

## Lecture 8: February 26, 2024

Lecturer: Yishay Mansour

Scribes: Eyal Orgad, Yair Vaknin<sup>1</sup>

# 1 Contextual Bandits Model

## 1.1 Motivation

Until now, the action we chose was the universally best option. However, suppose we are building a system to provide recommendations to users, perhaps to suggest products to buy or to show them relevant advertisements. The recommendation should depend on the ‘context’ of the user’s profile - information we know about the given user (e.g. gender, location, etc.). We would like to use this contextual information to personalize the recommendations.

Another example of this could be choosing a route for a drive. Even if their origins and destinations are identical, two different users might want to take a different route. In different times, they might also want to take different routes - The context could be things other than the user, such as different weather or time of week or day.

This lecture will focus on contexts, where the context will influence the correct action to be taken.

## 1.2 Model

At time  $t \in [T]$ , a contextual bandit model observes **context**  $x_t \in X$ . The algorithm then chooses action  $a_t \in A$  and receives loss  $\ell_t(x_t, a_t) \in [0, 1]$ . When defining the model, we need to take note of a few things. First, how do we choose the context? Second, how do we set the loss?

Although the loss will always be a function of both the context ( $x_t$ ) and the chosen action ( $a_t$ ), we will usually write  $\ell_t(a_t)$  for the loss, omitting the context.

The set of possible contexts  $X$  might be finite (and even small), or it could be infinite, as we will see. In this setting, we shall take a look at two models - the stochastic model and the adversarial model, as described below.

### 1.2.1 Stochastic Model

In this model, the contexts are sampled from some distribution  $D$ , i.e, the contexts  $x_t \sim D$  are i.i.d. Given  $x_t$  and  $a_t$ , there is also a distribution over 1-sub-Gaussian loss functions which yields  $\ell_t(x_t, a_t)$ .

Recall that a random variable  $X$  is 1-sub-Gaussian if  $\forall \lambda \in \mathbb{R} : \mathbb{E} [e^{\lambda X}] \leq e^{\frac{\lambda^2}{2}}$ . The reason for the sub-Gaussian assumption on the loss functions here will allow the use of a Chernoff bound.

### 1.2.2 Adversarial Model

In this model, the contexts  $x_t$  are chosen by an adversary. It also selects a loss functions  $\ell_t(x_t, \cdot)$  at each time  $t$ , which is defined for all actions  $a \in A$ . Note that the loss function is selected independently of the chosen action  $a_t$ .

## 1.3 Regret Definition

The model’s online loss is given by  $\mathbb{E} \left[ \sum_{t=1}^T \ell_t(a_t) \right]$ .

To define regret, we must pick a benchmark value to compare this to. Previously we have seen such a benchmark of the form  $\min_a \mathbb{E} \left[ \sum_{t=1}^T \ell_t(a) \right]$ . However, this is too weak for a contextual bandit model,

<sup>1</sup>Based on the 2021/22 scribe notes by Morris Alper and Naama Yochai.

because it treats action  $a$  as independent from context  $x_t$ . We would instead like to measure how well we are able to pick actions in context relative to an excellent choice of an action in context.

In order to do this, we define a **policy class**  $\Pi = \{\pi: X \rightarrow A\}$  whose elements are policies that map contexts to actions.

For example, in the basic MAB model actions are chosen without respect to any context, so the policy class is  $\Pi = \{\pi_a : a \in A\}$  where  $\pi_a(x) = a \quad \forall x \in X$  - Every action has a policy that always chooses said action (Since context is irrelevant in basic MAB, we can model it as a policy that maps context to action, but the mapping is trivial).

### 1.3.1 Stochastic Model Regret

In the stochastic model,

$$\text{Regret} = \mathbb{E} \left[ \sum_{t=1}^T \ell_t(a_t) \right] - \min_{\pi \in \Pi} \mathbb{E} \left[ \sum_{t=1}^T \ell_t(\pi(x_t)) \right]$$

In other words, the regret measures our loss relative to the expected loss of the **best possible policy**, where the expectation is relative to distributions over contexts and loss functions.

