Lecture 21

Instructor: Haipeng Luo

1 Softening Policy Elimination

In this lecture we are finally ready to discuss the state-of-the-art algorithm for the i.i.d. contextual bandit problem, which is both optimal and oracle-efficient [Agarwal et al., 2014]. Recall that the idea of Policy Elimination is to find $P_t \in \Delta(\Pi_t)$ such that $V(P_t, \pi) \leq 2K$ for all $\pi \in \Pi_t$ where

$$V(P, \pi) = \mathbb{E}_x \left[\frac{1}{P^\mu(\pi(x)|x)} \right]$$

is essentially the variance of the loss estimators that we want to control. To obtain an efficient algorithm, we need to forget about the idea of removing policies from Π. So is it possible to ensure $V(P_t, \pi) \leq 2K$ for all $\pi \in \Pi$ while at the same time P_t puts most of the weights on good policies?

Unfortunately this is too strong of a requirement. For example, if there is a bad policy π_{bad} which always picks a bad action a_{bad} with loss 1 and no other policy ever picks a_{bad}, then

$$2K \geq V(P_t, \pi_{\text{bad}}) = \mathbb{E}_x \left[\frac{1}{P_t^\mu(a_{\text{bad}}|x)} \right] = \frac{1}{(1 - K\mu)P_t(\pi_{\text{bad}}) + \mu}$$

which implies that $P_t(\pi_{\text{bad}})$ will be pretty large assuming μ is small. This is clearly not a good algorithm.

From this example, however, we can see that the condition $V(P_t, \pi) \leq 2K$ should be somehow relaxed for bad policies. Just as in Policy Elimination, whether a policy is good or bad can be roughly determined by its empirical performance compared to the empirically best policy. Specifically, recall the notation $\bar{\ell}_t(\pi) = \frac{1}{t} \sum_{\tau=1}^t \widehat{\ell}_\tau(\pi(x_\tau))$ for the empirical average loss and $\pi_t^\star = \operatorname{argmin}_{\pi \in \Pi} \bar{\ell}_t(\pi)$ for the empirically best policy up to time t. Define empirical average regret for a policy π to be

$$\text{Reg}_t(\pi) = \bar{\ell}_t(\pi) - \bar{\ell}_t(\pi_t^\star).$$

We now relax the low-variance condition as: find P_t such that

$$V(P_t, \pi) \leq 2K + \beta \text{Reg}_{t-1}(\pi) \quad \forall\, \pi \in \Pi$$

for some parameter $\beta > 0$ to be specified later. Now there is hope to impose exploitation simultaneously. Specifically, we want $\sum_{\pi \in \Pi} P_t(\pi)\text{Reg}_{t-1}(\pi)$ to be as small as possible. How small can it be? The following lemma answers this question.

Lemma 1. *For any $\beta > 0$, there always exists a distribution $P \in \Delta(\Pi)$ such that*

$$\sum_{\pi \in \Pi} P(\pi)\text{Reg}_{t-1}(\pi) \leq \frac{2K}{\beta}$$
$$V(P, \pi) \leq 2K + \beta \text{Reg}_{t-1}(\pi) \quad \forall\, \pi \in \Pi.$$

Proof. Define function $F_t : \Delta(\Pi) \to \mathbb{R}_+$ as

$$F_t(P) = \sum_{\pi \in \Pi} P(\pi)\text{Reg}_{t-1}(\pi) + \frac{2}{\beta} \mathbb{E}_x \left[\sum_{a=1}^K \ln \frac{1}{P^\mu(a|x)} \right].$$

The claim is that the minimizer of $F_t(P)$, which always exists due to compactness of $\Delta(\Pi)$ and continuousness of F_t, satisfies both conditions. To see this, first notice that we can extend the function to a set of "sub-distributions" $\Delta(\Pi)' = \{P \in \mathbb{R}_+^N : \sum_{\pi\in\Pi} P(\pi) \leq 1\}$ and still have

$$\min_{P\in\Delta(\Pi)} F_t(P) = \min_{P\in\Delta(\Pi)'} F_t(P).$$

This is because for any sub-distribution $P \in \Delta(\Pi)'$, one can make it a distribution by increasing the weight for policy $\pi^\star_{t-1}$ until the weights sum up to 1. This will only decrease the function value since $\mathrm{Reg}_{t-1}(\pi^\star_{t-1}) = 0$ and the second term of F_t is decreasing in any coordinate of P.

