CSCI 659 Lecture 6

Fall 2022

Instructor: Haipeng Luo

1 The Adversarial Multi-Armed Bandit Problem

All the topics we have discussed so far consider problems with full information feedback. Starting from this lecture, we will move on to the more challenging settings with partial information feedback. The classical example of such problems is the *Multi-Armed Bandit* (MAB) problem [Lai and Robbins, 1985], and in this lecture we start with the adversarial version of MAB introduced in [Auer et al., 2002], which can be seen as a variant of the expert problem.

More specifically, the problem models the situation where a gambler sequentially pulls the arm of one of the slot machines in a casino, with the hope of maximizing reward. A slot machine is sometimes called a "one-armed bandit", and hence the name multi-armed bandit for this problem. Formally, there are K arms/actions available for a learner, and at each time $t = 1, \ldots, T$,

1. the learner picks an action $a_t \in [K]$ while simultaneously the environment decides the loss vector $\ell_t \in [0,1]^K$,
2. the learner then suffers and observes (only) the loss $\ell_t(a_t)$.

Clearly, this is simply a partial information version of the expert problem, with the difference being that the learner has to actually pick one action at each round and then observes only the loss for this action but not the whole loss vector ℓ_t. For convention, we move from the notation i and N to a and K to denote a specific action and the total number of actions respectively.

For simplicity, we focus on an oblivious adversary and only point out that dealing with an adaptive adversary only requires small modifications. An equivalent way to think about the oblivious case is that the adversary, knowing the learner's algorithm, decides the sequence of loss vector $\ell_1, \ldots, \ell_T$ ahead of time (possibly in a randomized way). For any such a loss sequence, we measure the algorithm's performance by the expected regret

$$\mathbb{E}\left[\mathcal{R}_T\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a),$$

where the expectation is with respect to the randomness of the algorithm.

While simple, MAB captures the essence of many real-life applications such as clinical trials and recommendation systems, where arms correspond to available choices of medicines or products. MAB is also the foundational model to study the well-known *exploitation versus exploration* trade-off, one of the key challenges for all problems with partial information feedback. Indeed, on the one hand, it is tempting to select arms that have suffered small losses in the past (exploitation), but on the other hand, there is also an incentive to select other actions just to find out whether they can lead to even smaller losses (exploration). These two incentives are in conflict when one can only select one action at each round, and thus having a good balance between them is the key to design good algorithms.

1.1 Loss Estimators and Exp3

Since the only difference between MAB and the expert problem is the incomplete information about the loss vector ℓ_t, a natural idea to solve MAB is to construct a loss estimator $\widehat{\ell}_t$ and then feed it to any expert algorithm in a blackbox manner. While it might seem impossible to accurately estimate ℓ_t, a vector that can be arbitrarily different from the past loss vectors, by only seeing one of its coordinates, this is indeed doable in a sense via the help of randomness. More specifically, suppose that at time t we pick a_t randomly according to a distribution $p_t \in \Delta(K)$ with a full support, then after seeing $\ell_t(a_t)$, we can construct an *inverse importance weighted* estimator $\widehat{\ell}_t \in \mathbb{R}_+^K$ such that

$$\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)} \mathbf{1}\{a = a_t\} = \begin{cases} \frac{\ell_t(a_t)}{p_t(a_t)} & \text{if } a = a_t, \\ 0 & \text{else.} \end{cases}$$

While written in terms of $\ell_t(a)$ for each a, the indicator $\mathbf{1}\{a = a_t\}$ makes sure that we indeed only use the information $\ell_t(a)$ when we pick this arm and observe its loss. In other words, the estimator is well defined and computable using the information available to the learner.

Unbiasedness The reason that inverse importance weighting makes sense is because it leads to an unbiased estimator.

Lemma 1. *Let $\mathbb{E}_t[\cdot]$ denote the conditional expectation with respect to the random draw of a_t given the past. Then we have for any $a \in [K]$, $\mathbb{E}_t[\widehat{\ell}_t(a)] = \ell_t(a)$.*

Proof. This is by direct calculation: $\mathbb{E}_t[\widehat{\ell}_t(a)] = (1 - p_t(a)) \times 0 + p_t(a) \frac{\ell_t(a)}{p_t(a)} = \ell_t(a)$. □

As mentioned, with such an unbiased estimator, it is natural to feed it to any expert algorithm. Indeed, the very first MAB algorithm is obtained by feeding this estimator to Hedge, resulting to the following Exp3 algorithm [Auer et al., 2002] (short for Exponential-weight for Exploration and Exploitation): at time t, sample $a_t \sim p_t \in \Delta(K)$ where

$$\forall a \in [K], \quad p_t(a) \propto \exp\left(-\eta \sum_{s<t} \widehat{\ell}_s(a)\right) \tag{Exp3}$$

for some learning rate $\eta > 0$.