### 1.3.2 Adversarial Model Regret

In the adversarial model,

$$\text{Regret} = \mathbb{E} \left[ \sum_{t=1}^T \ell_t(a_t) \right] - \min_{\pi \in \Pi} \sum_{t=1}^T \ell_t(\pi(x_t))$$

This is similar to the regret in the stochastic case, but since contexts and loss functions are selected by the adversary at each time step  $t$  there is no expected value applied to the benchmark term. Instead, our observed loss is compared to the hypothetical loss of the best-in-hindsight policy. We assume that the series of contexts are the exact same for both terms.

## 1.4 Motivation

Note that in contextual bandit models the set of contexts  $X$  and/or the set of policies  $\Pi$  may be very large. As we can see in the above expressions for regret, large  $|X|$  and/or  $|\Pi|$  will make the contextual bandit problem more challenging. We would like to minimize the dependence on these magnitudes.

## 2 Small Number of Contexts

### 2.1 Algorithm

In this case, suppose that  $|X|$  is small.

This case is “easy” - for every context  $x \in X$  we maintain a separate algorithm  $ALG_x$ :

---

**Algorithm 1:** Small Number of Contexts

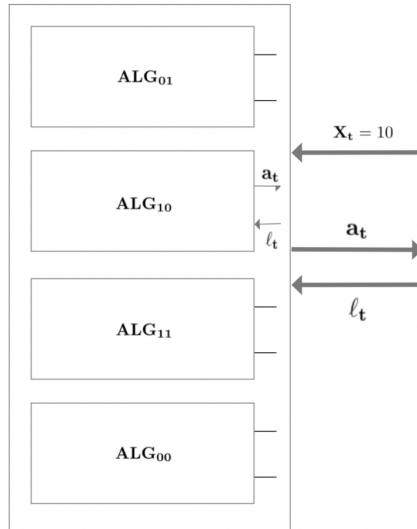
---

```

for each time  $t$  do
  Receive context  $x_t \in X$ 
  Run algorithm  $ALG_{x_t}$  to pick action  $a_t$ 
  Use action  $a_t$  and receive loss  $\ell_t(a_t)$ 
  Return  $\ell_t(a_t)$  to  $ALG_{x_t}$ 
end

```

---

Figure 1: Small Number of Contexts illustrated for  $X = \{00, 01, 10, 11\}$ 

## 2.2 Analysis

For each  $x \in X$ , define  $T_x = \{t : x_t = x\}$  the time steps where context  $x$  is seen. Note that  $\sum_{x \in X} T_x = T$ .

We may assume that the algorithm  $ALG$  (operated on each context separately) has regret bound  $Regret = R(T, K) = O(\sqrt{KT \log K})$  - for example, we know this bound holds if the EXP3 algorithm is used.

Then we may calculate the total regret over all contexts:

$$\begin{aligned}
 Regret &= \sum_{x \in X} Regret(ALG_x) \\
 &\leq \sum_{x \in X} R(|T_x|, K) \\
 &= \sum_{x \in X} O(\sqrt{K |T_x| \log K}) \\
 &= O(\sqrt{|X| K |T| \log K})
 \end{aligned}$$

If the number of contexts is small, the regret bound doesn't change. However, if the number of contexts is very large (let's say larger than  $T$ ), then this bound becomes irrelevant.

Notice that the last equality follows from the fact that  $\sum_{x \in X} \sqrt{|T_x|} \leq \sqrt{|X| T}$ , which is a corollary of the following general inequality with  $\sqrt{|T_x|}$  in place of  $x_i$ :

**Lemma 1**

$$\sum_{i=1}^n x_i \leq \sqrt{n \sum_i x_i^2}, \quad x_1, \dots, x_n \in \mathbb{R}$$

**Proof:** Let  $X$  be a random variable which equals  $x_i$  with probability  $\frac{1}{n}$ . Then

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0 \\
&\Rightarrow (\mathbb{E}[X])^2 \leq \mathbb{E}[X^2] \\
&\Rightarrow \left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2 \leq \frac{1}{n} \sum_{i=1}^n x_i^2 \\
&\Rightarrow \left(\sum_{i=1}^n x_i\right)^2 \leq n \sum_{i=1}^n x_i^2 \\
&\Rightarrow \sum_{i=1}^n x_i \leq \sqrt{n \sum_{i=1}^n x_i^2}
\end{aligned}$$

