Lecture 22

Instructor: Haipeng Luo

1 Contextual Bandit with Adversarial Losses

We have seen oracle-efficient and optimal algorithms for the i.i.d. contextual bandit problems. Going beyond the i.i.d. assumption remains a challenging question and we will discuss some recent progress from [Rakhlin and Sridharan, 2016, Syrgkanis et al., 2016] in this lecture.

Specifically we study a hybrid setting where the contexts are i.i.d samples from a fixed but unknown distribution $\mathcal{D}$, while the losses can be adversarial. This could be a reasonable assumption for the personalized recommendation problem: users' contextual information such as gender might be relatively stationary (e.g. 40% men and 60% women for a shopping website), while the actual preferences for items might be changing more rapidly. We also assume that we can draw fresh examples from $\mathcal{D}$ as we want. This is a somewhat technical assumption but in some cases can still be very reasonable.

Recall that in Lecture 12 the very first attempt we tried for bandit was to reduce it to the expert problem in a blackbox way. Here we will take the same approach. Specifically, we first consider deriving oracle-efficient and optimal algorithms for the following full information problem: for each round $t = 1, \ldots, T$,

1. environment draws $x_t \sim \mathcal{D}$ and reveals x_t;
2. learner decides a distribution $p_t \in \Delta(K)$;
3. environment decides a loss vector $\widehat{\ell}_t \in \mathcal{L} \subset [0,1]^K$;
4. learner suffers loss $\left\langle p_t, \widehat{\ell}_t \right\rangle$ and observes $\widehat{\ell}_t$.

Here $\mathcal{L}$ is a special loss space to be discussed soon. The regret is defined as the difference between the loss of the algorithm and the best fixed policy from a policy class Π:

$$\mathcal{R}_T = \sum_{t=1}^{T} \left\langle p_t, \widehat{\ell}_t \right\rangle - \min_{\pi \in \Pi} \sum_{t=1}^{T} \widehat{\ell}_t(\pi(x_t)).$$

Now suppose we have an oracle-efficient algorithm to solve such a full information problem. Then for the hybrid contextual bandit problem, we can simply sample an action a_t according to $(1 - K\mu)p_t + \mu\mathbf{1}$, construct the usual importance-weighted estimator, and finally rescale it by μ and feed it to the full information algorithm as the loss vector $\widehat{\ell}_t$. It is then clear that the loss space $\mathcal{L}$ for this reduction has a special structure and can be defined as $\{ce_a : a \in [K], c \in [0,1]\}$ where $e_1, \ldots, e_K$ are the K-dimensional standard basis vectors.

Note that the optimal regret for this full information problem is $\mathcal{O}(\sqrt{T \ln N})$ (achieved by applying Hedge). Having an oracle-efficient algorithm for this problem with regret $\mathcal{O}(\sqrt{T \ln N})$ will then lead to expected regret $\mathcal{O}(T^{\frac{3}{4}} K^{\frac{1}{2}} (\ln N)^{\frac{1}{4}})$ [Rakhlin and Sridharan, 2016], by the exact same argument as in Lecture 12. This is clearly not the optimal regret. Recent work [Syrgkanis et al., 2016] improves the regret to $\mathcal{O}((TK)^{\frac{2}{3}} (\ln N)^{\frac{1}{3}})$, while getting the optimal regret with an oracle-efficient algorithm $\mathcal{O}(\sqrt{TK \ln N})$ is still open.

2 Relaxation-based Approach

We now discuss how to solve the full information problem. The approach is based on a minimax analysis. Specifically, first note that the optimal worst-case expected regret (with respect to the random contexts) can be written as a sequence of minimax expressions

$$\mathbb{E}_{x_1} \min_{p_1 \in \Delta(K)} \max_{\widehat{\ell}_1 \in \mathcal{L}} \cdots \mathbb{E}_{x_T} \min_{p_T \in \Delta(K)} \max_{\widehat{\ell}_T \in \mathcal{L}} \left(\sum_{\tau=1}^{T} \left\langle p_\tau, \widehat{\ell}_\tau \right\rangle - \min_{\pi \in \Pi} \sum_{\tau=1}^{T} \widehat{\ell}_\tau(\pi(x_\tau)) \right) \tag{1}$$

