Note on Contextual Bandits with Online Regression Oracles

Wen Sun

1 Setting

We consider the following Contextual bandit model. At round t, the environment presents a context $x_t \in \mathcal{X}$; the learner selects a distribution $p_t \in \Delta(\mathcal{A})$ over actions and samples an action $a_t \sim p_t$; the learner receives a reward signal $r_t \sim R(x_t, a_t)$ where $R(x, a)$ is the distribution of the reward under context-action pair (x, a). The game proceeds for T many rounds. Without loss of generality, we assume reward is always normalized in the sense that it is bounded between $[0, 1]$.

We now define regret. Denote $f^\star(x, a) = \mathbb{E}_{r\sim R(x,a)}[r]$ as the expected reward under (x, a). The optimal policy is defined as $\pi^\star(x) = \arg\max_a f^\star(x, a)$. Thus, the regret is defined as follows:

$$\text{Regret} = \sum_{t=0}^{T-1} f^\star(x_t, \pi^\star(x)) - \sum_{t=0}^{T-1} \mathbb{E}_{a\sim p_t} f^\star(x_t, a_t).$$

Function approximation setup and online regression oracle We will use function approximation. Define $\mathcal{F} \subset \mathcal{X} \times \mathcal{A} \mapsto [0, 1]$ as a class of functions which aim to capture $f^\star$.

We assume that we have an online regression oracle. More formally, at iteration t, given context x_t, a_t, before seeing the realized reward r_t, the regression oracle selects f_t and predicts reward $f_t(x_t, a_t)$; it then sees the reward r_t, and suffers loss $(f_t(x_t, a_t) - r_t)^2$. We assume that the online regression oracle achieves has bounded regret, i.e.,

$$\sum_{t=0}^{T-1}(f_t(x_t, a_t) - r_t)^2 - \min_{f\in\mathcal{F}} \sum_{t=0}^{T-1}(f(x_t, a_t) - r_t)^2 = \text{Reg}(T). \tag{1}$$

Here $\text{Reg}(T)$ typically grows sublinear. One example is that when $\mathcal{F}$ is a discrete class, then there exists algorithm which can have $\text{Reg}(T) = O(\ln(T)\ln(|\mathcal{F}|))$. Note the $\ln$-dependence on the size of the function class, which means that the function class can be exponentially large. For continuous function class, when $\mathcal{F}$ is convex, we can also have $\text{Reg}(T)$ scaling in the order of $\ln(T)$.

2 A general algorithmic framework

First of all, using the fact that $\mathbb{E}[r_t|x_t, a_t] = f^\star(x_t, a_t)$ and $a_t \sim p_t \in \Delta(\mathcal{A})$, the regret on the square loss in (1) implies the following. With probability at least $1 - \delta$, we have:

$$\forall t \leq T : \sum_{i=0}^{t-1} \mathbb{E}_{a\sim p_t} \left(f_t(x_t, a) - f^\star(x_t, a)\right)^2 = O\left(\text{Reg}(T) + \ln(1/\delta)\right). \tag{2}$$

This step is standard in minimizing square loss, and we will defer the proof to the appendix. Intuitively, this means that f_t is doing well compared to the Bayes optimal $f^\star$.

Now we consider the following meta algorithm which defines p_t in iteration t using the following min-max procedure. Given the context x_t, we perform

$$p_t = \operatorname*{argmin}_{p\in\Delta(\mathcal{A})} \max_{f\in\mathcal{F}} \underbrace{\left(\max_a f(x_t,a) - \mathbb{E}_{a\sim p}[f(x_t,a)]\right)}_{\text{“Regret" under } x_t \text{ and function } f} - \lambda \underbrace{\mathbb{E}_{a\sim p}\left(f(x_t,a) - f_t(x_t,a)\right)^2}_{\text{Regularization: constrain } f \text{ near } f_t} \tag{3}$$

The algorithm then will sample $a_t \sim p_t$, receive receive r_t, and call the online regression oracle to update function f_t to f_{t+1}, and then move on to the iteration $t+1$.

We define the Decision Estimation Coefficient β as follows.

$$\beta/\lambda := \max_{x\in\mathcal{X}, g\in\mathcal{F}} \min_p \max_{f\in\mathcal{F}} \left(\max_a f(x,a) - \mathbb{E}_{a\sim p_t} f(x,a)\right) - \lambda \mathbb{E}_{a\sim p}\left(f(x,a) - g(x,a)\right)^2 \tag{4}$$

The following theorem converts the DEC β and the online regression oracle regret $\text{Reg}(T)$ into the regret of our algorithm in (3).