□

### 3 EXP4 Algorithm for Contextual Bandits

#### 3.1 EXP4 Algorithm

Given that we have:

- a set  $A$  of  $k$  actions
- a set  $M$  of  $m$  experts s.t. for every round  $t$  each expert returns  $q_{t,i} \in \Delta(A)$

Given that:

$$\text{Regret} = E\left(\sum_{t=1}^T \ell_t(a_t)\right) - \min_{i \in M} \sum_{t=1}^T \ell_t \cdot q_{t,i}$$

Our goal is to achieve a regret bound of  $O(\sqrt{Tk \log m})$ . The naive solution is to pick  $M = A$ , and then we get  $R = \sqrt{Tm \log m}$  (usually  $m \gg k$ ).

Let  $S = \min(m, k)$ , the goal is to get  $\text{Regret} = O(\sqrt{ST \log m})$ .

This will allow us to have an exponential number of experts.

The idea for the algorithm is to maintain a weight  $w_t(i)$  for each expert  $i$ . At time  $t$ , we define the weight of expert  $i$  to be:  $P_t(i) = \frac{w_t(i)}{W_t}$  where  $W_t = \sum_j w_t(j)$ . We will then choose an action  $a \in A$  using the distribution  $Q_t$  which is the average of  $q_{t,i}$  according to  $P_t(i)$ . This will let us update many experts from one action.

---

#### Algorithm 2: EXP4

---

Initialization  $w_q(i) = 1; \forall 1 \leq i \leq m$

**for each time  $t$  do**

    Expert  $i \in M$  gives  $q_{t,i} \in \Delta(A)$  after seeing the context

$P_t(i) = \frac{w_t(i)}{W_t}; W_t = \sum_{i \in M} w_t(i)$

$\forall a: Q_t(a) = \sum_{i \in M} P_t(i) q_{t,i}(a)$

    Sample action  $a_t$  from the distribution  $Q_t$

    Define  $\hat{\ell}_t(a) = \frac{\ell_t(a)}{Q_t(a)} \mathbb{I}\{a = a_t\}$

**for  $i=1, \dots, m$  do do**

        Compute  $h_t(i) = \hat{\ell}_t \cdot q_{t,i} = \frac{\ell_t(a_t)}{Q_t(a_t)} q_{t,i}(a_t)$

        Update weights  $w_{t+1}(i) = w_t(i) e^{-\eta h_t(i)}$

**end**

**end**

---

### 3.2 Analysis

We show that  $h_t(i)$  is an unbiased estimate of the loss of expert  $i$ :

$$E[h_t(i) | Q_t] = E[\hat{\ell}_t \cdot q_{t,i}] = \sum_{a \in A} Q_t(a) \frac{q_{t,i}(a)\ell_t(a)}{Q_t(a)} = q_{t,i} \cdot \ell_t = \text{loss}_t(i)$$

### 3.3 Regret Bound

From the analysis of EXP3 (for  $h_t$ ):

$$E\left[\sum_{t=1}^T P_t \cdot h_t\right] - \sum_{t=1}^T h_t(i) \leq \frac{1}{\eta} \log m + \underbrace{\frac{\eta}{2} \sum_{t=1}^T \sum_{i=1}^m E[P_t(i)h_t^2(i)]}_{\text{Need to bound}}$$

We will bound  $\sum_{i=1}^m E[P_t(i)h_t^2(i)]$  :

$$E[h_t^2(i) | P_t] = \sum_{a \in A} Q_t(a) \left(\frac{q_{t,i}(a)\ell_t(a)}{Q_t(a)}\right)^2 \leq \sum_{a \in A} \frac{q_{t,i}^2(a)}{Q_t(a)} \quad (\ell_t^2(a) \leq 1)$$

Summing all the experts we get:

$$\sum_{i=1}^m E[P_t(i)h_t^2(i)] = \sum_{i=1}^m E[P_t(i)E[h_t^2(i) | P_t]] \leq E\left[\sum_{a \in A} \frac{\sum_i P_t(i)q_{t,i}^2(a)}{Q_t(a)}\right]$$