Before analyzing the regret of Exp3, we first address the following question: where is the aforementioned exploration versus exploitation trade-off in this algorithm? The exploitation part is basically executed by the Hedge algorithm: arms with smaller estimated losses are selected with higher probability. On the other hand, the exploration part is somewhat implicit. Indeed, whenever an arm a_t is selected (maybe due to exploitation), the probability of selecting this arm next time is always decreased (or at least not increased), which will then encourage the algorithm to explore other actions. This is due to the structure of the estimator $\widehat{\ell}_t$ so that only the selected action a_t can have non-zero loss, while all the other actions have zero estimated loss.

To better understand the importance of this implicit exploration, consider the case where the losses are negative: $\ell_t \in [-1, 0]^K$ (or equivalently their magnitude corresponds to reward). One can verify that this does not make any difference to Hedge for the expert problem. However, for MAB, Exp3 does not work anymore, since whenever an arm a_t is selected, its probability of being selected next time gets even larger (again due to the structure of $\widehat{\ell}_t$). This scheme clearly lacks sufficient exploration and will suffer linear regret in the worst case. Therefore, when using Exp3, it is important that losses are shifted so that they are all nonnegative.

1.2 Variance and Variance Cancellation

Given that the estimators are unbiased, does it mean that the expected regret of Exp3 is simply $\mathcal{O}(\sqrt{T \ln K})$, the same as Hedge for the expert problem? The answer is no — the $\mathcal{O}(\sqrt{T \ln K})$ bound only holds when the losses fed to Hedge is in $[0, 1]$, but in Exp3, the estimated losses fed to Hedge can be huge due to the inverse importance weighting (for example, $\widehat{\ell}_t(a)$ can be as large as

$1/p_t(a)$). One may try to fix this by enforcing a lower bound on $p_t(a)$ such that the estimators are not too large, but such methods would lead to $\omega(\sqrt{T})$ regret (you should try it yourself).

One may also wonder why it makes sense to have such large estimated losses given that we know the true losses are always in $[0,1]$. However, this is intuitively unavoidable since we are estimating an arbitrary vector by only seeing one of its coordinates. In fact, the importance weighted estimators not only have a large magnitude, but also have a large variance (or rather a large second moment):

Lemma 2. *Let $\mathbb{E}_t[\cdot]$ denote the conditional expectation with respect to the random draw of a_t given the past. Then we have for any $a \in [K]$, $\mathbb{E}_t[\widehat{\ell}_t(a)^2] = \frac{\ell_t(a)^2}{p_t(a)} \leq \frac{1}{p_t(a)}$.*

Proof. This is also by direct calculation: $\mathbb{E}_t[\widehat{\ell}_t(a)] = (1 - p_t(a)) \times 0 + p_t(a)\frac{\ell_t(a)^2}{p_t(a)^2} = \frac{\ell_t(a)^2}{p_t(a)}$. □

So how do we deal with such a large variance/magnitude? Somewhat surprisingly, there is nothing we need to do algorithmically once we realize that Exp3 itself enjoys a certain intrinsic *variance cancellation* effect, highlighted in the proof below.

Theorem 1. *With $\eta = \sqrt{\frac{\ln K}{TK}}$, Exp3 ensures $\mathbb{E}[\mathcal{R}_T] = \mathcal{O}(\sqrt{TK \ln K})$.*

Proof. We apply the following regret bound of Hedge proven in Lecture 1 (see Eq. (1) therein): for any $a^\star \in [K]$,

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

which holds as long as $\widehat{\ell}_t(a) \geq 0$ (in fact, as long as $\eta\widehat{\ell}_t(a) \geq -1$). On the other hand, using Lemma 1 and Lemma 2, we have $\mathbb{E}_t[\widehat{\ell}_t(a^\star)] = \ell_t(a^\star)$,

$$\mathbb{E}_t[\langle p_t, \widehat{\ell}_t \rangle] = \langle p_t, \mathbb{E}_t[\widehat{\ell}_t] \rangle = \langle p_t, \ell_t \rangle = \mathbb{E}_t[\ell_t(a_t)],$$

and

$$\mathbb{E}_t \left[\sum_{a=1}^{K} p_t(a) \widehat{\ell}_t(a)^2 \right] = \sum_{a=1}^{K} p_t(a) \frac{\ell_t(a)^2}{p_t(a)} \leq K. \qquad \text{(variance cancellation)}$$