Theorem 1. *Consider the CB algorithm which updates f_t using an online regression oracle, and computes p_t as in Eq. 3. Then with probability at least $1-\delta$, the regret of the algorithm is upper bounded by $O(\sqrt{T\beta(Reg(T) + \ln(1/\delta))})$.*

Proof. The definition of DEC in (4) implies that for our choice of p_t at iteration t under context x_t, we have:

$$\max_{f\in\mathcal{F}} \mathbb{E}_{a\sim p_t}\left(\max_a f(x_t,a) - f(x_t,a)\right) - \lambda \mathbb{E}_{a\sim p_t}\left(f(x_t,a) - f_t(x_t,a)\right)^2 \leq \beta/\lambda.$$

Now let us revisit the CB regret definition.

$$\begin{aligned}
\text{Regret} &= \sum_{t=0}^{T-1} \max_a f^\star(x_t,a) - \sum_{t=0}^{T-1} \mathbb{E}_{a\sim p_t} f^\star(x_t,a) \\
&= \sum_{t=0}^{T-1}\left(\max_a f^\star(x_t,a) - \mathbb{E}_{a\sim p_t} f^\star(x_t,a) - \lambda \mathbb{E}_{a\sim p_t}(f^\star(x_t,a) - f_t(x_t,a))^2\right) \\
&\qquad + \lambda \sum_{t=0}^{T-1} \mathbb{E}_{a\sim p_t}(f^\star(x_t,a) - f_t(x_t,a))^2 \\
&\leq T\beta/\lambda + \lambda(\text{Reg}(T) + \ln(1/\delta))
\end{aligned}$$

where the last inequality uses the fact that $f^\star \in \mathcal{F}$, and also the regret bound on $\sum_t \mathbb{E}_{a\sim p_t}(f^\star(x_t,a) - f_t(x_t,a)) \leq \text{Reg}(T) + \ln(1/\delta)$.

Set $\lambda = T\beta/(\text{Reg}(T) + \ln(1/\delta))$, we have:

$$\text{Regret} \leq 2\sqrt{T\beta(\text{Reg}(T) + \ln(1/\delta))}.$$

□

3 Inverse Gap Weighting

So far we have seen that if we can solve (3) – the minmax procedure, and the DEC defined in (4) is bounded, then we achieve a $\sqrt{T}$ regret bound (assuming $\text{Reg}(T) = O(\ln(T))$). However, solving the minmax problem formed in (3) can be computationally challenging in general – a naive approach is to search over all possible $p \in \Delta(\mathcal{A})$ and all

$f \in \mathcal{F}$ which clearly is not computationally efficient. Moreover, it also seems not that straightforward to check if β in (4) is small as it involves complicated $\max, \min, \max$.

Luickly, for contextual bandit, there is a simple approach to construct a distribution p_t which satisfies the following:

$$\max_{x\in\mathcal{X},g\in\mathcal{F}} \max_{f\in\mathcal{F}} \Big(\max_a f(x,a) - \mathbb{E}_{a\sim p_t} f(x,a)\Big) - \lambda \mathbb{E}_{a\sim p_t} \left(f(x,a) - g(x,a)\right)^2 \leq O(\frac{A}{\lambda}),$$

which implies that $\beta \leq O(A)$, where $A = |\mathcal{A}|$. The way to construct such a p_t is through the approach called *Inverse Gap Weighting* (IGW). IGW is formally defined as follows. Given any function $g \in \mathcal{F}$ and context x, $\text{IGW}(g,x) \in \Delta(\mathcal{A})$ is a distribution over actions defined as follows. Denote $\tilde{a} = \text{argmax}_{a\in\mathcal{A}} g(x,a)$.

$$\text{IGW}(g,x)[a] = \frac{1}{A + \lambda(g(x,\tilde{a}) - g(x,a))}, \quad \text{IGW}(g,x)[\tilde{a}] = 1 - \sum_{a\neq\tilde{a}} \text{IGW}(g,x)[a].$$

Using IGW, we can compupte p_t in iteration t as follows: $p_t = \text{IGW}(f_t, x_t)$. Note that p_t is not necessarily the minimizer in (3), instead, it should be considered as an approximated minimizer.

The following lemma shows that using IGW, we indeed can upper bound the DEC β by A.

Lemma 2. *For any $x \in \mathcal{X}$ and any $g \in \mathcal{G}$, define $p = IGW(g,x)$, we must have:*

$$\max_{f\in\mathcal{F}} \left[(\max_a f(x,a) - \mathbb{E}_{a\sim p} f(x,a)) - \lambda \mathbb{E}_{a\sim p}(f(x,a) - g(x,a))^2\right] \leq (4A)/\lambda,$$

for all $\lambda \in \mathbb{R}^+$.