Define:

$$S_t(a) = \max_i q_{t,i}(a); S_t = \sum_a S_t(a); S = \max_t S_t$$

We bound the second order term,

$$\sum_{i=1}^m E[P_t(i)h_t^2(i)] \leq E\left[\sum_{a \in A} \frac{\sum_i P_t(i)q_{t,i}(a)}{Q_t(a)} S_t(a)\right] = \sum_{a \in A} S_t(a) \leq S$$

Now finally bound the regret:

$$\text{Regret} = E\left[\sum_{t=1}^T P_t \cdot h_t\right] - \min_i \sum_{t=1}^T h_t(i) \leq \frac{1}{\eta} \log m + \frac{\eta}{2} TS = O(\sqrt{TS \log m})$$

For  $\eta = \sqrt{\frac{\log m}{TS}}$ .

### 3.4 Summary

To give a bound depending on S we will consider a few cases:

1. If all experts always agree ( $q_{t,i} = q_{t,j}$ ):

$$\text{Then } S_t = \sum_a q_{t,i}(a) = 1$$

The bound will be  $O(\sqrt{T \log m})$

2. If there are a lot of different minded experts:

$$S_t(a) \leq 1 \Rightarrow S_t \leq K \Rightarrow S \leq K$$

The bound will be  $O(\sqrt{TK \log m})$

3. If we have many actions, but only a few experts:

$$S_t(a) = \max_i q_{t,i}(a) \leq \sum_{i=1}^m q_{t,i}(a)$$

This means that:

$$S_t = \sum_a S_t(a) \leq \sum_{i=1}^m \sum_a q_{t,i}(a) = m$$

Thus the bound is:  $O(\sqrt{Tm \log m})$

Advantage: Optimal regret bound.

Disadvantage: Space and runtime bound  $O(m)$

## 4 Stochastic Contextual Bandits

Let  $\Pi$  be a class of policies:

$$\Pi = \{\pi \in \Pi \mid \pi : X \rightarrow A\}$$

We will show a greedy algorithm in the context of the full information model, s.t. for every time step  $t$  we observe all losses. The idea is: At each step, compute a policy  $\pi_t$  that achieves minimal loss relative to all actions. Then, use this policy to choose the next action. This is okay since there's no reason to explore due to the full information setting, and we can also be greedy, since the model is stochastic and not adversarial.

---

**Algorithm 3:** Greedy Stochastic Contextual Bandit (full information)

---

```

for each time  $t = 1, 2, 3, \dots, T$  do
  Compute policy with minimal loss so far:  $\pi_t = \operatorname{argmin}_{\pi \in \Pi} \sum_{i=1}^{t-1} \ell_t(\pi(x_i))$ ;
  Perform action:  $a_t = \pi_t(x_t)$ ;
  For every  $a \in A$  observe  $\ell_t(a)$ ;
end

```

---

**Theorem 2** *The regret bound of the algorithm:*

$$\sum_{t=1}^T \mathbb{E} [\ell_t(\pi^*(x_t))] \leq \sum_{t=1}^T \mathbb{E} [\ell_t(\pi_t(x_t))] + O(\sqrt{T \log(|\Pi| \cdot T)})$$

**Proof:** For every  $\pi$  and time  $t$ , with probability greater than  $1 - \delta$ :

$$\left| \frac{1}{t-1} \sum_{i=1}^{t-1} \ell_i(\pi(x_i)) - \mathbb{E} [\ell_i(\pi(x_i))] \right| \leq \sqrt{\frac{1}{2t} \log \left( \frac{2|\Pi|T}{\delta} \right)}$$

Therefore, for every time  $t$ , with probability greater than  $1 - \delta$ :

$$|\mathbb{E} [\ell_t(\pi_t(x_t))] - \mathbb{E} [\ell_t(\pi^*(x_t))]| \leq 2 \cdot \sqrt{\frac{1}{2t} \log \left( \frac{2|\Pi|T}{\delta} \right)} = \sqrt{\frac{2}{t} \log \left( \frac{2|\Pi|T}{\delta} \right)}$$

