Improved Regret Bounds for Bandits with Expert Advice

NICOLÒ CESA-BIANCHI, Università degli Studi di Milano, Italy and Politecnico di Milano, Italy
KHALED ELDOWA, Università degli Studi di Milano, Italy and Politecnico di Milano, Italy
EMMANUEL ESPOSITO, Università degli Studi di Milano, Italy
JULIA OLKHOVSKAYA, TU Delft, Netherlands

In this research note, we revisit the bandits with expert advice problem. Under a restricted feedback model, we prove a lower bound of order $\sqrt{KT\ln(N/K)}$ for the worst-case regret, where K is the number of actions, $N > K$ the number of experts, and T the time horizon. This matches a previously known upper bound of the same order and improves upon the best available lower bound of $\sqrt{KT(\ln N)/(\ln K)}$. For the standard feedback model, we prove a new instance-based upper bound that depends on the agreement between the experts and provides a logarithmic improvement compared to prior results.

JAIR Track: Research Note

JAIR Associate Editor: Sriraam Natarajan

JAIR Reference Format:
Nicolò Cesa-Bianchi, Khaled Eldowa, Emmanuel Esposito, and Julia Olkhovskaya. 2025. Improved Regret Bounds for Bandits with Expert Advice. *Journal of Artificial Intelligence Research* 83, Article 6 (July 2025), 15 pages. DOI: 10.1613/jair.1.16738

1 Introduction

The problem of bandits with expert advice provides a simple and general framework for incorporating contextual information into the non-stochastic multi-armed bandit problem. In this framework, the learner receives in every round a recommendation, in the form of a probability distribution over the actions, from each expert in a given set. This set of experts can be seen as a set of strategies each mapping an unobserved context to a (randomized) action choice. The goal of the learner is to minimize their expected regret with respect to the best expert in hindsight; that is, the difference between their expected cumulative loss and that of the best expert. This problem was formulated by Auer et al. [9], who proposed the EXP4 algorithm as a solution strategy that has since become an important baseline or building block for addressing many related problems; for example, sleeping bandits [17], online multi-class classification [12], online non-parametric learning [10], and non-stationary bandits [19]. Auer et al. [9] proved a bound of order $\sqrt{KT\ln N}$ on the expected regret incurred by the EXP4 strategy, where T denotes the number of rounds, K the number of actions, and N the number of experts. This result is of a worst-case nature, in that it holds for any sequence of losses assigned to the actions and any sequence of expert recommendations.

The appealing feature of this bound is that it exhibits only a logarithmic dependence on the number of experts, in addition to the $\sqrt{K}$ dependence on the number of actions known to be unavoidable in the classical bandit

Authors' Contact Information: Nicolò Cesa-Bianchi, ORCID: 0000-0001-8477-4748, nicolo.cesa-bianchi@unimi.it, Università degli Studi di Milano, Milan, Italy and Politecnico di Milano, Milan, Italy; Khaled Eldowa, ORCID: 0000-0001-9362-2252, khaled.eldowa3@gmail.com, Università degli Studi di Milano, Milan, Italy and Politecnico di Milano, Milan, Italy; Emmanuel Esposito, ORCID: 0000-0002-4748-3279, emmanuel@emmanuelesposito.it, Università degli Studi di Milano, Milan, Italy; Julia Olkhovskaya, ORCID: 0009-0002-8337-726X, julia.olkhovskaya@gmail.com, TU Delft, Delft, Netherlands.

DOI: 10.1613/jair.1.16738

problem, where the learner competes with the best fixed action. While the minimax regret[1] in the latter problem has been shown to be of order $\sqrt{KT}$ [6], a similar exact characterization remains missing for the expert advice problem. Kale [16] studied a generalized version of the bandits with expert advice problem—originally proposed by Seldin et al. [23]—where the learner is only allowed to query the advice of $M \leq N$ experts. When $M = N$, the results of Kale [16] imply an upper bound of order $\sqrt{\min\{K,N\}T(1+\ln(N/\min\{K,N\}))}$ on the minimax regret, improving upon the bound of Auer et al. [9]. Unlike the latter, the logarithmic factor in the bound of Kale [16] diminishes as K increases with respect to N, leading to a bound of order $\sqrt{NT}$ when $N \leq K$, which is tight in general as the experts in that case can be made to emulate an N-armed bandit problem. This improved bound was achieved via the PolyINF algorithm [6, 7] played on the expert set utilizing the importance-weighted loss estimators of EXP4. Later, Seldin and Lugosi [24] proved a lower bound of order $\sqrt{KT(\ln N)/(\ln K)}$ for $N \geq K$.

