Lecture 13

Instructor: Haipeng Luo

1 Optimal Multi-armed Bandit Algorithms

We have shown a lower bound of order $\Omega(\sqrt{TK})$ for the expected regret of any algorithm for the multi-armed bandit problem, and also that Exp3 ensures an expected regret of order $\mathcal{O}(\sqrt{TK \ln K})$. Can we close the $\sqrt{\ln K}$ gap in the upper and lower bounds?

The answer turns out to be yes, and the approach is again FTRL, but with special regularizers. Specifically, consider the following general FTRL algorithm for multi-armed bandit [Audibert and Bubeck, 2010, Abernethy et al., 2015]: draw $a_t \sim p_t$ with

$$p_{t+1} = \operatorname*{argmin}_{p \in \Delta(K)} \left\langle p, \sum_{\tau=1}^{t} \widehat{\ell}_\tau \right\rangle + \frac{1}{\eta}\psi(p)$$

where $\psi(p)$ is a regularizer and $\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)}\mathbf{1}\{a = a_t\}$ is the importance weighted estimator. Exp3 is clearly just a special case with $\psi(p)$ being the negative entropy. To derive the optimal algorithm, we will consider a family of FTRL instances by using the following regularizer

$$\psi(p) = \frac{1 - \sum_{a=1}^{K} p(a)^\beta}{1-\beta},$$

for a parameter $\beta \in (0,1)$. This is known as the *Tsallis entropy* and is in fact a generalization of the Shannon entropy since $\lim_{\beta \to 1} \frac{1-\sum_a p(a)^\beta}{1-\beta} = \sum_a p(a)\ln(p(a))$ by L'Hôpital's rule. Therefore the algorithm above can be seen as a generalization of the Exp3 algorithm. One can now verify that the algorithm admits the following update rule

$$\frac{1}{p_{t+1}(a)^{1-\beta}} = \frac{1-\beta}{\beta}\left(\lambda + \eta \sum_{\tau=1}^{t} \widehat{\ell}_\tau(a)\right), \quad \forall a \in [K] \tag{1}$$

for some constant λ such that p_{t+1} is a distribution. This constant λ comes from the Lagrangian multiplier and can be found efficiently by a simple binary search.

While it is possible to use the general FTRL analysis to derive the regret bound for this algorithm, it is in fact simpler to analyze it using the Online Mirror Descent (OMD) framework (see Homework 1). Recall that one way to write the OMD algorithm is

$$\begin{aligned} \nabla\psi(p'_{t+1}) &= \nabla\psi(p_t) - \eta\widehat{\ell}_t \\ p_{t+1} &= \operatorname*{argmin}_{p \in \Delta(K)} D_\psi(p, p'_{t+1}) \end{aligned}$$

where $D_\psi(p,q) = \psi(p) - \psi(q) - \langle \nabla\psi(q), p-q\rangle$ is the Bregman divergence associated with ψ. With ψ being the Tsallis entropy, one can verify $\nabla\psi(q)(a) = \frac{-\beta}{1-\beta}\frac{1}{q(a)^{1-\beta}}$ and

$$D_\psi(p,q) = \frac{1}{1-\beta}\sum_{a=1}^{K}\left(q(a)^\beta - p(a)^\beta + \frac{\beta}{q(a)^{1-\beta}}(p(a)-q(a))\right)$$

and the update rule becomes

$$\frac{1}{p'_{t+1}(a)^{1-\beta}} = \frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta}\eta\widehat{\ell}_t(a) \tag{2}$$

$$\frac{1}{p_{t+1}(a)^{1-\beta}} = \frac{1}{p'_{t+1}(a)^{1-\beta}} + \lambda \tag{3}$$

where λ is again such that p_{t+1} is a distribution (different from the λ in Eq. (1) though) and can be computed by a binary search. This update rule is in fact equivalent to the FTRL update rule (1) since combining (2) and (3) iteratively leads to

$$\frac{1}{p_{t+1}(a)^{1-\beta}} = \frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta}\eta\widehat{\ell}_t(a) + \lambda = \cdots = \frac{1-\beta}{\beta}\eta\left(\sum_{\tau=1}^{t}\widehat{\ell}_\tau(a)\right) + \lambda'$$

for some other normalization term λ'. This shows that the two algorithms are exactly the same and we can use the OMD analysis to analyze the algorithm, which is the focus for the rest of the section.