Now we can bound the regret:

$$\text{Regret} \leq \sum_{t=1}^T \sqrt{\frac{2}{t} \log \left( \frac{2|\Pi|T}{\delta} \right)} = \sqrt{2 \log \left( \frac{2|\Pi|T}{\delta} \right)} \underbrace{\sum_{t=1}^T \frac{1}{\sqrt{t}}}_{O(\sqrt{T})}$$

Setting  $\delta = \frac{1}{T}$  yields:

$$\text{Regret} = O(\sqrt{T \log(|\Pi|T)})$$

□

## 4.1 Explore Exploit

We will now look at a setting where we only have bandit feedback and not full information. That is, exploration will be needed now.

---

**Algorithm 4:** Contextual Explore Exploit

---

```

for each time  $t = 1, 2, 3, \dots, B$  do
  | Given  $x_t$ , choose random action  $a_t$ 
  | Perform action  $a_t$ 
  | Receive loss  $\ell_t(a)$ 
  | Save  $(x_t, a_t, \ell_t(a_t))$ 
end
Compute policy with minimal loss so far:  $\pi_B = \operatorname{argmin}_{\pi \in \Pi} \frac{1}{B} \sum_{i=1}^B \ell_B(\pi(x_i)) \mathbb{I}\{\pi(x_i) = a_i\}$ 
for each time  $t = B + 1, \dots, T$  do
  | Given  $x_t$ , compute  $a_t = \pi_B(x_t)$ 
  | Perform action:  $a_t$ 
  | Receive loss  $\ell_t(a_t)$ 
end

```

---

## 4.2 Analysis

Define:

$$L(\pi) = \mathbb{E}_x[\ell(\pi(x), x)]$$

For every policy  $\pi \in \Pi$ , in the explore part we have:

$$\mathbb{E}[\ell_t(a_t) \mathbb{I}\{a_t = \pi(x_t)\}] = \mathbb{E}_{x_t, a_t}[\mathbb{E}[\ell_t(a_t) \mathbb{I}\{a_t = \pi(x_t)\} \mid x_t]] = \frac{1}{k} \mathbb{E}_{x_t}[\mathbb{E}[\ell_t(a_t) \mid x_t]] = \frac{1}{k} L(\pi)$$

Thus:

$$\mathbb{E}[\hat{L}(\pi)] = \mathbb{E}\left[\frac{1}{B} \sum_{t=1}^B \ell_t(a_t) \mathbb{I}\{a_t = \pi(x_t)\}\right] = \frac{1}{k} L(\pi)$$

And so we have an unbiased estimation for  $L(\pi) = k\hat{L}(\pi)$ .

Using Hoeffding's inequality, w.p. greater than  $1 - \delta$ :

$$|\hat{L}(\pi) - \underbrace{\mathbb{E}[\hat{L}(\pi)]}_{L(\pi)/k}| \leq \sqrt{\frac{1}{2B} \log\left(\frac{2|\Pi|}{\delta}\right)}$$

Therefore w.p.  $1 - \delta$  we have:

$$L(\pi_B) \leq L(\pi^*) + 2k \cdot \sqrt{\frac{1}{2B} \log\left(\frac{2|\Pi|}{\delta}\right)}$$

So the regret bound is:

$$\mathbb{E}[\text{Regret}] = B + (T - B) (\mathbb{E}[L(\pi_B) - L(\pi^*)]) \leq B + 2Tk \sqrt{\frac{2}{B} \log\left(\frac{2|\Pi|}{\delta}\right)} + \delta T$$

Using  $\delta = \frac{1}{T}$  and  $B = (Tk)^{\frac{2}{3}} \left(\log \frac{|\Pi|}{\delta}\right)^{\frac{1}{3}}$  we get:

$$\mathbb{E}[\text{Regret}] = O\left(T^{\frac{2}{3}} k^{\frac{2}{3}} \log^{\frac{1}{3}}(T|\Pi|)\right)$$

## 5 Adversarial Contextual Bandits

In this case, we will see a use of the **Follow the Perturbed Leader** algorithm.