Therefore, taking expectation on both sides of Eq. (1) leads to $\mathbb{E}[\mathcal{R}_T] \leq \frac{\ln K}{\eta} + TK\eta$, and picking the optimal value of η finishes the proof. □

From this proof, it is clear that the key is in the specific bound $\sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) \widehat{\ell}_t(a)^2$ on the stability term of Hedge, which quite remarkably cancels the potentially large variance of the estimators automatically. Note that this bound does not necessarily hold if the true losses are negative, which is consistent with our earlier intuition that Exp3 with negative losses should not work due to the lack of exploration. Also note that we have derived other types of bound on the stability term for Hedge and its variant, including $\sum_{t=1}^{T} \|\widehat{\ell}_t\|_\infty^2$ (Section 3.2 of Lecture 2) and $\sum_{t=1}^{T} \widehat{\ell}_t(a^\star)^2$ when competing with $a^\star$ (Theorem 3 of Lecture 3), but none of these enjoys the same variance cancellation effect that is critical for MAB (verify this yourself).

Question 1. *Can you figure out what breaks in this proof when the adversary is adaptive?*

2 Lower Bounds

The regret bound of Exp3 shows that the price of having bandit feedback is an extra $\sqrt{K}$ factor compared to full information setting, which is quite intuitive since each round we only obtain $1/K$ fraction of information. So is this the optimal regret bound? To answer this question, we first show an $\Omega(\sqrt{TK})$ lower bound in this section.

The intuition of the lower bound is rather straightforward. For any fixed algorithm, first image running it in a simple world where losses for all arms are generated independently and uniformly from $\{0,1\}$. There must exist an arm that is selected no more than T/K times by this algorithm.

Now suppose that the adversary secretly modifies the environment so that the loss of this arm follows a Bernoulli distribution with parameter $1/2-\sqrt{K/T}$, which is not distinguishable from the uniform distribution with only T/K samples based on information theory. Thus, when run in this new environment, the same algorithm should not be aware of this change and will still pick this arm not often enough, say no more than $T/2$ rounds, leading to at least $\frac{T}{2}\sqrt{K/T} = \Omega(\sqrt{TK})$ regret.

The question is how to make this argument formal. In particular, how to formally argue that in the new environment the algorithm's behavior stays roughly the same. As we will see in the proof below, this can in fact be related to the KL divergence between two distributions corresponding to the two environments.

Theorem 2. *For any MAB algorithm $\mathcal{A}$, there exists a fixed sequence of loss vectors such that*

$$\mathbb{E}_{\mathcal{A}}[\mathcal{R}_T] = \Omega(\sqrt{TK})$$

where we use $\mathbb{E}_{\mathcal{A}}[\cdot]$ to denote the expectation with respect to the randomness of $\mathcal{A}$.

Proof. According to the informal argument mentioned earlier, we create two randomized environments $\mathcal{E}$ and $\mathcal{E}'$ in the following way (and use $\mathbb{E}$ and $\mathbb{E}'$ to denote the expectation in these two environments respectively). In $\mathcal{E}$, every loss $\ell_t(a)$ follows independently a Bernoulli distribution with parameter $1/2$, denoted by $\text{Ber}(1/2)$. There must exist $a' \in [K]$ such that $\mathbb{E}[n(a')] \le \frac{T}{K}$ where $n(a) = \sum_{t=1}^{T} \mathbf{1}\{a_t = a\}$ is the total number of times a is selected. Then $\mathcal{E}'$ is constructed such that the losses of arm a' follow $\text{Ber}(1/2-\epsilon)$ independently for some small $\epsilon \le 1/4$ to be specified later, and every other arm still follows $\text{Ber}(1/2)$ independently.

The rest of the proof argues that $\mathbb{E}'\mathbb{E}_{\mathcal{A}}[\mathcal{R}_T] = \Omega(\sqrt{TK})$, which implies that there exists a *fixed* sequence of loss vectors such that $\mathbb{E}_{\mathcal{A}}[\mathcal{R}_T] = \Omega(\sqrt{TK})$ and concludes the proof. Further note that $\mathbb{E}'\mathbb{E}_{\mathcal{A}}[\mathcal{R}_T] = \mathbb{E}_{\mathcal{A}}\mathbb{E}'[\mathcal{R}_T]$, so it is sufficient to prove that for any *deterministic* algorithm, $\mathbb{E}'[\mathcal{R}_T] = \Omega(\sqrt{TK})$. If we denote the observation of the learner at time t by $\widetilde{\ell}_t = \ell_t(a_t)$, then a deterministic algorithm selects a_t via some fixed function of $\widetilde{\ell}_{1:t-1}$, a shorthand for the sequence $\widetilde{\ell}_1, \ldots, \widetilde{\ell}_{t-1}$ (note that the information of $a_{1:t-1}$ is redundant since it is determined by $\widetilde{\ell}_{1:t-2}$ already).

