E1 245: Online Prediction & Learning **Fall 2014**

Lecture 15 — September 23

Lecturer: Aditya Gopalan *Scribe: Aadirupa Saha*

15.1 Recap

In the last lecture, we have seen the minimax regret bound for adversarial multi-armed bandits. In this lecture, we will study bandits with side informations which are popularly known as contextual bandits or associative bandits. Towards the end of this lecture, we will also analyze a contextual variant of adversarial multi-armed bandit problem called bandits with expert advice.

15.2 (Adversarial) Contextual Multi-Armed Bandits

Adversarial contextual multi-armed bandits (C-MAB) can be seen as a natural extension of adversarial multi-armed bandit (MAB) problems where at each round $t = 1, 2, \cdots$, the environment reveals a context s_t from a set of contexts $\mathscr{S}$, and based on which the learner selects a randomized action I_t from the set of actions $[N]$.[1] The objective of the learner is to minimize its (expected) regret w.r.t the best policy or function $g : \mathscr{S} \mapsto [N]$ that maps each context to some action. A more formal description of adversarial C-MAB is given below:

(Adversarial) C-MAB
Inputs:
Set of contexts or features: $\mathscr{S}$
Set of actions or decisions: $[N]$
For each round $t = 1, 2, \ldots$
– Environment reveals $s_t \in \mathscr{S}$
– Learner selects an action $I_t \in [N]$
– Learner suffers loss $l(I_t, t)$
End

The expected regret of the learner at the end of round T is defined as:

$$R_T := \max_{s_1, s_2, \ldots s_T} \max_{g: \mathscr{S} \mapsto [N]} \mathbb{E}\left[\sum_{t=1}^{T} l(I_t, t) - \sum_{t=1}^{T} l\big(g(s_t), t\big)\right]$$

[1] $[n] = \{1, 2, \cdots, n\}$.

Remarks:

1. One of the important application of C-MAB can be personalization of advertisements in recommender systems. The task of the system is to recommend an advertisement to each of its user, and the system gets rewarded whenever a user clicks on the recommended advertisement. Here the system plays the role of a learner, and each advertisement corresponds to one action. The recorded history of each user serves as the contextual information based on which the system recommends advertisements to its users.

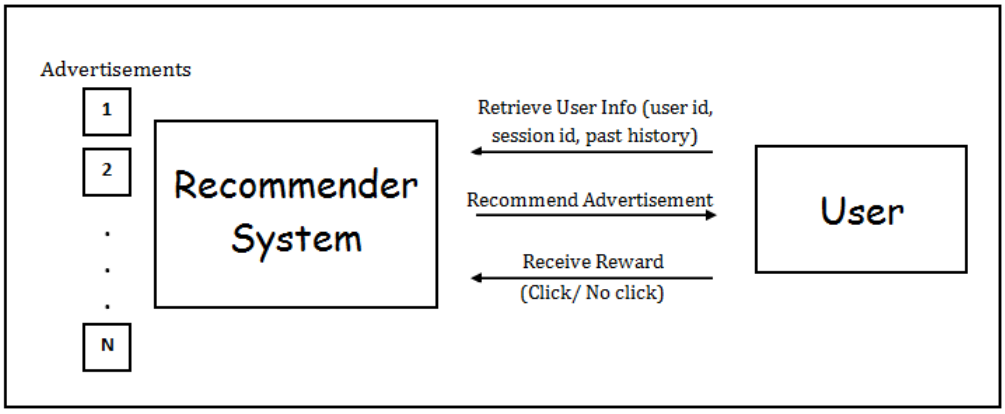

Figure 15.1. An application of C-MAB: Personalization of advertisements in recommender systems.

2. Basic MAB is a special case of C-MAB where $|\mathscr{S}| = 1$. Hence in this case the concept of competing against the best policy $g : \mathscr{S} \mapsto [N]$ simply boils down to competing against the best action in $[N]$.

3. In C-MAB problems, it is assumed that there exists some "good" mapping $g^* : \mathscr{S} \mapsto [N]$ in the hindsight. The goal of the learner is to exploit the contextual informations while selecting the actions.

4. Another popular variant of adversarial C-MAB is stochastic C-MAB where at each round t, the loss corresponding to each action is chosen probabilistically.

15.2.1 Baseline Approach: S-EXP3

A simple and naive approach to deal with C-MAB problems is to simultaneously run $|\mathscr{S}|$ independent instances of Exp3 algorithm for each distinct context $s \in \mathscr{S}$, and at each round t, predict as per the EXP3 instance corresponding to the current context s_t.