Next note that the derivate of F_t with respect to a policy π is

$$\nabla F_t(P)(\pi) = \mathrm{Reg}_{t-1}(\pi) - \frac{2(1-K\mu)}{\beta} V(P,\pi).$$

Let P^* be a minimizer of F_t over $\Delta(\Pi)'$. By KKT conditions, we have

$$\mathrm{Reg}_{t-1}(\pi) - \frac{2(1-K\mu)}{\beta} V(P^*,\pi) - \lambda_\pi + \lambda = 0 \tag{1}$$

for some Lagrangian multipliers $\lambda_\pi \geq 0$ and $\lambda \geq 0$. Multiply both sides by $P^*(\pi)$ and sum over $\pi \in \Pi$ gives

$$\begin{aligned}
\sum_{\pi\in\Pi} P^\star(\pi)\mathrm{Reg}_{t-1}(\pi) &= \frac{2(1-K\mu)}{\beta}\sum_{\pi\in\Pi} P^*(\pi)V(P^*,\pi) + \sum_{\pi\in\Pi} P^*(\pi)\lambda_\pi - \lambda \quad && (P^* \in \Delta(\Pi)) \\
&= \frac{2(1-K\mu)}{\beta}\sum_{\pi\in\Pi} P^*(\pi)V(P^*,\pi) - \lambda && \text{(complementary slackness)} \\
&\leq \frac{2}{\beta}\mathbb{E}_x\left[\sum_{\pi\in\Pi}\frac{P^*(\pi)}{P^*(\pi(x)|x)}\right] - \lambda = \frac{2K}{\beta} - \lambda \leq \frac{2K}{\beta},
\end{aligned}$$

showing that P^* satisfies the first condition. Moreover, the last equality above also implies $\lambda \leq \frac{2K}{\beta}$ since $\mathrm{Reg}_{t-1}(\pi) \geq 0$. Rearranging Eq. (1) thus gives

$$V(P^*,\pi) \leq \frac{\beta}{2(1-K\mu)}\left(\mathrm{Reg}_{t-1}(\pi) + \lambda\right) \leq 2K + \beta\mathrm{Reg}_{t-1}(\pi),$$

where we assume $\mu \leq \frac{1}{2K}$ so that $2(1-K\mu) \geq 1$ (since otherwise we trivially have $V(P^*,\pi) \leq 1/\mu \leq 2K$). This shows that $P^\star$ satisfies the second condition too. $\square$

The question is now what β we should use. Assuming $\mathrm{Reg}_{t-1}(\pi)$ concentrates well around the actual expected regret of π compared to $\pi^\star$,

$$\mathrm{Reg}(\pi) \stackrel{\text{def}}{=} \bar{\ell}(\pi) - \bar{\ell}(\pi^\star),$$

which is exactly what we hope for, $\mathrm{Reg}_{t-1}(\pi)$ should be at most a constant. A reasonable choice of β would then be of order $1/\mu$, since $V(P,\pi)$ is trivially bounded by $1/\mu$. In other words, when a policy π is good, which means $\mathrm{Reg}_{t-1}(\pi)$ is close to zero, we still require $V(P,\pi)$ to be close to $2K$, while when the policy is bad, which means $\mathrm{Reg}_{t-1}(\pi)$ is a large constant, then there is almost no requirement on $V(P,\pi)$ with this choice of β.

On the other hand, this means that the exploitation constraint is $\sum_{\pi\in\Pi} P(\pi)\mathrm{Reg}_{t-1}(\pi) = \mathcal{O}(K\mu)$, which also makes sense because μ should be of order $1/\sqrt{T}$, and if the per round regret is of order $1/\sqrt{T}$, then the overall regret over T rounds is of order $\sqrt{T}$. With some specific constant (chosen based on the analysis), this leads to the final algorithm called ILOVETOCONBANDITS (see Algorithm 1).