Clearly, in expectation a' is the best arm in $\mathcal{E}'$ and

$$\mathbb{E}'[\mathcal{R}_T] = \mathbb{E}'\left[\sum_{t=1}^{T} \ell_t(a_t) - \min_{a\in[K]} \sum_{t=1}^{T} \ell_t(a)\right] \ge \mathbb{E}'\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(a')\right] = (T - \mathbb{E}'[n(a')])\epsilon.$$

We next show that $\mathbb{E}'[n(a')]$ and $\mathbb{E}[n(a')]$ are close, that is, the number of times a' is selected in environment $\mathcal{E}$ and that in environment $\mathcal{E}'$ are similar (just as in the previous informal argument). Indeed, using $\mathbb{P}$ and $\mathbb{P}'$ to denote the distributions of the observation sequence $\widetilde{\ell}_{1:T}$ in $\mathcal{E}$ and $\mathcal{E}'$ respectively, we have

$$\begin{aligned}
\mathbb{E}'[n(a')] - \mathbb{E}[n(a')] &= \sum_{\widetilde{\ell}_{1:T}\in\{0,1\}^T} n(a')\left(\mathbb{P}'(\widetilde{\ell}_{1:T}) - \mathbb{P}(\widetilde{\ell}_{1:T})\right) \\
&\le T \sum_{\widetilde{\ell}_{1:T}\in\{0,1\}^T} \left|\mathbb{P}'(\widetilde{\ell}_{1:T}) - \mathbb{P}(\widetilde{\ell}_{1:T})\right| = T\left\|\mathbb{P}' - \mathbb{P}\right\|_1 \le T\sqrt{2\text{KL}(\mathbb{P} \parallel \mathbb{P}')},
\end{aligned}$$

where the last step is by the Pinsker's inequality.[1] To calculate $\text{KL}(\mathbb{P} \parallel \mathbb{P}')$, we apply a handy divergence decomposition lemma (Lemma 3, included after this proof):

$$\begin{aligned}
\text{KL}(\mathbb{P} \parallel \mathbb{P}') &= \mathbb{E}[n(a')] \cdot \text{KL}(\text{Ber}(1/2) \parallel \text{Ber}(1/2-\epsilon)) \\
&= \frac{\mathbb{E}[n(a')]}{2}\left(\ln\frac{1/2}{1/2+\epsilon} + \ln\frac{1/2}{1/2-\epsilon}\right) \\
&= \frac{\mathbb{E}[n(a')]}{2}\ln\left(\frac{1}{1-4\epsilon^2}\right) \\
&\le 8\mathbb{E}[n(a')]\,\epsilon^2,
\end{aligned}$$

[1] We in fact have proven Pinsker's inequality in Lecture 2 when arguing the strong convexity of negative entropy; see if you can make the connection.

where in the last step we use the fact $\ln\left(\frac{1}{1-x}\right) \le 4x$ for any $x \le \frac{1}{2}$. Finally we have shown

$$\mathbb{E}'\left[n(a')\right] \le \mathbb{E}\left[n(a')\right] + 4T\epsilon\sqrt{\mathbb{E}\left[n(a')\right]} \le \frac{T}{K} + 4T\epsilon\sqrt{\frac{T}{K}}$$

(recall a' is selected such that $\mathbb{E}\left[n(a')\right] \le T/K$) and thus

$$\mathbb{E}'[\mathcal{R}_T] \ge T\left(1 - \frac{1}{K} - 4\epsilon\sqrt{\frac{T}{K}}\right)\epsilon \ge T\left(\frac{1}{2} - 4\epsilon\sqrt{\frac{T}{K}}\right)\epsilon$$

Setting $\epsilon = \frac{1}{16}\sqrt{\frac{K}{T}}$ (to maximize the lower bound above) shows $\mathbb{E}'[\mathcal{R}_T] = \Omega(\sqrt{TK})$, finishing the proof. □

The following divergence decomposition lemma is very powerful and is used extensively in proving lower bounds.