Recall that the key in the proof of Exp3 is the following bound: $\forall a^\star \in [K]$,

$$\sum_{t=1}^{T}\left\langle p_t, \widehat{\ell}_t\right\rangle - \widehat{\ell}_t(a^\star) \le \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\widehat{\ell}_t(a)^2, \tag{4}$$

and the last term deals with the large variance issue of the estimator automatically. With ψ being the Tsallis entropy, we can prove a generalized version of the bound:

Theorem 1. *As long as $\widehat{\ell}_t(a) \ge 0$ for all t and a, FTRL* (1) *or OMD* (2) (3) *ensures $\forall a^\star \in [K]$,*

$$\sum_{t=1}^{T}\left\langle p_t, \widehat{\ell}_t\right\rangle - \widehat{\ell}_t(a^\star) \le \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta}{\beta}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^{2-\beta}\widehat{\ell}_t(a)^2. \tag{5}$$

Note that the theorem does not require $\widehat{\ell}_t(a)$ to be the specific importance weighted estimator. By L'Hôpital's rule, we have $\lim_{\beta\to 1}\frac{K^{1-\beta}-1}{(1-\beta)} = \ln K$ and thus the bound above exactly recovers Eq. (4). However, the bound is actually slightly better and allows one to obtain the optimal regret as shown by the following corollary.

Corollary 1. *With $\widehat{\ell}_t$ being the importance weighted estimator, FTRL* (1) *or OMD* (2) (3) *ensures*

$$\mathbb{E}[\mathcal{R}_T] \le \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta K^\beta T}{\beta}.$$

Therefore, by picking $\beta = 1/2$ and $\eta = 1/\sqrt{T}$, we obtain the optimal regret $\mathbb{E}[\mathcal{R}_T] = 4\sqrt{TK}$.

Proof. Recall that the conditional second moment of the estimator $\mathbb{E}_t[\widehat{\ell}_t(a)^2]$ is bounded by $1/p_t(a)$. Therefore, by taking expectation on both sides of Eq. (5), we arrive at

$$\mathbb{E}[\mathcal{R}_T] \le \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta}{\beta}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^{1-\beta}.$$

Applying Hölder's inequality to the last term

$$\sum_{a=1}^{K} p_t(a)^{1-\beta} \le \left(\sum_{a=1}^{K}(p_t(a)^{1-\beta})^{\frac{1}{1-\beta}}\right)^{1-\beta}\left(\sum_{a=1}^{K}1^{\frac{1}{\beta}}\right)^{\beta} \le K^\beta$$

finishes the proof. □

Clearly picking other constants such $\beta = 1/3$ (along with the optimal η) can also lead to a bound of the optimal order $\mathcal{O}(\sqrt{TK})$. It remains to prove Theorem 1.

Proof of Theorem 1. According to the OMD analysis, for any $q \in \Delta(K)$ we have

$$\begin{aligned}\eta \left\langle p_t - q, \widehat{\ell}_t \right\rangle &= D_\psi(q, p_t) - D_\psi(q, p'_{t+1}) + D_\psi(p_t, p'_{t+1}) \\ &\le D_\psi(q, p_t) - D_\psi(q, p_{t+1}) + D_\psi(p_t, p'_{t+1}),\end{aligned}$$

and thus

$$\sum_{t=1}^{T} \left\langle p_t - q, \widehat{\ell}_t \right\rangle \le \frac{D_\psi(q, p_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{T} D_\psi(p_t, p'_{t+1}).$$