2 Oracle-Efficiency

To discuss oracle-efficiency, keep in mind that as in Policy Elimination, the true context distribution in the definition of V can be replaced by the empirical distribution of observed contexts, that is, a

Algorithm 1: ILOVETOCONBANDITS (colloquially referred as Mini-monster)

Input: failure probability $\delta \in (0, 1)$

Initialization: let $\mu = \min\left\{\frac{1}{K}, \sqrt{\frac{\ln(TN/\delta)}{TK}} \ln T\right\}$

for $t = 1, \ldots, T$ **do**

find P_t such that

$$\sum_{\pi \in \Pi} P_t(\pi)\text{Reg}_{t-1}(\pi) \leq 20K\mu$$

$$V(P_t, \pi) \leq 2K + \frac{\text{Reg}_{t-1}(\pi)}{10\mu} \quad \forall\, \pi \in \Pi.$$

play $a_t \sim P_t^\mu(\cdot|x_t)$

uniform distribution over $x_1, \ldots, x_{t-1}$ at time t. (For simplicity, the analysis of next section will assume that the true context distribution is known instead.)

According to the proof of Lemma 1, to find distribution P_t it suffices to solve the optimization problem $\text{argmin}_{P \in \Delta(\Pi)} F_t(P)$ (in fact an approximate solution is enough). This is in fact very similar to FTRL with a special regularizer. To see how to solve it efficiently with the oracle, notice that the derivative of $F_t(P)$ with respect to a policy π can be written as (with $\beta = 1/(10\mu)$)

$$\nabla F_t(P)(\pi) = \frac{1}{t-1}\sum_{\tau=1}^{t-1}\left(\widehat{\ell}_\tau(\pi(x_\tau)) - \widehat{\ell}_\tau(\pi^\star_{t-1}(x_\tau))\right) - \frac{20\mu(1-K\mu)}{(t-1)}\sum_{\tau=1}^{t-1}\frac{1}{P^\mu(\pi(x_\tau)|x_\tau)}.$$

Since the part involving $\pi^\star_{t-1}$ is independent of π, if we feed the oracle with a training set

$$\mathcal{S} = \left\{\left(x_1, \widehat{\ell}_1 - \frac{20\mu(1-K\mu)}{P^\mu(\,\cdot\,|x_1)}\right), \cdots, \left(x_{t-1}, \widehat{\ell}_{t-1} - \frac{20\mu(1-K\mu)}{P^\mu(\,\cdot\,|x_{t-1})}\right)\right\},$$

we have ERM$(\mathcal{S}) = \text{argmin}_{\pi \in \Pi} \nabla F_t(P)(\pi)$. In other words, the oracle can tell us the minimum coordinate of the gradient of $F_t(P)$ for any P, which opens up many possibilities to utilize the theory of optimization to find P_t. For example, one can directly apply the Frank-Wolfe algorithm (also known as conditional gradient method). Specifically, for a constraint convex optimization problem $\min_{w \in \Omega} f(w)$, the Frank-Wolfe algorithm performs the following iterative updates (staring with an arbitrary $w_1 \in \Omega$):

$$v_k = \operatorname*{argmin}_{v \in \Omega}\, \langle v, \nabla f(w_k)\rangle$$
$$w_{k+1} = (1-\gamma_k)w_k + \gamma_k v_k$$

for some step-size γ_k (default choice is $2/(k+1)$). When Ω is the simplex, the first step is exactly to find the minimum coordinate of the gradient. Therefore, with the oracle we can implement the Frank-Wolfe algorithm to solve $\text{argmin}_{P \in \Delta(\Pi)} F_t(P)$ efficiently. We omit the details on how many iterations are needed but it will be polynomial in T, K, and $\ln N$.

Importantly, notice that unlike gradient descent, Frank-Wolfe leads to a sparse solution: when Ω is the simplex, after k rounds w_k has only k non-zero coordinates (assuming w_1 concentrates on one element to start with). This means that P_t's are all sparse distributions and operations involving P_t, such as constructing the training set $\mathcal{S}$ and sampling a_t, are all efficient.