Lemma 3 (Divergence decomposition)**.** *Let $\mathcal{E}$ and $\mathcal{E}'$ be two stochastic MAB environments where for each $a \in [K]$ the losses of arm a are i.i.d. samples of $\mathcal{P}_a$ and $\mathcal{P}'_a$ respectively. Let $\widetilde{\ell}_t = \ell_t(a_t)$ be the observation of a deterministic learner at time t and $\mathbb{P}$ and $\mathbb{P}'$ be the distributions of $\widetilde{\ell}_{1:T}$ for environments $\mathcal{E}$ and $\mathcal{E}'$ respectively. Then we have*

$$\mathrm{KL}(\mathbb{P} \parallel \mathbb{P}') = \sum_{a=1}^{K} \mathbb{E}\left[n(a)\right] \mathrm{KL}(\mathcal{P}_a \parallel \mathcal{P}'_a).$$

where $n(a)$ is the total number of times a is selected in $\mathcal{E}$.

Proof. By definition and direct calculation we have (for conciseness, we omit the range in the subscript of a summation; for example, $\sum_{\widetilde{\ell}_{1:T} \in \{0,1\}^T}$ is simply written as $\sum_{\widetilde{\ell}_{1:T}}$)

$$\begin{aligned}
\mathrm{KL}(\mathbb{P} \parallel \mathbb{P}') &= \sum_{\widetilde{\ell}_{1:T}} \mathbb{P}(\widetilde{\ell}_{1:T}) \ln\left(\frac{\mathbb{P}(\widetilde{\ell}_{1:T})}{\mathbb{P}'(\widetilde{\ell}_{1:T})}\right) = \sum_{\widetilde{\ell}_{1:T}} \mathbb{P}(\widetilde{\ell}_{1:T}) \ln\left(\frac{\prod_{t=1}^{T} \mathbb{P}(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\prod_{t=1}^{T} \mathbb{P}'(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \sum_{t=1}^{T} \sum_{\widetilde{\ell}_{1:T}} \mathbb{P}(\widetilde{\ell}_{1:T}) \ln\left(\frac{\mathbb{P}(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\mathbb{P}'(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \sum_{t=1}^{T} \sum_{\widetilde{\ell}_{1:t}} \left(\sum_{\widetilde{\ell}_{t+1:T}} \mathbb{P}(\widetilde{\ell}_{t+1:T}|\widetilde{\ell}_{1:t})\right) \mathbb{P}(\widetilde{\ell}_{1:t}) \ln\left(\frac{\mathbb{P}(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\mathbb{P}'(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \sum_{t=1}^{T} \sum_{\widetilde{\ell}_{1:t}} \mathbb{P}(\widetilde{\ell}_{1:t}) \ln\left(\frac{\mathbb{P}(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\mathbb{P}'(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \sum_{a=1}^{K} \sum_{t=1}^{T} \sum_{\widetilde{\ell}_{1:t}: a_t = a} \mathbb{P}(\widetilde{\ell}_{1:t}) \ln\left(\frac{\mathbb{P}(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\mathbb{P}'(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \sum_{a=1}^{K} \sum_{t=1}^{T} \sum_{\widetilde{\ell}_{1:t-1}: a_t = a} \mathbb{P}(\widetilde{\ell}_{1:t-1}) \sum_{\widetilde{\ell}_t} \mathbb{P}(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1}) \ln\left(\frac{\mathbb{P}(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\mathbb{P}'(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \sum_{a=1}^{K} \sum_{t=1}^{T} \mathbb{P}(a_t = a) \mathrm{KL}(\mathcal{P}_a \parallel \mathcal{P}'_a) = \sum_{a=1}^{K} \mathbb{E}\left[n(a)\right] \mathrm{KL}(\mathcal{P}_a \parallel \mathcal{P}'_a),
\end{aligned}$$

which completes the proof. □

3 Minimax Optimal MAB Algorithms

Given the gap between the Exp3 regret upper bound $\mathcal{O}(\sqrt{TK \ln K})$ and the lower bound $\Omega(\sqrt{TK})$ from the last section, it is more than natural to ask whether this gap can be closed and which one is the exact minimax optimal bound. While a gap of $\sqrt{\ln K}$ might seem quite negligible especially given that we already pay for $\sqrt{K}$ in the bound, the effort in understanding and closing this gap in the literature turns out to be highly fruitful (we will see some examples in the future).

