \leq 1$, now we have two questions to explore:

1. how can we design a good algorithm such that the pseudo-regret is small? Recall that if we know θ^* exactly, we can simply take the $a_t = \text{argmax}_{a \in \mathcal{A}} \langle \theta^*, \phi(x, a) \rangle$. Can we estimate θ^*?
   Specifically, At round t, we know allx_s, a_s, r_s for all $s = 1, 2, ...t - 1$. Given these information, can we construct a good estimator of θ^*, denoted by $\hat{\theta}_t$?
2. given $\hat{\theta}_t$,how to take actions?
   $a_t = \text{argmax}_a \langle \hat{\theta}_t, \phi(x_t, a) \rangle + exploration(a)$, find a that has the biggest upper confidence bound

1.1 How to accurately estimate θ

We can view this problem as a regression problem. We use the following ridge estimator:

$$\theta_t(\lambda) = \underset{\theta \in R^d}{\text{argmin}} \sum_{s=1}^{t-1} (\langle \theta, \phi_s \rangle - r_s)^2 + \lambda ||\theta||_2^2$$

Informally we have the following result:

$$\theta_t(0) = \theta^* + V_{t-1}^{-1} (\sum_{s=1}^{t-1} \phi_s \epsilon_s)$$

Since ϵ_s is Gaussian with variance 1,we should have $V_{t-1}^{-1}(\sum_{s=1}^{t-1}\phi_s\epsilon_s)$ as Gaussian with variance V_{t-1}^{-1} , we should also have $\langle\theta_t(0),\phi(x,a)\rangle$ distributed as a Gaussian with mean $\langle\theta^*,\phi(x,a)\rangle$ and variance $||\phi(x,a)||^2_{V_{t-1}^{-1}}$.Given new x and a, we should be able to construct a confidence interval of $\langle\theta^*,\phi(x,a)\rangle$ based on $\theta_t(0)$ as follows. Since

$$|\langle\theta_t(0),\phi(x,a)\rangle - \langle\theta^*,\phi(x,a)\rangle| \leq ||\phi(x,a)||_{V_{t-1}^{-1}}\sqrt{\ln\frac{1}{\sigma}}$$

With probability $1-\sigma$,By sub-Gaussian tail property. Therefore we can define a high probability upper confidence bound of $f(x_t,a)$ as:

$$UCB_t(a) = \langle\theta_t(0),\phi(x,a)\rangle + ||\phi(x,a)||_{V_{t-1}^{-1}}\sqrt{\ln\frac{1}{\sigma}}$$

However, the above reasoning is inherently flawed because of the following problem:

1. ϕ_t is somehow dependent on ϵ_t through a_t
2. V_{t-1} may not be full rank.

There are papers devoted to solve the first problem using Specialized binning procedure which we won't elaborate. For problem 2 we can simply set $\lambda > 0$ in $\hat{\theta}_t(\lambda)$ – specifically we set $\lambda = 1$. Then we will work with $V_{t-1}(1) = V_{t-1} + I$, as our ridge estimator is $\hat{\theta}_t(1) = V_{t-1}(1)^{-1}(\sum_{s=1}^{t-1}\phi_s\epsilon_s)$. In this case we can guarantee that the matrix $V_{t-1}(1)$ is full rank. The new bound for $\hat{\theta}_t(1)$ will be given in the following lemma:

Lemma 1. *define*

$$\beta_t(\sigma) = 1 + \sqrt{2\ln\frac{1}{\sigma} + d\ln(1+\frac{t}{d})}$$

then exists event E, with probability $P(E) \geq 1-\sigma$, and on E we have

$$||\hat{\theta}_t(1) - \theta^*||_{V_{t-1}(1)} \leq \beta_t(\sigma)$$

for all t.

We now demonstrate the proof of this lemma (Aner, et al. 2002).

$$\begin{aligned}
\hat{\theta}_t(1) - \theta^* &= V_t^{-1}(1)(V_{t-1}\theta^* + \sum_{s=1}^{t-1}\phi_s\epsilon_s) - \theta^* \\
&= V_t^{-1}(1)(V_{t-1}\theta^* + \sum_{s=1}^{t-1}\phi_s\epsilon_s) - V_t^{-1}(V_{t-1}+I)\theta^* \qquad (1)\\
&= -V_t^{-1}(1)\theta^* + V_{t-1}^{-1}(1)\sum_{s=1}^{t-1}\phi_s\epsilon_s
\end{aligned}$$

Applying the Mahalanobis norm

$$\begin{aligned}
||\hat{\theta}_t(1) - \theta^*||_{V_{t-1}(1)} &\leq ||V_t^{-1}(1)\theta^*||_{V_{t-1}(1)} + ||V_{t-1}^{-1}(1)\sum_{s=1}^{t-1}\phi_s\epsilon_s||_{V_{t-1}(1)} \\
&= ||\theta^*||_{V_{t-1}(1)} + ||\sum_{s=1}^{t-1}\phi_s\epsilon_s||_{V_{t-1}^{-1}(1)} \qquad (2)\\
&\leq 1 + ||\sum_{s=1}^{t-1}\phi_s\epsilon_s||_{V_{t-1}(1)}
\end{aligned}$$