When q concentrates on one particular action, $D_\psi(q, p_1) = \frac{K^{1-\beta}-1}{(1-\beta)}$. Therefore, it remains to prove

$$D_\psi(p_t, p'_{t+1}) \le \frac{\eta^2}{\beta} \sum_{a=1}^{K} p_t(a)^{2-\beta} \widehat{\ell}_t(a)^2. \tag{6}$$

Indeed, by definition we have

$$\begin{aligned}D_\psi(p_t, p'_{t+1}) &= \frac{1}{1-\beta} \sum_{a=1}^{K} \left(p'_{t+1}(a)^\beta - p_t(a)^\beta + \frac{\beta}{p'_{t+1}(a)^{1-\beta}} (p_t(a) - p'_{t+1}(a)) \right) \\ &= \frac{1}{1-\beta} \sum_{a=1}^{K} \left((1-\beta) p'_{t+1}(a)^\beta - p_t(a)^\beta + \beta \left(\frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta} \eta \widehat{\ell}_t(a) \right) p_t(a) \right) \\ &= \sum_{a=1}^{K} \left(p'_{t+1}(a)^\beta - p_t(a)^\beta + \eta p_t(a) \widehat{\ell}_t(a) \right). \end{aligned} \tag{7}$$

Now notice that

$$p'_{t+1}(a)^\beta = p_t(a)^\beta \left(\frac{p'_{t+1}(a)^{\beta-1}}{p_t(a)^{\beta-1}} \right)^{\frac{\beta}{\beta-1}} = p_t(a)^\beta \left(1 + \frac{1-\beta}{\beta} \eta p_t(a)^{1-\beta} \widehat{\ell}_t(a) \right)^{\frac{\beta}{\beta-1}},$$

and thus using the fact $(1+x)^\alpha \le 1 + \alpha x + \alpha(\alpha-1)x^2$ for any $x \ge 0$ and $\alpha < 0$,[1] we have

$$\begin{aligned}p'_{t+1}(a)^\beta &\le p_t(a)^\beta \left(1 - \eta p_t(a)^{1-\beta} \widehat{\ell}_t(a) + \frac{\eta^2}{\beta} p_t(a)^{2-2\beta} \widehat{\ell}_t(a)^2 \right) \\ &= p_t(a)^\beta - \eta p_t(a) \widehat{\ell}_t(a) + \frac{\eta^2}{\beta} p_t(a)^{2-\beta} \widehat{\ell}_t(a)^2.\end{aligned}$$

Plugging this into Eq. (7) proves Eq. (6) and thus the theorem. □

2 High Probability Bounds

So far we have only proven that the expected regret of Exp3 or the more general algorithm is nicely bounded. However, since online learning focuses more on sequentially playing the game without going back, it seems that the *expected* regret does not really say much about the performance of the algorithm for a particular run. To address this issue, we need to derive a bound on the actual regret that holds with high probability.

Due to the high variance of the importance weighted estimator, without any modification the approaches we have discussed cannot ensure the same regret bound with high probability. There are many fixes for this, but they all share the same idea of sacrificing a little bit of unbiasedness to lower the variance. Here, we discuss a simple strategy introduced in [Neu, 2015], which constructs loss estimators as

$$\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a) + \gamma} \mathbf{1}\{a = a_t\}, \quad \forall a \in [K] \tag{8}$$

[1] This is because with $y = \ln(1+x)$, one has $(1+x)^\alpha = e^{\alpha y} \le 1 + \alpha y + \alpha^2 y^2$ due to $\alpha y < 0$. Further using inequalities $y = \ln(1+x) \ge x - x^2$ and $y = \ln(1+x) \le x$ proves the fact.

for some parameter $\gamma > 0$. The rest of the algorithm remains exactly the same: plug this new estimator into the update rule of Exp3 or FTRL/OMD with Tsallis entropy to obtain p_t and then sample $a_t \sim p_t$.

The new estimator makes a difference mostly when $p_t(a_t)$ is small – in this case the extra term γ makes the estimator much less dramatic. In general the estimator is underestimating the losses and the following important property holds (the proof can be found in [Neu, 2015] and is omitted here).

