**STATS 710 – Seq. Dec. Making with mHealth Applications** Oct 11, 2016

# Lecture 10: Contextual Bandits

*Instructors: Susan Murphy and Ambuj Tewari* *Scribe: Peng Liao*

## 1 Contextual Bandits Problem

Contextual bandit problems are also known as "bandit problems with side-information", "bandit problems with covariates", and "associative reinforcement learning".

The first paper on contextual bandit was written by Woodroofe in 1979 [Woo79]. The term "Contexual Bandit" was coined by Langford and Zhang in 2008 [LZ08]. The general flow of information for a multi-armed bandit with contexts drawn from a set $\mathcal{X}$ is:

1: **for** $t = 1$ to $T$ **do**
2: Learner sees $X_t \in \mathcal{X}$
3: Learner selects action $A_t \in \mathcal{A}$
4: Learner receives reward $R_t = R_t^{A_t}$
5: **end for**

## 2 Definition of Regret

There are different ways to define the regret in contextual bandit setting. In the stochastic setting, we assume that the context-reward pair $(X_t, R_t) = \{X_t, R_t^a, a \in \mathcal{A}\}$ are i.i.d. across time. Given a joint distribution $D$ on $(X, R) = \{X, R^a, a \in \mathcal{A}\}$ and a class of (deterministic) policies $\Pi \subseteq \mathcal{A}^{\mathcal{X}}$, the expected regret of learning algorithm $\mathcal{L}$ is defined as

$$\mathcal{R}_T(\mathcal{L}, D, \Pi) := \sup_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} R_t^{\pi(X_t)} - \sum_{t=1}^{T} R_t^{A_t}\right] = TV(\pi^*) - \mathbb{E}\left[\sum_{t=1}^{T} R_t^{A_t}\right]$$

where $V(\pi) := \mathbb{E}_{(X,R)\sim D}\left[R^{\pi(X)}\right]$ is the value function and $\pi^* = \text{argmax}_{\pi \in \Pi} V(\pi)$ is the optimal policy of the class $\Pi$. One can also generalize the regret competing with stochastic policies. The learner could have access to the policy class he is competing against, but not the underlying distribution.

As opposed to the stochastic MAB, here the notation of regret is with respect to the best policy in a class of policies rather than the best arm. Even though it may seem like contextual bandits is an easier problem than MAB (we have additional information), it is in fact a harder problem from the regret perspective (we want to compete against policies, not pulling the same optimal arm).

There are two main approaches (or strategies) to solve the contextual bandits problem, depending on the assumptions one is willing to make. In the first approach, one does not want to make any simplifying assumption on the underlying distribution $D$ of $(X, R)$, but only wants to compete against a relatively simple class of policies. This is similar to a classification problem: the classifier now becomes policy and we are interested in quantifying the risk w.r.t. the best classifier in the model. On the other hand, if one is comfortable putting some structure on the distribution $D$, e.g. assuming a linear model $\mathbb{E}[R|X, A] = \beta' f(X, A)$, we could then compete with the optimal policy over all possible policies. This is the second approach and is related to a regression problem in the sense that the learner needs to estimate the mean function $\mathbb{E}[R|X, A]$ and use it to take actions.

# 3 Reducing to MAB

When the context space $\mathcal{X}$ is finite (and small), one natural idea is to reduce the contextual bandits to a multiarm bandits problem. To be specific, suppose we have assess to a MAB algorithm with (non-stochastic) regret bound of $f(K)g(T)$, e.g. $f(K) = \sqrt{K \log K}$ and $g(T) = \sqrt{T}$ in EXP3 when rewards are within $[0, 1]$, and run $|\mathcal{X}|$ copies of this algorithm, one for each element of $\mathcal{X}$. The expected regret against all policies can be bounded by

$$\begin{aligned}
&\max_{\pi:\mathcal{X}\to\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} R_t^{\pi(X_t)} - \sum_{t=1}^{T} R_t^{A_t}\right] \\
&= \max_{\pi:\mathcal{X}\to\mathcal{A}} \mathbb{E}\left[\sum_{x\in\mathcal{X}} \sum_{t:X_t=x} (R_t^{\pi(X_t)} - R_t^{A_t})\right] \\
&= \max_{\pi:\mathcal{X}\to\mathcal{A}} \sum_{x\in\mathcal{X}} \mathbb{E}\left[\sum_{t:X_t=x} R_t^{\pi(x)} - R_t^{A_t}\right] \\
&= \sum_{x\in\mathcal{X}} \max_{a\in\mathcal{A}} \mathbb{E}\left[\sum_{t:X_t=x} R_t^{a} - R_t^{A_t}\right] \qquad (1) \\
&\le \sum_{x\in\mathcal{X}} f(K)g(T_x) \qquad (2)
\end{aligned}$$

where $T_x = \sum_{t=1}^{T} \mathbf{1}_{\{X_t=x\}}$ is the number of rounds that we see context $x$ in $T$ round, the equality (1) holds because the maximization is over all policies and thus we can exchange the sum and max, and we use the regret bound of the underlying MAB algorithm to get inequality (2). Note however that MAB copy for context $x$ runs for a random number $T_x$ of time steps.

Now assuming that $g(T)$ is concave, then applying Jensen's inequality gives

$$\begin{aligned}
\sum_{x\in\mathcal{X}} f(K)g(T_x) = f(K)\sum_{x\in\mathcal{X}} g(T_x) &= f(K)|\mathcal{X}| \sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|} g(T_x) \\
&\le f(K)|\mathcal{X}| g\left(\sum_x \frac{1}{|\mathcal{X}|} T_x\right) \\
&= f(K)|\mathcal{X}| g\left(\frac{T}{|\mathcal{X}|}\right)
\end{aligned}$$

Thus in the case of EXP3, we get $\sqrt{K \log K}\sqrt{T|\mathcal{X}|}$. One can show that the factor $\sqrt{|\mathcal{X}|}$ can not be improved by using the lower bound of MAB and thus the above bound is tight (up to the $\sqrt{\log K}$ factor).

## References

[LZ08] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

[Woo79] Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.