To finish the proof we need to introduce another lemma called self-normalized bound

Lemma 2. *there exists an event E, with probability $P(E) \geq 1 - \sigma$, we have the following inequality:*

$$||\sum_{s=1}^{t} \phi_s \epsilon_s||_{V_t^{-1}(\lambda)} \leq \sqrt{a \ln \frac{1}{\sigma} + d \ln 1 + \frac{t}{d}}$$

Using the above lemma with $\lambda = 1$, can conclude that with probability $1 - \sigma$,

$$||\hat{\theta}_t(1) - \theta^*||_{V_{t-1}(1)} \leq \beta_t(\sigma) = 1 + \sqrt{2 \ln \frac{1}{\sigma} + d \ln(1 + \frac{t}{d})}$$

which finishes the proof of Lemma 1.

1.2 How to take actions after estimating θ^*

The question now is how do we use the previous lemma to define a Upper confidence bound for actions' rewards. To do that we define a set Θ_t as following:

$$\Theta_t = \theta : ||\hat{\theta}_t(1) - \theta||_{V_{t-1}(1)} \leq \beta_t(\sigma)$$

we know for event E, the θ^* lies in this set. At step t we want to construct a UCB for $\langle \theta^*, \phi(x_t, a) \rangle$ for all a.

The UCB of step t for action a is defined as following:

$$UCB_t(a) = \max_{\theta \in \Theta_t} \langle \theta, \phi(x_t, a) \rangle$$

After simplification and derivation we have

$$UCB_t(a) = \langle \hat{\theta}_t(1), \phi(x_t, a) \rangle + \beta_t(\sigma) * ||\phi(x_t, a)||_{V_{t-1}^{-1}}$$

The first term of the previous equation is considered as predicted reward and the second term is the uncertainty for this prediction.

The algorithm will simply be the following:

Algorithm 1 LinUCB / OFUL for Stochastic linear contextual bandits

for $t = 1, 2, \cdots, T$ **do**
  $a_t = \text{argmax}_{a \in A} UCB_t(a)$
end for

2 Analysis of LinUCB

We state the main theorem as following:

Theorem 3. *For LinUCB we have*

$$R_T \leq \mathcal{O}(T\sigma + \beta_T(\sigma) \sqrt{dT \ln(1 + \frac{T}{d})})$$

If we set $\sigma = \frac{1}{T}$, $R_T \leq \hat{\mathcal{O}}(d\sqrt{T})$.

Note that for MAB define $\phi(x, a) = e_a$ the bound is $K\sqrt{T}$, and recall that our direct analysis of UCB gives regret $\sqrt{KT}$, which is of lower order.

Proof. On event $\bar{E}$,contributions to R_T is bounded by $2T\sigma$. On event E we have the instantaneous regret as:

$$\begin{aligned}\rho_t &= \max_a f(x_t, a) - f(x_t, a_t) \\ &\leq \max_a UCB_t(a) - f(x_t, a_t) \\ &\leq UCB_t(a_t) - f(x_t, a_t) \\ &\leq \langle \hat{\theta}_t, \phi_t \rangle + \beta_t(\delta)||\phi_t||^{-1}_{V_{t-1}} - \langle \theta^*, \phi_t \rangle\end{aligned} \tag{3}$$

Applying the Cauchy Schwartz in equality and also the fact that$||\hat{\theta}_t - \theta^*||_{V_{t-1}(1)} \leq \beta_t$ we have the firs term and the last term bounded by the middle term using the V_{t-1}norm.

$$\langle \hat{\theta}_t, \phi_t \rangle - \langle \theta^*, \phi_t \rangle \leq \beta_t(\delta)||\phi_t||^{-1}_{V_{t-1}} \tag{4}$$

so we have the following:

$$\rho_t = \max_a f(x_t, a) - f(x_t, a_t) \leq 2\beta_t(\delta)||\phi_t||^{-1}_{V_{t-1}} \tag{5}$$

since $V_t(1) \preceq 2V_{t-1}(1)$ (left as an exercise), we have

$$\rho_t \leq 4\beta_t(\sigma)||\phi_t||_{V_t^{-1}(1)}$$

Summing all the instantaneous regrets, we have

$$\sum_{t=1}^{T} \rho_t \leq 4\beta_t(\sigma) \sum_{t=1}^{T} ||\phi_t||_{V_t^{-1}(1)}$$

Using Cauchy Schwartz and Elliptic potential we have

$$\sum_{t=1}^{T} \rho_t \leq 4\beta_t(\sigma)\sqrt{Td\ln(1 + \frac{T}{d})}$$

This concludes the proof.

□