Proof. Let us consider any $f \in \mathcal{F}$ and show that the above holds for $f \in \mathcal{F}$.

Denote $a^\star = \arg\max_a f(x,a)$ and recall $\tilde{a} = \arg\max_a g(x,a)$. For the regret on x and f, we have:

$$\mathbb{E}_{a\sim p}(f(x,a^\star) - f(x,a)) = \sum_{a\neq a^\star} p(a)(f(x,a^\star) - f(x,a))$$
$$= \sum_{a\neq a^\star} p(a) \left[\underbrace{f(x,a^\star) - g(x,a^\star)}_{T_1} + \underbrace{g(x,a^\star) - g(x,\tilde{a})}_{T_2} + \underbrace{g(x,\tilde{a}) - g(x,a)}_{T_3} + \underbrace{g(x,a) - f(x,a)}_{T_4}\right]$$

Let us first bound T_4. For T_4, apply AM-GM we have:

$$\sum_{a\neq a^\star}\left[\frac{p(a)}{4\lambda} + p(a)\lambda(g(x,a) - f(x,a))^2\right] = \frac{1 - p(a^\star)}{4\lambda} + \lambda \sum_{a\neq a^\star} p(a)(g(x,a) - f(x,a))^2 \tag{5}$$
$$\leq \frac{1}{4\lambda} + \lambda \sum_{a\neq a^\star} p(a)(g(x,a) - f(x,a))^2. \tag{6}$$

Now let us apply AM-GM on T_1. We have:

$$(1 - p(a^\star))(f(x,a^\star) - g(x,a^\star)) \leq \frac{(1 - p(a^\star))^2}{4\lambda p(a^\star)} + \lambda p(a^\star)(f(x,a^\star) - g(x,a^\star))^2 \tag{7}$$
$$\leq \frac{1}{4\lambda p(a^\star)} + \lambda p(a^\star)(f(x,a^\star) - g(x,a^\star))^2. \tag{8}$$

The $\frac{1}{4\lambda p(a^\star)}$ term will be used together with the term T_2 below. The $\lambda p(a^\star)(f(x,a^\star) - g(x,a^\star))^2$ term can be combined together with the term $\lambda \sum_{a\neq a^\star} p(a)(g(x,a) - f(x,a))^2$ in Eq. 6, to cancel out the $\lambda \mathbb{E}_{a\sim p}(f(x,a) - g(x,a))^2$ term in the key inequality in the lemma.

Now let us bound the term T_3.

$$\sum_{a\neq a^\star} p(a)(g(x,\tilde{a}) - g(x,a)) = \sum_{a\neq a^\star} \frac{1}{A + \lambda(g(x,\tilde{a}) - g(x,a))}(g(x,\tilde{a}) - g(x,a)) \leq \frac{A-1}{\lambda}.$$

Now let us consider term T_2, combined it with the term $1/(4\lambda p(a^\star))$ left from Eq. 8.

$$(1 - p(a^\star))(g(x,a^\star) - g(x,\tilde{a})) + \frac{1}{4\lambda p(a^\star)}.$$

To proceed, we consider two cases below.

First case: when $a^\star \neq \tilde{a}$, then we have

$$\begin{aligned}
&(1 - p(a^\star))(g(x,a^\star) - g(x,\tilde{a})) + \frac{1}{4\lambda p(a^\star)} \\
&= (1 - \frac{1}{A + \lambda(g(x,\tilde{a}) - g(x,a^\star))})(g(x,a^\star) - g(x,\tilde{a})) + \frac{A + \lambda(g(x,\tilde{a}) - g(x,a^\star))}{4\lambda} \\
&\leq (g(x,a^\star) - g(x,\tilde{a})) + \frac{1}{\lambda} + \frac{A}{4\lambda} + \frac{g(x,\tilde{a}) - g(x,a^\star)}{4} \leq \frac{1}{\lambda} + \frac{A}{4\lambda}.
\end{aligned}$$

where the first inequality comes from the fact that $\frac{1}{A+\lambda(g(x,\tilde{a})-g(x,a^\star))}(g(x,\tilde{a}) - g(x,a^\star)) \leq 1/\lambda$.