As these upper and lower bounds still do not match, the correct minimax rate remains unclear. In this work, we take a step towards resolving this issue by showing that the upper bound is not improvable in general under a restricted feedback model in which the importance weighted loss estimators used by EXP4 or PolyINF remain implementable. In this restricted model, without observing the experts' recommendations, the learner picks an expert (possibly at random) at the beginning of each round, and the environment subsequently samples the action to be executed from the chosen expert's distribution. Afterwards, the learner only observes the distributions of the experts that had assigned positive probability to the chosen action. Via a reduction from the problem of multi-armed bandits with feedback graphs, we use the recent results of Chen et al. [11] to obtain a lower bound of order $\sqrt{KT\ln(N/K)}$ for $N > K$.

Departing from the worst-case results discussed thus far, a few works have obtained instance-dependent bounds for this problem. The dependence on the instance can be in terms of the assigned sequence of losses through small loss bounds [2], or in terms of the sequence of expert recommendations through bounds that reflect the similarity between the recommended expert distributions; e.g., see [21, 13] or [18, Theorem 18.3]. Our focus here is on the latter case, where to the best of our knowledge the state of the art is a bound of order $\sqrt{\sum_{t=1}^{T} C_t \ln N}$, shown in the recent work of Eldowa et al. [13] for the EXP4 algorithm. Here, C_t is the (chi-squared) capacity of the recommended distributions at round t. This quantity measures the dissimilarity between the experts' recommendations and satisfies $0 \leq C_t \leq \min\{K,N\} - 1$. Improving upon this result, we illustrate that it is possible to achieve a bound of order $\sqrt{\sum_{t=1}^{T} C_t\big(1+\ln(N/\max\{\overline{C}_T, 1\})\big)}$, where $\overline{C}_T = \sum_{t=1}^{T} C_t/T$ is the average capacity. This bound combines the best of the bound of Eldowa et al. [13] (its dependence on the agreement between the experts) and that of Kale [16] (its improved logarithmic factor), simultaneously outperforming both.

Road map. We formalize the problem setting in the next section. In Section 3, as a preliminary building block, we present Algorithm 1, an instance of the follow-the-regularized-leader (FTRL) algorithm with the (negative) q-Tsallis entropy as the regularizer. This algorithm is essentially equivalent to the PolyINF algorithm [8, 1], which was used by Kale [16] to achieve the best known worst-case upper bound. We then show in Section 4 that combining this algorithm with a doubling trick allows us to achieve the improved instance-based bound mentioned above. The lower bound for the restricted feedback setting is presented in Section 5. Finally, we provide some concluding remarks in Section 6.

2 Preliminaries

Notation. For a positive integer n, $[n]$ denotes the set $\{1,\ldots,n\}$. For $x, y \in \mathbb{R}$, let $x \vee y \coloneqq \max\{x,y\}$ and $x \wedge y \coloneqq \min\{x,y\}$. Moreover, we define $x_+ \coloneqq x \vee 0$.

[1] The best achievable worst-case regret guarantee.

Algorithm 1 q-FTRL for bandits with expert advice

```
**input:** $q \in (0, 1)$, $\eta > 0$
**initialization:** $p_1(i) \leftarrow 1/N$ for all $i \in V$
**for** $t = 1, \dots, T$ **do**
    receive expert advice $(\theta_t^i)_{i \in V}$
    draw expert $I_t \sim p_t$ and action $A_t \sim \theta_t^{I_t}$
    construct $\widehat{y}_t \in \mathbb{R}^N$ where $\widehat{y}_t(i) \coloneqq \frac{\theta_t^i(A_t)}{\sum_{j \in V} p_t(j)\theta_t^j(A_t)} \ell_t(A_t)$ for all $i \in V$
    let $p_{t+1} \leftarrow \arg\min_{p \in \Delta_N} \eta \langle \sum_{s=1}^t \widehat{y}_s, p \rangle + \psi_q(p)$
**end for**
```