Theorem 15.1. *For any set of contexts $\mathscr{S}$,*

$$R_T(\textbf{\textit{S-EXP3}}) \leq \sqrt{2TN|\mathscr{S}|\log N}$$

Proof: Let $T_s = \sum_{t=1}^{T} \mathbf{1}(s_t = s)$ denotes the number of times context s is encountered in T rounds. Then, for any sequence of contextual informations $(s_1, s_2, \cdots, s_T)$,

$$\begin{aligned}
\max_{g:\mathscr{S}\mapsto[N]} \mathbb{E}\left[\sum_{t=1}^{T} l(I_t,t) - \sum_{t=1}^{T} l\big(g(s_t),t\big)\right] &= \max_{g:\mathscr{S}\mapsto[N]} \mathbb{E}\left[\sum_{s\in\mathscr{S}}\left(\sum_{t=1}^{T}\big(l(I_t,t) - l(g(s_t),t)\big)\mathbf{1}(s_t=s)\right)\right] \\
&= \max_{g:\mathscr{S}\mapsto[N]} \mathbb{E}\left[\sum_{s\in\mathscr{S}}\left(\sum_{t:s_t=s}\big(l(I_t,t) - l(g(s),t)\big)\right)\right] \\
&= \sum_{s\in\mathscr{S}} \max_{i\in[N]} \mathbb{E}\left[\left(\sum_{t:s_t=s}\big(l(I_t,t) - l(i,t)\big)\right)\right] \\
&\leq \sum_{s\in\mathscr{S}} \sqrt{2T_sN\log N} \quad \text{[By regret bound of EXP3 algorithm]} \\
&= \sqrt{2N\log N}\sum_{s\in\mathscr{S}}\sqrt{T_s} \\
&\leq \sqrt{2TN|\mathscr{S}|\log N} \quad \text{[By Cauchy Schwartz Inequality]}. \quad \square
\end{aligned}$$

Remarks:

1. Similar to original Exponential Weights algorithm, if the time horizon T is not known in advance, the optimal learning rate η for each instance of EXP3 algorithm can be chosen by so called doubling trick. Another popular approach for this purpose is to use time varying learning rate $\eta_t = \frac{1}{\sqrt{t}}$, $\forall t = 1,2,\ldots$, which is independent of the time horizon T.

15.3 A variant of C-MAB: Bandits with Expert Advice

In this setting, it is assumed that there exists a finite set of M randomized policies (each of them predicts over the set of N actions) which are treated as M experts. Based on the contextual information revealed at round $t = 1,2,\cdots$, each expert $j \in [M]$ makes a randomized prediction $\pi_{j,t} \in \Delta_N$ which in turn used by the learner for selecting an action $I_t \in [N]$. This problem can be viewed as a variant of C-MAB, although the learner's decision only indirectly depends on the contextual information through the prediction made by each expert. Similar to basic expert advice setting here also the objective of the learner is to minimize the (expected) regret w.r.t the best expert. A more formal description of problem setting is given below:

Bandits with Expert Advice

Inputs:

Set of randomized policies or experts' advice: $\{\pi_{j,t} \in \Delta_N : j \in [M], t = 1,2,..\}$

Set of actions or decisions: $[N]$

For each round $t = 1,2,\ldots$

– Learner receives the experts' advice $\{\pi_{1,t}, \pi_{2,t}, \cdots, \pi_{M,t}\}$

– Learner selects an action $I_t \in [N]$

– Learner suffers loss $l(I_t,t)$

End

The expected regret of the learner at the end of round T is defined as:

$$R_T := \max_{m\in[M]} \mathbb{E}\left[\sum_{t=1}^{T} l(I_t,t) - \sum_{t=1}^{T} \mathbb{E}_{i\sim\pi_{m,t}}[l(i,t)]\right]$$

Remarks:

1. Note that, $\mathbb{E}_{i\sim\pi_{m,t}}[l(i,t)] = \sum_{i=1}^{N} \pi_{m,t}(i)l(i,t) = <\pi_{m,t}, l_t>$ denotes the expected loss incurred by expert m at round t, where $l_t = (l(1,t), l(2,t), \cdots, l(N,t))$.
2. It is assumed that each expert's advice depends on the learner's past history, i.e. $\pi_{m,t} \in \sigma\big(I_1, I_2, \cdots, I_{t-1}, l(I_1,1), l(I_2,2), \cdots, l(I_{t-1}, t-1)\big)$.

15.3.1 Baseline Approach: EXP3 over Experts

A simple approach is to treat each of the M experts as one action (or bandit arm) and run EXP3 algorithm over these higher order bandits. More formally, at each round t, the algorithm first draws one action (expert) $I_t^{'} \in [M]$ from a distribution $q_t \in \Delta_M$ which is maintained by running EXP3 algorithm over the set of M experts, and predicts $I_t \in [N]$ according to the prediction of the selected expert $I_t^{'}$, i.e $I_t \sim \pi_{I_t^{'},t}$. The detailed description of the algorithm is given below:

Algorithm: **EXP3 over Experts**

Parameter:

$\eta > 0$

Initialize:

Probability distribution over M experts $q_1 \sim \text{Unif}(0,1)$, i.e. $q_{m,1} = \frac{1}{M}, \forall m \in [M]$

Cumulative loss for each expert $\widetilde{Y}_0 = \mathbf{0}$, i.e. $\widetilde{Y}_{m,0} = 0, \forall m \in [M]$