More generally, at the beginning of time t, having observed $x_{1:t-1}$ and $\widehat{\ell}_{1:t-1}$,[1] and assuming the player and the environment will both play optimally afterwards, the conditional expected regret is

$$\begin{aligned}
&\mathbb{E}_{x_t} \min_{p_t \in \Delta(K)} \max_{\widehat{\ell}_t \in \mathcal{L}} \cdots \mathbb{E}_{x_T} \min_{p_T \in \Delta(K)} \max_{\widehat{\ell}_T \in \mathcal{L}} \left(\sum_{\tau=1}^{T} \left\langle p_\tau, \widehat{\ell}_\tau \right\rangle - \min_{\pi \in \Pi} \sum_{\tau=1}^{T} \widehat{\ell}_\tau(\pi(x_\tau)) \right) \\
&= \sum_{\tau=1}^{t-1} \left\langle p_\tau, \widehat{\ell}_\tau \right\rangle + \underbrace{\mathbb{E}_{x_t} \min_{p_t \in \Delta(K)} \max_{\widehat{\ell}_t \in \mathcal{L}} \cdots \mathbb{E}_{x_T} \min_{p_T \in \Delta(K)} \max_{\widehat{\ell}_T \in \mathcal{L}} \left(\sum_{\tau=t}^{T} \left\langle p_\tau, \widehat{\ell}_\tau \right\rangle - \min_{\pi \in \Pi} \sum_{\tau=1}^{T} \widehat{\ell}_\tau(\pi(x_\tau)) \right)}_{\Phi(x_{1:t-1}, \widehat{\ell}_{1:t-1})}.
\end{aligned}$$

We denote the last term as $\Phi(x_{1:t-1}, \widehat{\ell}_{1:t-1})$, which can be seen as the optimal "regret" (in fact only the part of the regret that we can still control), starting from the state $(x_{1:t-1}, \widehat{\ell}_{1:t-1})$. Note that we have the following recursive relationship:

$$\Phi(x_{1:t-1}, \widehat{\ell}_{1:t-1}) = \mathbb{E}_{x_t} \min_{p_t \in \Delta(K)} \max_{\widehat{\ell}_t \in \mathcal{L}} \left(\left\langle p_t, \widehat{\ell}_t \right\rangle + \Phi(x_{1:t}, \widehat{\ell}_{1:t}) \right).$$

Also note that $\Phi(x_{1:T}, \widehat{\ell}_{1:T})$ is the negative benchmark $-\min_{\pi \in \Pi} \sum_{\tau=1}^{T} \widehat{\ell}_\tau(\pi(x_\tau))$ and $\Phi(\emptyset)$ is exactly the minimax regret Eq. (1). Moreover, using Φ we can derive the minimax optimal algorithm: pick p_t to be

$$\operatorname*{argmin}_{p \in \Delta(K)} \max_{\widehat{\ell}_t \in \mathcal{L}} \left(\left\langle p, \widehat{\ell}_t \right\rangle + \Phi(x_{1:t}, \widehat{\ell}_{1:t}) \right).$$

It is clear that the expected regret of this algorithm is bounded by $\Phi(\emptyset)$ since

$$\begin{aligned}
&\mathbb{E}_{x_t} \left[\left\langle p_t, \widehat{\ell}_t \right\rangle + \Phi(x_{1:t}, \widehat{\ell}_{1:t}) \right] \le \mathbb{E}_{x_t} \max_{\widehat{\ell}_t \in \mathcal{L}} \left(\left\langle p_t, \widehat{\ell}_t \right\rangle + \Phi(x_{1:t}, \widehat{\ell}_{1:t}) \right) \\
&= \mathbb{E}_{x_t} \min_{p_t \in \Delta(K)} \max_{\widehat{\ell}_t \in \mathcal{L}} \left(\left\langle p_t, \widehat{\ell}_t \right\rangle + \Phi(x_{1:t}, \widehat{\ell}_{1:t}) \right) = \Phi(x_{1:t-1}, \widehat{\ell}_{1:t-1})
\end{aligned} \tag{2}$$

and thus

$$\begin{aligned}
\mathbb{E}_{x_{1:T}}[\mathcal{R}_T] &= \mathbb{E}_{x_{1:T}} \left[\sum_{\tau=1}^{T} \left\langle p_\tau, \widehat{\ell}_\tau \right\rangle + \Phi(x_{1:T}, \widehat{\ell}_{1:T}) \right] \\
&\le \mathbb{E}_{x_{1:T-1}} \left[\sum_{\tau=1}^{T-1} \left\langle p_\tau, \widehat{\ell}_\tau \right\rangle + \Phi(x_{1:T-1}, \widehat{\ell}_{1:T-1}) \right] \le \cdots \le \Phi(\emptyset).
\end{aligned}$$