### 5.1 Model

The model consists of

- Actions  $a \in \{0, 1\}^k$  ( $k$  bits each,  $2^k$  possible actions)
- A set of contexts  $Z \subset X$  which distinguishes between elements of policy class  $\Pi$ , i.e.:

$$\forall \pi_1, \pi_2 \in \Pi \exists z \in Z \text{ s.t. } \pi_1(z) \neq \pi_2(z)$$

- Optimizer oracle  $M((\ell_\tau, x_\tau)_{\tau=1}^{t-1})$  which returns an optimal policy given the history up to time  $t-1$ .

### 5.2 Contextual FTPL Algorithm

The Contextual-FTPL( $Z, \eta$ ) algorithm is as follows:

---

**Algorithm 5:** Contextual-FTPL( $Z, \eta$ )

---

Initialization:

**for** each  $z \in Z$  **do**

    | Sample random vector  $\ell_z \in [-\frac{1}{\eta}, \frac{1}{\eta}]^k$  uniformly

**end**

Set  $S := \{(z, \ell_z) : z \in Z\}$

**for** each time  $t$  **do**

    | Select policy  $\pi_t = M(S \cup (\ell_\tau, x_\tau)_{\tau=1}^{t-1})$

    | Perform action  $a_t = \pi_t(x_t)$

    | Receive loss  $\ell_t(\cdot)$

**end**

---

Note: Losses are defined as linear functions  $\ell_z(a) = a \cdot \ell_z$ .

### 5.3 Regret Bound

**Theorem 3** *We can bound the regret of the algorithm using the stability and error bounds:*

$$\text{Regret} \leq \text{Stability} + \text{Error}$$

$$\begin{aligned} \text{Stability} &= \sum_{t=1}^T \mathbb{E}_z [\ell_t(\pi_t(x_t))] - \mathbb{E}_z [\ell_t(\pi_{t+1}(x_t))] \\ \text{Error} &= \mathbb{E} \left[ \max_{\pi \in \Pi} \sum_{\zeta \in Z} \ell_\zeta(\pi(\zeta)) \right] - \mathbb{E} \left[ \min_{\pi \in \Pi} \sum_{\zeta \in Z} \ell_\zeta(\pi(\zeta)) \right] \end{aligned}$$

**Proof:** Considering the losses of all time steps, and marking  $\pi^*$  as the optimal policy:

$$\begin{aligned} \text{Regret} &= \sum_{t=1}^T \ell_t(\pi_t(x_t)) - \ell_t(\pi^*(x_t)) = \underbrace{\sum_{t=1}^T \ell_t(\pi_t(x_t)) - \ell_t(\pi_{t+1}(x_t))}_{\text{Stability}} + \underbrace{\sum_{t=1}^T \ell_t(\pi_{t+1}(x_t)) - \ell_t(\pi^*(x_t))}_{\text{Error}} \end{aligned}$$

To prove the theorem we will first prove the following lemma:

**Lemma 4**  $\sum_{t=1}^T \ell_t(\pi_{t+1}(x_t)) - \ell_t(\pi^*(x_t)) \leq \text{Error}$

**Proof:** We will show by induction on  $T$  that:

$$\sum_{\zeta \in Z} \ell_\zeta(\pi_1(\zeta)) + \sum_{t=1}^T \ell_t(\pi_{t+1}(x_t)) \leq \sum_{\zeta \in Z} \ell_\zeta(\pi^*(\zeta)) + \sum_{t=1}^T \ell_t(\pi^*(x_t))$$

The induction step: Substituting  $\pi^*$  with  $\pi_{T+2}$  and adding  $\ell_{T+1}(\pi_{T+2}(x_{T+1}))$  implies:

$$\begin{aligned} \sum_{\zeta \in Z} \ell_{\zeta}(\pi_1(\zeta)) + \sum_{t=1}^{T+1} \ell_t(\pi_{t+1}(x_t)) &\leq \\ \sum_{\zeta \in Z} \ell_{\zeta}(\pi_{T+2}(\zeta)) + \sum_{t=1}^T \ell_t(\pi_{T+2}(x_t)) + \ell_{T+1}(\pi_{T+2}(x_{T+1})) &= \\ \sum_{\zeta \in Z} \ell_{\zeta}(\pi_{T+2}(\zeta)) + \sum_{t=1}^{T+1} \ell_t(\pi_{T+2}(x_t)) &= (*) \end{aligned}$$