Instead of using Frank-Wolfe, another possibility is to do some kind of coordinate descent: iteratively use the oracle to fine the coordinate with minimum derivative and adjust the weight for this coordinate appropriately. This is exactly the method taken in [Agarwal et al., 2014]. In fact, with additional tricks that are specialized for this task, it was shown that over T rounds only $\mathcal{O}(\sqrt{T})$ oracle calls are needed, which also implies that all P_t's are $\mathcal{O}(\sqrt{T})$-sparse.

3 Regret Analysis

Finally in this section we prove that ILOVETOCONBANDITS enjoys optimal regret. The key is to show the following concentration results on regret.

Lemma 2. *With probability* $1-\delta/2$*, Algorithm 1 ensures that for all* $t \in [T]$ *and all* $\pi \in \Pi$*,*

$$\mathrm{Reg}(\pi) \le 2\mathrm{Reg}_t(\pi) + \epsilon_t \quad \textit{and} \quad \mathrm{Reg}_t(\pi) \le 2\mathrm{Reg}(\pi) + \epsilon_t$$

where $\epsilon_t = \frac{20C}{\mu t} + 15K\mu$ *and* $C = \ln\left(\frac{4NT}{\delta}\right)\ln T$.

Proof. By Freedman's inequality and a union bound, we have with probability $1-\delta/2$, for all $t \in [T]$, all $\pi \in \Pi$, and any $\lambda \in [0,\mu]$,

$$|\bar{\ell}_t(\pi) - \bar{\ell}(\pi)| \le \frac{\lambda}{t}\sum_{\tau=1}^{t} V(P_\tau, \pi) + \frac{\ln\left(\frac{4NT}{\delta}\right)}{\lambda t}.$$

Specifically picking $\lambda = \frac{\mu}{\ln T}$ gives

$$|\bar{\ell}_t(\pi) - \bar{\ell}(\pi)| \le \frac{\mu}{t\ln T}\sum_{\tau=1}^{t} V(P_\tau, \pi) + \frac{C}{\mu t}. \tag{2}$$

Now we use induction to prove the lemma. The base case $t=0$ is trivial. Assuming the statement holds for all rounds before time t, we have by the algorithm

$$V(P_\tau, \pi) \le 2K + \frac{\mathrm{Reg}_{\tau-1}(\pi)}{10\mu} \le 2K + \frac{\mathrm{Reg}(\pi)}{5\mu} + \frac{\epsilon_{\tau-1}}{10\mu} \tag{3}$$

for all $\tau = 2, \ldots, t$ and $V(P_1, \pi) \le 2K$. Therefore, we have

$$\begin{aligned}
\mathrm{Reg}(\pi) - \mathrm{Reg}_t(\pi) &= \bar{\ell}(\pi) - \bar{\ell}(\pi^\star) - \bar{\ell}_t(\pi) + \bar{\ell}_t(\pi_t^\star) \\
&\le \bar{\ell}(\pi) - \bar{\ell}(\pi^\star) - \bar{\ell}_t(\pi) + \bar{\ell}_t(\pi^\star) && \text{(by optimality of } \pi_t^\star) \\
&\le \frac{2C}{\mu t} + \frac{\mu}{t\ln T}\sum_{\tau=1}^{t}\left(V(P_\tau,\pi) + V(P_\tau,\pi^\star)\right) && \text{(by Eq. (2))} \\
&\le \frac{2C}{\mu t} + \frac{4K\mu}{\ln T} + \frac{\mathrm{Reg}(\pi)}{5\ln T} + \frac{1}{5t\ln T}\sum_{\tau=2}^{t}\epsilon_{\tau-1} && \text{(by Eq. (3) and Reg}(\pi^\star)=0) \\
&\le \frac{2C}{\mu t} + \frac{4K\mu}{\ln T} + \frac{\mathrm{Reg}(\pi)}{5\ln T} + \frac{8C}{\mu t} + \frac{3K\mu}{\ln T} && \text{(by plugging in } \epsilon_\tau) \\
&\le \frac{10C}{\mu t} + 7K\mu + \frac{\mathrm{Reg}(\pi)}{2} \le \frac{\epsilon_t}{2} + \frac{\mathrm{Reg}(\pi)}{2}.
\end{aligned}$$