Therefore, the notion of Φ completely characterizes the optimal regret and algorithm. However, in general Φ is highly complicated and intractable, making the approach above only theoretically interesting. Nevertheless, in light of Eq. (2), if we can come up with a different and tractable function Rel and a strategy such that

$$\mathbb{E}_{x_t} \max_{\widehat{\ell}_t \in \mathcal{L}} \left(\left\langle p_t, \widehat{\ell}_t \right\rangle + \mathrm{Rel}(x_{1:t}, \widehat{\ell}_{1:t}) \right) \le \mathrm{Rel}(x_{1:t-1}, \widehat{\ell}_{1:t-1}) \tag{3}$$

and in addition $\Phi(x_{1:T}, \widehat{\ell}_{1:T}) \le \mathrm{Rel}(x_{1:T}, \widehat{\ell}_{1:T})$, then by the exact same argument we have $\mathbb{E}_{x_{1:T}}[\mathcal{R}_T] \le \mathrm{Rel}(\emptyset)$. Such function Rel is called a relaxation of Φ and is indeed an upper bound of Φ (check it yourself by a simple induction). The hope is thus Rel should be as small as possible so that the final regret bound $\mathrm{Rel}(\emptyset)$ is still of order $\mathcal{O}(\sqrt{T \ln N})$.

The question is now how to come up with a reasonable relaxation. To see this, we will simply let $\mathrm{Rel}(x_{1:T}, \widehat{\ell}_{1:T}) = \Phi(x_{1:T}, \widehat{\ell}_{1:T})$ and first see what $\mathrm{Rel}(x_{1:T-1}, \widehat{\ell}_{1:T-1})$ should be.

[1] We use the notation $z_{1:t}$ to denote the sequence $z_1, \ldots, z_t$.

2.1 Warm-up: Finding $\mathrm{Rel}(x_{1:T-1}, \widehat{\ell}_{1:T-1})$

Note that the existence of strategy p_t such that Eq. (3) holds implies

$$\mathbb{E}_{x_T} \min_{p_T \in \Delta(K)} \max_{\widehat{\ell}_T \in \mathcal{L}} \left[\left\langle p_T, \widehat{\ell}_T \right\rangle + \mathrm{Rel}(x_{1:T}, \widehat{\ell}_{1:T}) \right] \le \mathrm{Rel}(x_{1:T-1}, \widehat{\ell}_{1:T-1}).$$

We now work on the term on the left-hand side and relax it step by step:

$$\begin{aligned}
&\mathbb{E}_{x_T} \min_{p_T \in \Delta(K)} \max_{\widehat{\ell}_T \in \mathcal{L}} \left[\left\langle p_T, \widehat{\ell}_T \right\rangle + \Phi(x_{1:T}, \widehat{\ell}_{1:T}) \right] \\
&= \mathbb{E}_{x_T} \min_{p_T \in \Delta(K)} \max_{q_T \in \Delta(\mathcal{L})} \mathbb{E}_{\widehat{\ell}_T \sim q_T} \left[\left\langle p_T, \widehat{\ell}_T \right\rangle + \Phi(x_{1:T}, \widehat{\ell}_{1:T}) \right] \\
&= \mathbb{E}_{x_T} \max_{q_T \in \Delta(\mathcal{L})} \min_{p_T \in \Delta(K)} \mathbb{E}_{\widehat{\ell}_T \sim q_T} \left[\left\langle p_T, \widehat{\ell}_T \right\rangle + \Phi(x_{1:T}, \widehat{\ell}_{1:T}) \right] && \text{(Sion's minimax theorem)} \\
&= \mathbb{E}_{x_T} \max_{q_T \in \Delta(\mathcal{L})} \left(\left(\min_{p_T \in \Delta(K)} \mathbb{E}_{\widehat{\ell}'_T \sim q_T} \left[\left\langle p_T, \widehat{\ell}'_T \right\rangle \right] \right) + \mathbb{E}_{\widehat{\ell}_T \sim q_T} \left[\max_{\pi \in \Pi} - \sum_{t=1}^{T} \widehat{\ell}_t(\pi(x_t)) \right] \right) \\
&= \mathbb{E}_{x_T} \max_{q_T \in \Delta(\mathcal{L})} \mathbb{E}_{\widehat{\ell}_T \sim q_T} \left[\max_{\pi \in \Pi} \left(\left(\min_{p_T \in \Delta(K)} \mathbb{E}_{\widehat{\ell}'_T \sim q_T} \left[\left\langle p_T, \widehat{\ell}'_T \right\rangle \right] \right) - \widehat{\ell}_T(\pi(x_T)) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] \\
&\le \mathbb{E}_{x_T} \max_{q_T \in \Delta(\mathcal{L})} \mathbb{E}_{\widehat{\ell}_T \sim q_T} \left[\max_{\pi \in \Pi} \left(\mathbb{E}_{\widehat{\ell}'_T \sim q_T} \left[\widehat{\ell}'_T(\pi(x_T)) \right] - \widehat{\ell}_T(\pi(x_T)) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] \\
&\le \mathbb{E}_{x_T} \max_{q_T \in \Delta(\mathcal{L})} \mathbb{E}_{\widehat{\ell}_T, \widehat{\ell}'_T \sim q_T} \left[\max_{\pi \in \Pi} \left(\widehat{\ell}'_T(\pi(x_T)) - \widehat{\ell}_T(\pi(x_T)) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] \\
&= \mathbb{E}_{x_T} \max_{q_T \in \Delta(\mathcal{L})} \mathbb{E}_{\widehat{\ell}_T, \widehat{\ell}'_T \sim q_T, \sigma} \left[\max_{\pi \in \Pi} \left(\sigma \left(\widehat{\ell}'_T(\pi(x_T)) - \widehat{\ell}_T(\pi(x_T)) \right) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] \\
&\qquad \text{(σ is Rademacher variable, that is, uniformly drawn from $\{-1, 1\}$)} \\
&\le \mathbb{E}_{x_T} \max_{q_T \in \Delta(\mathcal{L})} \mathbb{E}_{\widehat{\ell}_T, \widehat{\ell}'_T \sim q_T, \sigma} \left[\max_{\pi \in \Pi} \left(\sigma \widehat{\ell}'_T(\pi(x_T)) - \frac{1}{2} \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) + \right. \\
&\qquad \left. \max_{\pi \in \Pi} \left(-\sigma \widehat{\ell}_T(\pi(x_T)) - \frac{1}{2} \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] \\
&= \mathbb{E}_{x_T} \max_{q_T \in \Delta(\mathcal{L})} \mathbb{E}_{\widehat{\ell}_T \sim q_T, \sigma} \left[\max_{\pi \in \Pi} \left(2\sigma \widehat{\ell}_T(\pi(x_T)) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] \\
&\qquad \text{(σ and $-\sigma$ follow the same distribution)} \\
&= \mathbb{E}_{x_T} \max_{\widehat{\ell}_T \in \mathcal{L}} \mathbb{E}_{\sigma} \left[\max_{\pi \in \Pi} \left(2\sigma \widehat{\ell}_T(\pi(x_T)) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] \\
&= \mathbb{E}_{x_T} \max_{\widehat{\ell}_T \in \mathcal{L}'} \mathbb{E}_{\sigma} \left[\max_{\pi \in \Pi} \left(2\sigma \widehat{\ell}_T(\pi(x_T)) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right]
\end{aligned}$$

where $\mathcal{L}' = \{\mathbf{0}, e_1, \ldots, e_K\}$ and last step is because maximizers of a convex function are always on the boundary. To continue, note that if $\widehat{\ell}_T = e_a$ for some $a \in [K]$, we can construct e'_a such that $e'_a(a) = 1$ and $e'_a(a')$ when $a' \neq a$ is an independent Rademacher variable, so that $\mathbb{E}_{e'_a}[e'_a] = e_a$ and

$$\begin{aligned}
\mathbb{E}_{\sigma} \left[\max_{\pi \in \Pi} \left(2\sigma e_a(\pi(x_T)) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] &= \mathbb{E}_{\sigma} \left[\max_{\pi \in \Pi} \left(2\sigma \mathbb{E}_{e'_a}[e'_a(\pi(x_T))] - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right] \\
&\le \mathbb{E}_{\sigma, e'_a} \left[\max_{\pi \in \Pi} \left(2\sigma e'_a(\pi(x_T)) - \sum_{t=1}^{T-1} \widehat{\ell}_t(\pi(x_t)) \right) \right].
\end{aligned}$$

Now if one looks at the random vector $\sigma e'_a$, it is clear that each coordinate of it is an independent Rademacher variable. We denote such random vector by ϵ_T and continue the bound with

$$\mathbb{E}_{\epsilon_T}\left[\max_{\pi\in\Pi}\left(2\epsilon_T(\pi(x_T)) - \sum_{t=1}^{T-1}\widehat{\ell}_t(\pi(x_t))\right)\right].$$