The importance of local norms. To see if we can derive a better algorithm to match the lower bound, we take a closer look at the stability term $\sum_{a=1}^K p_t(a)\widehat{\ell}_t(a)^2$ in regret bound (1), which as we showed is at most K in expectation. Recall that the $p_t(a)$ factor is critical in canceling the large variance, so wouldn't it be nice if we have even "more" $p_t(a)$ in this bound? For example, if the stability term is instead $\sum_{a=1}^K p_t(a)^{3/2}\widehat{\ell}_t(a)^2$, then after taking expectation, this is at most $\sum_{a=1}^K \sqrt{p_t(a)} \le \sqrt{K}$, better than the original bound K.

If such a bound on the stability term indeed exists, then intuitively we should pay more for the penalty term (too good to be true otherwise). To better understand this trade-off, first note that the term $\sum_{a=1}^K p_t(a)\widehat{\ell}_t(a)^2$ for Hedge can in fact be written as a norm of the loss vector: $\|\widehat{\ell}_t\|^2_{\nabla^{-2}\psi(p_t)}$, that is, the quadratic norm weighted by the inverse Hessian of ψ at p_t, where ψ is the negative entropy regularizer (verify it yourself). This is often called the *local norm* of the loss vector, since it depends on some local value of the Hessian, in contrast to the $\|\widehat{\ell}_t\|^2_\star$ bound we proved in Lecture 2 which is in terms of some fixed norm $\|\cdot\|_\star$ (and is not good enough for MAB as discussed). Assuming for a moment that such a local-norm bound $\|\widehat{\ell}_t\|^2_{\nabla^{-2}\psi(p_t)}$ on the stability term holds generally for FTRL with any regularizer, then we can reverse-engineer the correct regularizer ψ that leads to the aforementioned smaller bound $\sum_{a=1}^K p_t(a)^{3/2}\widehat{\ell}_t(a)^2$ — it should be $\psi(p) = -4\sum_{a=1}^K \sqrt{p(a)}$.

Now that we have figured out the regularizer, it is easy to find out how large the penalty term is: the range of this regularizer is $B_\psi = 4(\sqrt{K} - 1)$ and the penalty term is bounded by B_ψ/η as shown in Lecture 2. Combining everything, the overall regret would be $\mathcal{O}(\frac{\sqrt{K}}{\eta} + \eta T\sqrt{K})$, which, after picking the optimal η, is $\mathcal{O}(\sqrt{TK})$, exactly matching the lower bound!

It remains to understand whether such a local-norm bound always holds for FTRL. Let's start with some encouraging evidence. Recall the (stability+negative) term $\langle p_t - p_{t+1}, \widehat{\ell}_t\rangle - \frac{1}{\eta}D_\psi(p_{t+1}, p_t)$ proven in the general regret bound of Lemma 3 in Lecture 2. Also recall that by definition, the Bregman divergence $D_\psi(p_{t+1}, p_t)$ is equal to $\frac{1}{2}\|p_t - p_{t+1}\|^2_{\nabla^2\psi(\xi)}$ for some ξ between p_t and p_{t+1}. Therefore, applying Hölder's inequality, we have

$$\langle p_t - p_{t+1}, \widehat{\ell}_t\rangle - \frac{1}{\eta}D_\psi(p_{t+1}, p_t) \le \|p_t - p_{t+1}\|_{\nabla^2\psi(\xi)}\|\widehat{\ell}_t\|_{\nabla^{-2}\psi(\xi)} - \frac{1}{2\eta}\|p_t - p_{t+1}\|^2_{\nabla^2\psi(\xi)},$$

which is at most $\frac{\eta}{2}\|\widehat{\ell}_t\|^2_{\nabla^{-2}\psi(\xi)}$ by the fact $2xy \le x^2 + y^2$. This is very encouraging since it remains to connect $\|\widehat{\ell}_t\|^2_{\nabla^{-2}\psi(\xi)}$ and $\|\widehat{\ell}_t\|^2_{\nabla^{-2}\psi(p_t)}$.

Unfortunately, these two quantities can be quite different in general, especially when p_t and p_{t+1} are too far away from each other (which could happen if the loss estimator $\widehat{\ell}_t$ is too "large" in some sense). Connecting these two quantities and finally getting a local-norm bound on the stability term is always the central piece of designing and analyzing online learning algorithms with bandit feedback, and we will see many more examples beyond MAB in the future.