By the definition of  $\pi_{T+2}$ , it holds that for all  $\pi^*$ :

$$(*) \leq \sum_{\zeta \in Z} \ell_{\zeta}(\pi^*(\zeta)) + \sum_{t=1}^{T+1} \ell_t(\pi^*(x_t))$$

This completes the induction claim. Now, by using the previous equations we get:

$$\sum_{\zeta \in Z} \ell_{\zeta}(\pi_1(\zeta)) + \sum_{t=1}^T \ell_{t+1}(\pi_{t+1}(x_t)) \leq \sum_{\zeta \in Z} \ell_{\zeta}(\pi^*(\zeta)) + \sum_{t=1}^T \ell_t(\pi^*(x_t))$$

Thus:

$$\begin{aligned} \sum_{t=1}^T \ell_{t+1}(\pi_{t+1}(x_t)) - \ell_t(\pi^*(x_t)) &\leq \\ \sum_{\zeta \in Z} \ell_{\zeta}(\pi^*(\zeta)) - \ell_{\zeta}(\pi_1(\zeta)) &\leq \\ \max_{\pi \in \Pi} \sum_{\zeta \in Z} \ell_{\zeta}(\pi(\zeta)) - \min_{\pi \in \Pi} \sum_{\zeta \in Z} \ell_{\zeta}(\pi(\zeta)) &= \text{Error} \end{aligned}$$

□

Using this lemma we can complete the proof of the theorem:

$$\begin{aligned} \text{Regret} &= \underbrace{\sum_{t=1}^T \ell_t(\pi_t(x_t)) - \ell_t(\pi_{t+1}(x_t))}_{\text{Stability}} + \underbrace{\sum_{t=1}^T \ell_t(\pi_{t+1}(x_t)) - \ell_t(\pi^*(x_t))}_{\leq \text{Error}} \\ &\leq \text{Stability} + \text{Error} \end{aligned}$$

□

### 5.3.1 Error Bound of Contextual-FTPL

**Lemma 5**

$$\text{Error} = O\left(\frac{1}{\eta} \sqrt{|Z|k \log(|Z||\Pi|)}\right)$$

**Proof:**

$$\text{For any policy } \pi \in \Pi: \mathbb{E} \left[ \sum_{\zeta \in Z} \ell_{\zeta}(\pi(\zeta)) \right] = 0$$

W.p.  $\geq 1 - \delta$ :

$$\left| \sum_{\zeta \in Z} \ell_{\zeta}(\pi(\zeta)) - \mathbb{E} \left[ \sum_{\zeta \in Z} \ell_{\zeta}(\pi(\zeta)) \right] \right| = \left| \sum_{\zeta \in Z} \ell_{\zeta}(\pi(\zeta)) - 0 \right| = O\left(\frac{1}{\eta} \sqrt{|Z|k \log(|Z||\Pi|)}\right)$$

We choose  $\delta = \frac{1}{|Z||\Pi|}$  and get:

$$\text{Error} = O\left(\frac{1}{\eta} \sqrt{|Z|k \log(|Z||\Pi|)}\right)$$

□

### 5.3.2 Stability Bound of Contextual-F'TPL

**Lemma 6**

$$\mathbb{E}_Z [\ell_t(\pi_t(x_t)) - \ell_t(\pi_{t+1}(x_t))] \leq \eta k |Z|$$

**Proof:** Since:

$$\mathbb{E}_Z [\ell_t(\pi_t(x_t)) - \ell_t(\pi_{t+1}(x_t))] \leq P[\pi_t(x_t) \neq \pi_{t+1}(x_t)] \leq P[\pi_t \neq \pi_{t+1}]$$