Note that this is also an upper bound when $\widehat{\ell}_T = \mathbf{0}$ since

$$\begin{aligned}\mathbb{E}_{\epsilon_T}\left[\max_{\pi\in\Pi}\left(2\epsilon_T(\pi(x_T)) - \sum_{t=1}^{T-1}\widehat{\ell}_t(\pi(x_t))\right)\right] &\geq \max_{\pi\in\Pi}\left(\mathbb{E}_{\epsilon_T}\left[2\epsilon_T(\pi(x_T)) - \sum_{t=1}^{T-1}\widehat{\ell}_t(\pi(x_t))\right]\right)\\ &= \max_{\pi\in\Pi}\left(-\sum_{t=1}^{T-1}\widehat{\ell}_t(\pi(x_t))\right).\end{aligned}$$

Therefore, we have shown

$$\mathbb{E}_{x_T}\min_{p_T\in\Delta(K)}\max_{\widehat{\ell}_T\in\mathcal{L}}\left[\left\langle p_T, \widehat{\ell}_T\right\rangle + \mathrm{Rel}(x_{1:T}, \widehat{\ell}_{1:T})\right] \leq \mathbb{E}_{x_T,\epsilon_T}\left[\max_{\pi\in\Pi}\left(2\epsilon_T(\pi(x_T)) - \sum_{t=1}^{T-1}\widehat{\ell}_t(\pi(x_t))\right)\right],$$

and can thus denote the last term by $\mathrm{Rel}(x_{1:T-1}, \widehat{\ell}_{1:T-1})$.

2.2 Generalizing the Argument

Compared to $\mathrm{Rel}(x_{1:T}, \widehat{\ell}_{1:T})$, one can see that in $\mathrm{Rel}(x_{1:T-1}, \widehat{\ell}_{1:T-1})$ the loss for the last round $-\widehat{\ell}_T(\pi(x_T))$ is replaced by the random loss $2\epsilon_T(\pi(x_T))$. This motivates us to define

$$\begin{aligned}\mathrm{Rel}(x_{1:t}, \widehat{\ell}_{1:t}) &= \mathbb{E}_{x_{t+1,T},\epsilon_{t+1,T}}\left[\max_{\pi\in\Pi}\left(2\sum_{\tau=t+1}^{T}\epsilon_\tau(\pi(x_\tau)) - \sum_{\tau=1}^{t}\widehat{\ell}_\tau(\pi(x_\tau))\right)\right]\\ &= \mathbb{E}_{x_{t+1,T},\epsilon_{t+1,T}}\left[\Phi(x_{1:T}, \widehat{\ell}_{1:t}, 2\epsilon_{t+1,T})\right], \qquad (4)\end{aligned}$$

which is saying that we should replace all the future losses by $2\epsilon_\tau$ and this leads to the worst-case regret. To make this algorithmic, we need to find a strategy such that Eq. (3) holds. While the most natural choice is

$$p_t = \operatorname*{argmin}_{p\in\Delta(K)}\max_{\widehat{\ell}_t\in\mathcal{L}}\left(\left\langle p, \widehat{\ell}_t\right\rangle + \mathrm{Rel}(x_{1:t}, \widehat{\ell}_{1:t})\right).$$

This still does not lead to an oracle-efficient algorithm since computing Rel is intractable. However, it turns out that the following strategy suffices:

$$p_t = \mathbb{E}_{x_{t+1,T},\epsilon_{t+1,T}}\left[p_t(x_{t+1,T}, \epsilon_{t+1,T})\right] \qquad (5)$$

where

$$p_t(x_{t+1,T}, \epsilon_{t+1,T}) = \operatorname*{argmin}_{p\in\Delta(K)}\max_{\widehat{\ell}_t\in\mathcal{L}}\left(\left\langle p, \widehat{\ell}_t\right\rangle + \Phi(x_{1:T}, \widehat{\ell}_{1:t}, 2\epsilon_{t+1,T})\right). \qquad (6)$$

Before discussing the oracle-efficiency of this strategy, let's first verify Eq. (3) is indeed satisfied.