Rearranging proves $\mathrm{Reg}(\pi) \le 2\mathrm{Reg}_t(\pi) + \epsilon_t$. Similarly, we also have

$$\begin{aligned}
\mathrm{Reg}_t(\pi) - \mathrm{Reg}(\pi) &= \bar{\ell}_t(\pi) - \bar{\ell}_t(\pi_t^\star) - \bar{\ell}(\pi) + \bar{\ell}(\pi^\star) \\
&\le \bar{\ell}_t(\pi) - \bar{\ell}_t(\pi_t^\star) - \bar{\ell}(\pi) + \bar{\ell}(\pi_t^\star) && \text{(by optimality of } \pi^\star) \\
&\le \frac{2C}{\mu t} + \frac{\mu}{t\ln T}\sum_{\tau=1}^{t}\left(V(P_\tau,\pi) + V(P_\tau,\pi_t^\star)\right) && \text{(by Eq. (2))} \\
&\le \frac{2C}{\mu t} + \frac{4K\mu}{\ln T} + \frac{\mathrm{Reg}(\pi)}{5\ln T} + \frac{\mathrm{Reg}(\pi_t^\star)}{5\ln T} + \frac{1}{5t\ln T}\sum_{\tau=2}^{t}\epsilon_{\tau-1} && \text{(by Eq. (3))} \\
&\le \frac{2C}{\mu t} + \frac{4K\mu}{\ln T} + \frac{\mathrm{Reg}(\pi)}{5\ln T} + \frac{\epsilon_t}{5\ln T} + \frac{8C}{\mu t} + \frac{3K\mu}{\ln T} && (4) \\
&\le \frac{14C}{\mu t} + 10K\mu + \mathrm{Reg}(\pi) \le \epsilon_t + \mathrm{Reg}(\pi),
\end{aligned}$$

where Step (4) uses the fact $\mathrm{Reg}(\pi) \le 2\mathrm{Reg}_t(\pi) + \epsilon_t$ just proven above with π set to $\pi_t^\star$, and also the fact $\mathrm{Reg}_t(\pi_t^\star) = 0$. Rearranging then proves $\mathrm{Reg}_t(\pi) \le 2\mathrm{Reg}(\pi) + \epsilon_t$ as well. $\square$

The final regret bound is now a simple application of this lemma and the exploitation constraint of the algorithm.

Theorem 1. *Algorithm 1 ensures that with probability* $1-\delta$*, we have* $\mathcal{R}_T = \widetilde{\mathcal{O}}\left(\sqrt{TK\ln(N/\delta)}\right)$.

Proof. The first step is exactly the same as analyzing Policy Elimination: by Azuma's inequality we have with probability $1-\delta/2$,

$$\sum_{t=1}^{T} \ell_t(a_t) \leq \sum_{t=1}^{T} \sum_{\pi\in\Pi} P_t(\pi)\bar{\ell}(\pi) + TK\mu + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right).$$

Conditioning on this event and the event stated in Lemma 2, which happen simultaneously with probability $1-\delta$, we have

$$\begin{aligned}
\mathcal{R}_T &\leq \sum_{t=1}^{T} \sum_{\pi\in\Pi} P_t(\pi)\mathrm{Reg}(\pi) + TK\mu + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right) \\
&\leq 2\sum_{t=1}^{T} \sum_{\pi\in\Pi} P_t(\pi)\mathrm{Reg}_{t-1}(\pi) + \sum_{t=2}^{T} \epsilon_{t-1} + TK\mu + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right) \\
&\leq 56TK\mu + \frac{40C\ln T}{\mu} + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right), \qquad \text{(by the exploitation constraint)}
\end{aligned}$$

which is of order $\widetilde{\mathcal{O}}\left(\sqrt{TK\ln(N/\delta)}\right)$ with the optimal tuning of μ. $\square$

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.