Lemma 1. *Let $c_1, \ldots, c_T \in [0, 2\gamma]^K$ be such that $c_t(a)$ is fixed given everything up to the beginning of time t. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\sum_{t=1}^{T}\sum_{a=1}^{K} c_t(a)\left(\widehat{\ell}_t(a) - \ell_t(a)\right) \le \ln\left(1/\delta\right).$$

Note that before when $\widehat{\ell}_t$ was unbiased, such inequality would not hold because the (large) variance plays a role in the martingale concentration bound. The key is that the new estimator is now an underestimation, making such one-sided inequality possible. Moreover, such one-sided inequality turns out to be all we need to prove a high probability bound.

Theorem 2. *With $\widehat{\ell}_t$ defined as in Eq.* (8)*, FTRL* (1) *or OMD* (2) (3) *ensures that for a fixed $a^\star \in [K]$, we have with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \ell_t(a) - \ell_t(a^\star) \le \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta K^\beta T}{\beta} + \gamma T K + \frac{1}{2}\left(\frac{\eta}{\beta\gamma} + \frac{1}{\gamma} + 1\right)\ln\left(\frac{3}{\delta}\right).$$

Picking $\beta = 1/2$, $\eta = 1/\sqrt{T}$ and $\gamma = \sqrt{\frac{\ln(1/\delta)}{TK}}$ leads to $\mathcal{R}_T = \mathcal{O}(\sqrt{TK\ln(1/\delta)} + \ln(1/\delta))$.

Proof. Note that

$$\left\langle p_t, \widehat{\ell}_t \right\rangle = p_t(a_t)\frac{\ell_t(a_t)}{p_t(a_t)+\gamma} = \ell_t(a_t) - \gamma\frac{\ell_t(a_t)}{p_t(a_t)+\gamma} = \ell_t(a_t) - \gamma\sum_{a=1}^{K}\widehat{\ell}_t(a).$$

Therefore, by applying Theorem 1 which holds here since $\widehat{\ell}_t(a) \ge 0$, we have

$$\begin{aligned}
\sum_{t=1}^{T}\ell_t(a_t) &\le \sum_{t=1}^{T}\left(\left\langle p_t, \widehat{\ell}_t\right\rangle + \gamma\sum_{a=1}^{K}\widehat{\ell}_t(a)\right) \\
&\le \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \sum_{t=1}^{T}\left(\widehat{\ell}_t(a^\star) + \frac{\eta}{\beta}\sum_{a=1}^{K} p_t(a)^{2-\beta}\widehat{\ell}_t(a)^2 + \gamma\sum_{a=1}^{K}\widehat{\ell}_t(a)\right) \\
&\le \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \sum_{t=1}^{T}\left(\widehat{\ell}_t(a^\star) + \frac{\eta}{\beta}\sum_{a=1}^{K} p_t(a)^{1-\beta}\widehat{\ell}_t(a) + \gamma\sum_{a=1}^{K}\widehat{\ell}_t(a)\right)
\end{aligned}$$

where the last step is due to $p_t(a)\widehat{\ell}_t(a) \le 1$. We can now apply Lemma 1 to the last three terms with $c_t(a) \le 2\gamma$ being $2\gamma\mathbf{1}\{a = a^\star\}$, $2\gamma p_t(a)^{1-\beta}$, and 2γ respectively and a union bound to conclude that with probability at least $1-\delta$, the last three terms are bounded by

$$\sum_{t=1}^{T}\left(\ell_t(a^\star) + \frac{\eta}{\beta}\sum_{a=1}^{K} p_t(a)^{1-\beta}\ell_t(a) + \gamma\sum_{a=1}^{K}\ell_t(a)\right) + \frac{1}{2}\left(\frac{\eta}{\beta\gamma} + \frac{1}{\gamma} + 1\right)\ln\left(\frac{3}{\delta}\right).$$

Rearranging, using $\ell_t(a) \le 1$, and applying Hölder's inequality as in the proof of Corollary 1 finish the proof. □

References

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28*, 2015.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems 28*, 2015.