Theorem 1. *The relaxation defined in Eq.* (4) *and the strategy defined in Eq.* (5) *satisfy Eq.* (3)*.*

Proof. The left-hand side of Eq. (3) can be bounded as follows:

$$\begin{aligned}&\mathbb{E}_{x_t}\max_{\widehat{\ell}_t\in\mathcal{L}}\left(\left\langle p_t, \widehat{\ell}_t\right\rangle + \mathrm{Rel}(x_{1:t}, \widehat{\ell}_{1:t})\right)\\ &= \mathbb{E}_{x_t}\max_{\widehat{\ell}_t\in\mathcal{L}}\left(\mathbb{E}_{x_{t+1,T},\epsilon_{t+1,T}}\left[\left\langle p_t(x_{t+1,T}, \epsilon_{t+1,T}), \widehat{\ell}_t\right\rangle\right] + \mathbb{E}_{x_{t+1,T},\epsilon_{t+1,T}}\left[\Phi(x_{1:T}, \widehat{\ell}_{1:t}, 2\epsilon_{t+1,T})\right]\right)\\ &\leq \mathbb{E}_{x_{t:T},\epsilon_{t+1,T}}\max_{\widehat{\ell}_t\in\mathcal{L}}\left(\left\langle p_t(x_{t+1,T}, \epsilon_{t+1,T}), \widehat{\ell}_t\right\rangle + \Phi(x_{1:T}, \widehat{\ell}_{1:t}, 2\epsilon_{t+1,T})\right)\\ &= \mathbb{E}_{x_{t:T},\epsilon_{t+1,T}}\min_{p\in\Delta(K)}\max_{\widehat{\ell}_t\in\mathcal{L}}\left(\left\langle p, \widehat{\ell}_t\right\rangle + \Phi(x_{1:T}, \widehat{\ell}_{1:t}, 2\epsilon_{t+1,T})\right).\end{aligned}$$

By repeating the exact same argument in the warm-up section, the last quantity is bounded by $\mathrm{Rel}(x_{1:t-1}, \widehat{\ell}_{1:t-1})$. □

Algorithm 1: Water-filling

Input: $B_1 \leq \cdots \leq B_K$
Ouput: solution of $\operatorname{argmin}_{p \in \Delta(K)} \max_{a \in [K]} (p(a) + B_a)$
Initialization: let $p = \mathbf{0}$, $S = 1$, $B_{K+1} = +\infty$
for $i = 1, \ldots, K$ **do**
  $s = \min\{(B_{i+1} - B_i)i, S\}$
  for $j = 1, \ldots, i$ **do** $p(j) \leftarrow p(j) + s/i$
  $S \leftarrow S - s$

Finally, is this relaxation tight enough? In other words, how large is the regret bound $\mathrm{Rel}(\emptyset)$? In fact, $\mathrm{Rel}(\emptyset)$ is a variant of the *Rademacher complexity* of the policy class Π and can be shown to be at most $2\sqrt{2T \ln N}$. This means that the relaxation is very tight and the strategy above enjoys the optimal regret.

2.3 Oracle-efficiency

To see why the strategy is efficient, first note that since we only care about expected regret, there is no difference in playing p_t or $p_t(x_{t+1,T}, \epsilon_{t+1,T})$ with a random draw of $x_{t+1,T}, \epsilon_{t+1,T}$. It thus suffices to solve the optimization problem defined in Eq. (6). It is not hard to see that the maximum over $\mathcal{L}$ can only be obtained by one of the K basis vectors $e_1, \ldots, e_K$.[2] Therefore, with

$$B_a = \Phi(x_{1:T}, \widehat{\ell}_{1:t-1}, e_a, 2\epsilon_{t+1,T}),$$

which clearly can be computed by calling the oracle once, the optimization problem becomes

$$\operatorname*{argmin}_{p \in \Delta(K)} \max_{a \in [K]} (p(a) + B_a).$$

This can be simply solved by a so-called water-filling procedure. Specifically, assuming $B_1 \leq \cdots \leq B_K$ without loss of generality, the solution can be found by Algorithm 1. It is therefore clear that for each round the algorithm makes K oracle-calls and all the other operations run in time $\mathrm{poly}(T, K)$.

Note that the algorithm shares some similarity with FTPL, both of which hallucinate some fake data and solve an offline optimization problem involving both the observed data and the hallucinated data.

References

Alexander Rakhlin and Karthik Sridharan. Bistro: An efficient relaxation-based method for contextual bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Vasilis Syrgkanis, Haipeng Luo, Akshay Krishnamurthy, and Robert E Schapire. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in Neural Information Processing Systems 29*, 2016.

[2] Indeed, the objective is convex in $\widehat{\ell}_t$ and thus the maximum over $\mathcal{L}$ can only be obtained by $\mathcal{L}'$. Moreover, one can verify that $\mathbf{0}$ can not be the maximum since the average of the other K choices is larger.