Second case: when $a^\star = \tilde{a}$, then $p(a^\star) = p(\tilde{a}) = 1 - \sum_{a\neq\tilde{a}} \frac{1}{A+\lambda(g(x,\tilde{a})-g(x,a))} \geq 1/A$. Then,

$$(1 - p(a^\star))(g(x,a^\star) - g(x,\tilde{a})) + \frac{1}{4\lambda p(a^\star)} \leq \frac{A}{4\lambda}.$$

So combine all these terms together, we arrive that:

$$\mathbb{E}_{a\sim\rho}(f(x,a^\star) - f(x,a)) \leq \lambda \mathbb{E}_{a\sim p}(g(x,a) - f(x,a))^2 + \frac{1}{4\lambda} + \frac{A-1}{\lambda} + \frac{1}{\lambda} + \frac{A}{4\lambda},$$

which implies that:

$$\mathbb{E}_{a\sim\rho}(f(x,a^\star) - f(x,a)) - \lambda \mathbb{E}_{a\sim p}(g(x,a) - f(x,a))^2 \leq \frac{4A}{\lambda}.$$

□

The above lemma shows that $\beta \leq 4A$. Plug in this into the general theorem, we see that our algorithm which uses IGW gives a regret bound $O(\sqrt{TA(\mathrm{Reg}(T) + \ln(1/\delta))})$.

A Appendix

Here we show that the regret bound in (1) leads to (2)

The regret form in (1) and the realizability condition $f^\star \in \mathcal{F}$ implies that:

$$\sum_{t=0}^{T-1} \left((f_t(x_t,a_t) - r_t)^2 - (f^\star(x_t,a_t) - r_t)^2\right) = \mathrm{Reg}(T).$$

Denote $z_t := (f_t(x_t, a_t) - r_t)^2 - (f^\star(x_t, a_t) - r_t)^2$. Note that f_t and p_t does not depend on a_t and r_t (i.e., a_t and r_t are generated given f_t and p_t). Denote $\mathbb{E}_t$ as the condition expectation which conditions on history $x_0, a_0, r_0, \ldots, x_{t-1}, a_{t-1}, r_{t-1}, x_t$ (so conditioned on this history, the only randomness here is from $a_t \sim p_t$ and $r_t \sim R(x_t, a_t)$).

$$\begin{aligned}
\mathbb{E}_t[z_t] &= \mathbb{E}_t\left[(f_t(x_t, a_t) - f^\star(x_t, a_t))(f_t(x_t, a_t) + f^\star(x_t, a_t) - 2r_t)\right] \\
&= \mathbb{E}_t\left[(f_t(x_t, a_t) - f^\star(x_t, a_t))(f_t(x_t, a_t) + f^\star(x_t, a_t) - 2f^\star(x_t, a_t))\right] \\
&= \mathbb{E}_t\left(f_t(x_t, a_t) - f^\star(x_t, a_t)\right)^2 = \mathbb{E}_{a\sim p_t}\left(f_t(x_t, a) - f^\star(x_t, a)\right)^2
\end{aligned}$$

Also note that

$$\begin{aligned}
\mathbb{E}_t[z_t^2] &= \mathbb{E}_t\left[(f_t(x_t, a_t) - f^\star(x_t, a_t))^2(f_t(x_t, a_t) + f^\star(x_t, a_t) - 2r_t)^2\right] \\
&\leq 4\mathbb{E}_t[(f_t(x_t, a_t) - f^\star(x_t, a_t))^2] \\
&= 4\mathbb{E}_{a\sim p_t}\left[(f_t(x_t, a) - f^\star(x_t, a))^2\right].
\end{aligned}$$

The sequence $z_t - \mathbb{E}_t[z_t]$ forms a sequence of Martingale difference, which allows us to use Azuma-Bernstein's inequality, i.e., with probability at least $1 - \delta$, we have:

$$\sum_t (\mathbb{E}_t[z_t] - z_t) \leq \sqrt{8\sum_t \mathbb{E}_{a\sim p_t}(f_t(x_t, a) - f^\star(x_t, a))^2 \cdot \ln(1/\delta)} + 4\ln(1/\delta).$$

Now use the fact that $\sum_t z_t \leq \text{Reg}_T$, and $\sum_t \mathbb{E}_t[z_t] = \sum_t \mathbb{E}_{a\sim p_t}(f_t(x_t, a) - f^\star(x_t, a))^2$, we have:

$$\sum_t \mathbb{E}_t z_t \leq \sqrt{8\ln(1/\delta)\sum_t \mathbb{E}_t z_t} + 4\ln(1/\delta) + \text{Reg}(T).$$

Solve for $\sum_t \mathbb{E}_t z_t$, we arrive at:

$$\sum_t \mathbb{E}_t z_t \leq 2\text{Reg}(T) + 16\ln(1/\delta).$$

References