If  $\pi_t \neq \pi_{t+1}$ , then there exists  $\zeta \in Z$  s.t  $\pi_t(\zeta) \neq \pi_{t+1}(\zeta)$ . We denote action  $\pi(x) \in A$  as a vector  $\pi(x) \in \{0, 1\}^k$  and as a set  $\{j : \pi(x)[j] = 1\}$ . Thus, from the equation above:

$$P[\pi_t \neq \pi_{t+1}] \leq \sum_{\zeta \in Z} P[\pi_t(\zeta) \neq \pi_{t+1}(\zeta)] \leq \sum_{\zeta \in Z} \sum_{j \in [k]} P[j \in \pi_t(\zeta), j \notin \pi_{t+1}(\zeta)] + P[j \notin \pi_t(\zeta), j \in \pi_{t+1}(\zeta)]$$

We want to bound  $P[j \in \pi_t(\zeta), j \notin \pi_{t+1}(\zeta)]$  and  $P[j \notin \pi_t(\zeta), j \in \pi_{t+1}(\zeta)]$ . Given  $\zeta, j$ , we will define a vector  $\ell_{\zeta, j}$  s.t:

$$\ell_{\zeta, j} = (0, \dots, 0, \ell_{\zeta}(j), 0, \dots, 0)$$

Define  $\Phi(\pi)$  equal to all the losses of  $\pi$  until  $t$ , and loses on all  $\zeta' \neq \zeta$  and the loss on  $\zeta$  while excluding the  $j$ 'th coordinate:

$$\Phi(\pi) = \sum_{\tau=1}^{t-1} \ell_{\tau}(\pi(x_{\tau})) + \sum_{\zeta' \neq \zeta} \pi(\zeta') \cdot \ell_{\zeta'} + \pi(\zeta) \cdot (\ell_{\zeta} - \ell_{\zeta, j})$$

Furthermore, define:

$$\pi^* = \arg \min_{\pi: j \in \pi(\zeta)} \Phi(\pi), \tilde{\pi} = \arg \min_{\pi: j \notin \pi(\zeta)} \Phi(\pi)$$

The event  $\{j \in \pi_t(\zeta)\}$  occurs only if:

$$\Phi(\pi^*) + \ell_{\zeta}(j) \leq \Phi(\tilde{\pi}) \quad \Rightarrow \quad \ell_{\zeta}(j) \leq \Phi(\tilde{\pi}) - \Phi(\pi^*) = v$$

If  $v - 1 > \ell_{\zeta}(j)$ , then  $\{j \in \pi_{t+1}(\zeta)\}$ . This implies that  $\ell_{\zeta}(j) \in [v - 1, v]$  and:

$$P[\ell_{\zeta}(j) \in [v - 1, v]] = \frac{\eta}{2}$$

Therefore,

$$P[j \in \pi_t(\zeta), j \notin \pi_{t+1}(\zeta)] \leq \frac{\eta}{2}$$

The second case is similar. Thus, we can complete the lemma:

$$\mathbb{E}_Z [\ell_t(\pi_t(x_t)) - \ell_t(\pi_{t+1}(x_t))] \leq P[\pi_t \neq \pi_{t+1}] \leq \eta k |Z|$$

□

## 5.4 General Regret Bound

We saw that the regret of this algorithm is bounded as

$$\text{Regret} \leq \underbrace{\text{Stability}}_{O(\eta k |Z| T)} + \underbrace{\text{Error}}_{O(\frac{1}{\eta} \sqrt{|Z| k \log |Z| |\Pi|})}$$

Optimizing over  $\eta$  we get  $\eta = \left(\frac{\log |Z| |\Pi|}{k |Z|}\right)^{\frac{1}{4}} \frac{1}{\sqrt{T}}$ . Plugging this into both terms gives the bound

$$\text{Regret} = O\left(\sqrt{T} K^{\frac{3}{4}} |Z|^{\frac{3}{4}} \log^{\frac{1}{4}}(|Z| |\Pi|)\right)$$

## References

- [1] Aleksandrs Slivkins. [Introduction to multi-armed bandits](#). *arXiv:1904.07272*, 2019.
- [2] Syrgkanis, Krishnamurthy & Schapire. [Efficient Algorithms for Adversarial Contextual Learning](#), *arXiv:1602.02454* 2016.