Tsallis Entropy. Fortunately, for MAB, such a local-norm bound does hold for a broad family of regularizers, including the one obtained earlier via reverse engineering. Specifically, consider the following family of regularizers called (negative) *Tsallis entropy*, parameterized by $\beta \in (0, 1)$:

$$\psi(p) = \frac{1 - \sum_{a=1}^K p(a)^\beta}{1 - \beta}. \tag{2}$$

When $\beta = 1/2$, this essentially recovers the previously reverse-engineered regularizer (up to some constants that do not affect FTRL). Moreover, when β approaches 1, applying L'Hôpital's rule we

have $\psi(p) \to \sum_a p(a) \ln p(a)$, exactly recovering the negative Shannon entropy regularizer. Using such a regularizer in FTRL is first proposed by [Audibert and Bubeck, 2010] and later simplified by [Abernethy et al., 2015]. Below, we prove that this algorithm indeed enjoys the desired local-norm bound.

Theorem 3. *Consider the following FTRL algorithm*

$$p_t = \operatorname*{argmin}_{p \in \Delta(K)} \left\langle p, \sum_{s<t} \widehat{\ell}_s \right\rangle + \frac{1}{\eta}\psi(p) \tag{3}$$

where $\eta > 0$ is a learning rate, ψ is the Tsallis entropy defined in Eq. (2) with a parameter $\beta \in (0,1)$, and $\widehat{\ell}_1, \ldots, \widehat{\ell}_T \in \mathbb{R}_+^K$ are arbitrary loss vectors. Then the following holds for any $a^\star \in [K]$,

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

$$= \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta}{2\beta} \sum_{t=1}^T \sum_{a=1}^K p_t(a)^{2-\beta} \widehat{\ell}_t(a)^2. \tag{5}$$

Proof. Eq. (5) is by direct calculations: $\psi(p)$ is maximized when p concentrates on one action, and minimized when p is uniform, leading to $B_\psi = \max_p \psi(p) - \min_p \psi(p) = \frac{K^{1-\beta}-1}{1-\beta}$; the Hessian of ψ is a diagonal matrix with the a-th diagonal entry being $\beta p(a)^{\beta-2}$.

To prove Eq. (4), as mentioned, with Lemma 3 of Lecture 2 it suffices to show

$$\langle p_t - p_{t+1}, \widehat{\ell}_t \rangle - \tfrac{1}{\eta} D_\psi(p_{t+1}, p_t) \le \tfrac{\eta}{2} \|\widehat{\ell}_t\|^2_{\nabla^{-2}\psi(p_t)}.$$

To this end, we extend the definition of Tsallis entropy to the entire nonnegative orthant $\mathbb{R}_+^K$ following the same formula (2), and first bound $\langle p_t - p_{t+1}, \widehat{\ell}_t \rangle - \frac{1}{\eta} D_\psi(p_{t+1}, p_t)$ by

$$\max_{q \in \mathbb{R}_+^K} \langle p_t - q, \widehat{\ell}_t \rangle - \tfrac{1}{\eta} D_\psi(q, p_t).$$

Let q_t a maximizer of the above (which exists as we will show). Then, we repeat the analysis in our earlier discussion:

$$\begin{aligned}
&\langle p_t - q_t, \widehat{\ell}_t \rangle - \tfrac{1}{\eta} D_\psi(q_t, p_t) \\
&= \langle p_t - q_t, \widehat{\ell}_t \rangle - \tfrac{1}{2\eta} \|q_t - p_t\|^2_{\nabla\psi^2(\xi)} && \text{(for some } \xi \text{ between } p_t \text{ and } q_t\text{)} \\
&\le \|p_t - q_t\|_{\nabla\psi^2(\xi)} \|\widehat{\ell}_t\|_{\nabla\psi^{-2}(\xi)} - \tfrac{1}{2\eta} \|q_t - p_t\|^2_{\nabla\psi^2(\xi)} && \text{(Hölder's inequality)} \\
&\le \tfrac{\eta}{2} \|\widehat{\ell}_t\|^2_{\nabla\psi^{-2}(\xi)} && (2xy \le x^2 + y^2) \\
&\le \tfrac{\eta}{2} \|\widehat{\ell}_t\|^2_{\nabla\psi^{-2}(p_t)},
\end{aligned}$$

where the last step uses the monotonicity of the local norm and the fact $\xi(a) \le p_t(a)$ for all a. This is true because by the definition of q_t, we have by setting the gradient to zero:

$$\nabla\psi(q_t) = \nabla\psi(p_t) - \eta\widehat{\ell}_t,$$

which is equivalent to

$$\frac{1}{q_t(a)^{1-\beta}} = \frac{1}{p_t(a)^{1-\beta}} + \frac{1-\beta}{\beta}\eta\widehat{\ell}_t(a).$$

Since $\widehat{\ell}_t(a) \ge 0$, we must have $q_t(a) \le p_t(a)$ for all a, and being a point between p_t and q_t, ξ must also satisfy $\xi(a) \le p_t(a)$ for all a. This finishes the proof for Eq. (4). □

By L'Hôpital's rule, we have $\lim_{\beta \to 1} \frac{K^{1-\beta}-1}{1-\beta} = \ln K$ and thus Eq. (5) generalizes our analysis Eq. (1) for Hedge. One might notice that Theorem 2 holds for any nonnegative losses, while Eq. (1) only requires a slightly weaker condition $\eta\widehat{\ell}_t(a) \ge -1$. This can be addressed by a more careful analysis, which you will need to do in HW3.