For each round $t = 1, 2, \ldots$

– Receive the experts' advice $\{\pi_{t,1}, \pi_{2,t}, \cdots, \pi_{M,t}\}$

– Select an expert $I_t^{'} \sim q_t$

– Predict an action $I_t \sim \pi_{I_t^{'},t}$

– Incur loss $l(I_t, t)$

– Compute estimated loss of M experts:

$$\widetilde{y}(j,t) = \frac{\mathbb{E}_{i\sim\pi_{j,t}}[l(i,t)]}{q_{j,t}}\mathbf{1}(j = I_t^{'}), \forall j \in [M]$$

– Update estimated cumulative loss of M experts:

$$\widetilde{Y}_{j,t} = \widetilde{Y}_{j,t-1} + \widetilde{y}(j,t), \forall j \in [M]$$

– Update probability distribution over M experts:

$$q_{j,t+1} = \frac{\exp(-\eta\widetilde{Y}_{j,t})}{\sum_{m=1}^{M}\exp(-\eta\widetilde{Y}_{m,t})}, \forall j \in [M]$$

End

Theorem 15.2. *With* $\eta = \sqrt{\frac{2\log M}{TM}}$,

$$R_T(\textit{EXP3 over Experts}) \leq \sqrt{2TM\log M}$$

Proof of the above theorem follows directly from the regret bound of EXP3 algorithm as we have seen in Lecture 12. However, this algorithm have two major drawbacks:

1. Due to the $O(\sqrt{M\log M})$ dependency of M in the regret guarantee, if M is really large, the regret bound of the above algorithm becomes trivial.

2. The algorithm does not make use of the structure of N actions, and it is independent of size of the action set N. Indeed it can be shown that making a subtle modification to this algorithm one can achieve a regret bound of $O(\sqrt{TN\log M})$ which is much more competitive when $M >> N$. This modified algorithm is known as EXP4, which we will study next.

15.3.2 EXP4

Similar to the above algorithm, EXP4 also maintains a probability distribution $q_t \in \Delta_M$ over the set of M experts, but while the above algorithm predicts an action $I_t \in [N]$ according to the prediction of the selected expert $\pi_{I_t^{'},t} \in \Delta_N$, EXP4 first mixes the prediction of each expert $\pi_{m,t}$ with q_t, and

then predicts I_t according to the resulting composite distribution $p_t \in \Delta_N$. Moreover, the cumulative loss $\widetilde{Y}_{m,t}$ of each expert m is now estimated using p_t instead of $q_{I_t',t}$. A more formal description of EXP4 algorithm is given below:

Algorithm: **EXP4**

Parameter:

$\eta > 0$

Initialize:

Probability distribution over M experts $q_1 \sim \text{Unif}(0,1)$, i.e. $q_{m,1} = \frac{1}{M}, \forall m \in [M]$

Cumulative loss for each expert $\widetilde{Y}_0 = \mathbf{0}$, i.e. $\widetilde{Y}_{m,0} = 0, \forall m \in [M]$

For each round $t = 1, 2, \ldots$

– Receive the experts' advice $\{\pi_{t,1}, \pi_{2,t}, \cdots, \pi_{M,t}\}$

– Compute probability distribution over N actions $p_t \equiv (p_{1,t}, p_{2,t}, \cdots, p_{N,t}) \in \Delta_N$, s.t.

$$p_{i,t} = \sum_{m=1}^{M} q_{m,t} \pi_{t,m}(i), \forall i \in [N]$$

– Predict an action $I_t \sim p_t$

– Incur loss $l(I_t, t)$

– Compute estimated loss of N actions:

$$\widetilde{l}(i,t) := \frac{l(i,t)}{p_{i,t}} \mathbf{1}(i = I_t), \forall i \in [N]$$

– Compute estimated loss of M experts:

$$\widetilde{y}(j,t) = \mathbb{E}_{i \sim \pi_{j,t}}[\widetilde{l}(i,t)], \forall j \in [M]$$

– Update estimated cumulative loss of M experts:

$$\widetilde{Y}_{j,t} = \widetilde{Y}_{j,t-1} + \widetilde{y}(j,t), \forall j \in [M]$$

– Update probability distribution over M experts:

$$q_{j,t+1} = \frac{\exp(-\eta \widetilde{Y}_{j,t})}{\sum_{m=1}^{M} \exp(-\eta \widetilde{Y}_{m,t})}, \forall j \in [M]$$

End

Theorem 15.3. *With* $\eta = \sqrt{\frac{2\log M}{TN}}$,

$$R_T(EXP4) \leq \sqrt{2TN\log M}$$

15.4 Next Lecture

In the next lecture, we will study the proof of Theorem 15.3, and also start with the problem of stochastic multi-armed bandits.

References

1. Chapter 2, Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Bubeck, S., and Cesa-Bianchi, N. *Foundations and Trends in Machine Learning*, 2012.

2. The nonstochastic multiarmed bandit problem. Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. *SIAM Journal on Computing*, 32:4877, 2002.