With the help of this local-norm bound, we immediately obtain the following results for MAB.

Corollary 1. *Consider the following MAB algorithm: at time t sample a_t from p_t defined in Eq. (3) with $\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)}\mathbf{1}\{a = a_t\}$ being the inverse importance weighted estimator. Then we have*

$$\mathbb{E}[\mathcal{R}_T] \le \frac{K^{1-\beta} - 1}{\eta(1-\beta)} + \frac{\eta K^\beta T}{2\beta}.$$

Therefore, by picking $\beta = 1/2$ and $\eta = 1/\sqrt{T}$, we obtain the minimax optimal regret $\mathcal{O}(\sqrt{TK})$.

Proof. We directly take expectation on both sides of Eq. (5) and apply Lemma 2, arriving at

$$\mathbb{E}[\mathcal{R}_T] \le \frac{K^{1-\beta} - 1}{\eta(1-\beta)} + \frac{\eta}{2\beta}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)^{1-\beta}.$$

Applying Hölder's inequality to the last term

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

finishes the proof (or you can just argue that the LHS above is maximized when p is uniform). □

Clearly, picking other constants such $\beta = 1/3$ (along with the optimal η) also leads to the same minimax optimal bound $\mathcal{O}(\sqrt{TK})$. Compared to Hedge, whose penalty-stability trade-off is $\frac{\ln K}{\eta}$ versus ηTK, the trade-off here for any constant β is essentially $\frac{K^{1-\beta}}{\eta}$ versus $\eta T K^\beta$, leading a slight improvement in their product. Beside this improvement, in the next lecture we will also discuss one surprising application of this algorithm.

Question 2. *On the other hand, is Tsallis entropy a good regularizer for the expert problem with full information?*

Finally, note that this algorithm does not admit a closed-form update. Nevertheless, by writing down the Lagrangian and setting the gradient to zero, one can find that p_t satisfies

$$\frac{1}{p_t(a)^{1-\beta}} = \frac{1-\beta}{\beta}\left(\lambda + \eta\sum_{s<t}\widehat{\ell}_s(a)\right), \quad \forall a \in [K] \tag{6}$$

for some constant λ (the Lagrangian multiplier) such that p_t is a distribution, which can be found efficiently using a simple binary search.

4 Comparisons between Full-Information and Bandit Feedback

We conclude by highlighting some connections and differences between online learning problems with full information and those with bandit feedback. First, even with the more challenging bandit feedback, many problems still admit $\mathcal{O}(\sqrt{T})$ regret bound, though with extra dependence on other parameters characterizing the price of having less information. Almost all such $\mathcal{O}(\sqrt{T})$ bounds are obtained by feeding a full-information counterpart algorithm with some loss estimators and a careful analysis that makes use of certain local norm to handle the potentially large variance of the loss estimators. Thus, the critical part in the algorithm design is to find the "right" combination of the loss estimator and the regularizer to enable such variance cancellation effect. Most of the adaptive bounds we discussed in Lecture 3 (e.g. small-loss bounds, path-length bounds, and almost constant regret for stochastic losses) also have their counterparts for the bandit setting.

Despite theses similarities, there are also important distinctions between these two settings. For example, in Lecture 5 we derived strongly adaptive OCO algorithms with $\mathcal{O}(\sqrt{|\mathcal{I}|})$ regret for all interval $\mathcal{I}$ simultaneously. It turns out that the same is *impossible* for bandit problems (see HW3). As for achieving $\mathcal{O}(\sqrt{ST})$ switching regret for all S when competing against a sequence with S switches (which is possible for the full-information setting), it has also been proven impossible lately for MAB with an adaptive adversary and more than two arms [Marinov and Zimmert, 2021]. Interestingly, the case with exactly two arms or an oblivious adversary remains open.

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.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Teodor Vanislavov Marinov and Julian Zimmert. The pareto frontier of model selection for general contextual bandits. *Advances in Neural Information Processing Systems*, 34:17956–17967, 2021.