Problem setting. Let $V = [N]$ be a set of N experts and $\mathcal{A} = [K]$ be a set of K actions. We consider a sequential decision-making problem where a learner interacts with an unknown environment for T rounds. The environment is characterized by a fixed and unknown sequence of loss vectors $(\ell_t)_{t \in [T]}$, where $\ell_t \in [0, 1]^K$ is the assignment of losses for the actions at round t, and a fixed and unknown sequence of expert advice $(\theta_t^i)_{i \in V, t \in [T]}$, where $\theta_t^i \in \Delta_K$ is the distribution over actions recommended by expert i at round t.[2] At the beginning of each round $t \in [T]$, the expert recommendations $(\theta_t^i)_{i \in V}$ are revealed to the learner, who then selects (possibly at random) an action $A_t \in \mathcal{A}$ and subsequently suffers and observes the loss $\ell_t(A_t)$. For an expert $i \in V$, we define $y_t(i) \coloneqq \sum_{a \in \mathcal{A}} \theta_t^i(a)\ell_t(a)$ as its loss in round t. The goal is to minimize the expected regret with respect to the best expert in hindsight:

$$R_T \coloneqq \mathbb{E}\left[\sum_{t=1}^T \ell_t(A_t)\right] - \min_{i \in V} \sum_{t=1}^T y_t(i)\,,$$

where the expectation is taken with respect to the randomization of the learner.

3 q-FTRL for Bandits with Expert Advice

The EXP4 algorithm can be seen as an instance of the FTRL framework (see, e.g., [22, Chapter 7]) where a distribution p_t over the experts is maintained at each round t and updated as follows

$$p_{t+1} \leftarrow \arg\min_{p \in \Delta_N} \eta \left\langle \sum_{s=1}^t \widehat{y}_s, p \right\rangle + \sum_{i \in V} p(i) \ln p(i)\,,$$

where $\eta > 0$ is the learning rate, the second term is the negative Shannon entropy of p, and $\widehat{y}_s(i)$ is an importance-weighted estimate of $y_s(i)$. The action A_t is drawn from the mixture distribution $\sum_{i \in V} p_t(i)\theta_t^i(\cdot)$. Consider a more general algorithm (outlined in Algorithm 1) where the negative Shannon entropy is replaced with the negative q-Tsallis entropy, which for $q \in (0, 1)$ is given by

$$\psi_q(x) \coloneqq \frac{1}{1-q}\left(1 - \sum_{i \in V} x(i)^q\right) \qquad \forall x \in \Delta_N\,.$$

In the limit when $q \to 1$, the negative Shannon entropy is recovered. The following theorem provides a regret bound for the algorithm. This result is not novel, a similar bound is implied by Theorem 2 in [16] for a closely related algorithm in a more general setting. We provide a concise proof of the result for completeness. As mentioned before, when $N \le K$, this bound is trivially tight in general. While, when $N > K$, we prove an order-wise matching minimax lower bound in Section 5 under additional restrictions on the received feedback.

[2] For a positive integer d, we let Δ_d denote the probability simplex in $\mathbb{R}^d$ defined as $\{u \in \mathbb{R}^d : \sum_{j=1}^d u(j) = 1 \text{ and } u(j) \ge 0\ \forall j \in [d]\}$.

Theorem 3.1. *Algorithm 1 run with*

$$q = \frac{1}{2}\left(1 + \frac{\ln\bigl(N/(K \wedge N)\bigr)}{\sqrt{\ln\bigl(N/(K \wedge N)\bigr)^2 + 4} + 2}\right) \in [1/2, 1) \quad \textit{and} \quad \eta = \sqrt{\frac{2qN^{1-q}}{T(1-q)(K \wedge N)^q}}\,,$$

satisfies

$$R_T \le 2\sqrt{e(K \wedge N)T\bigl(2 + \ln\bigl(N/(K \wedge N)\bigr)\bigr)}\,.$$

Proof. Let $i^* \in \arg\min_{i \in V} \sum_{t=1}^T y_t(i)$, and note that $R_T = \mathbb{E}\sum_{t=1}^T \bigl(y_t(I_t) - y_t(i^*)\bigr)$ as $\mathbb{E}\,\ell_t(A_t) = \mathbb{E}\,y_t(I_t)$. For round $t \in [T]$, let $\mathcal{F}_t \coloneqq \sigma(I_1, A_1, \dots, I_t, A_t)$ denote the σ-algebra generated by the random events up to the end of round t, and let $\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ with $\mathcal{F}_0$ being the trivial σ-algebra. For action $a \in \mathcal{A}$, let $\phi_t(a) \coloneqq \sum_{i \in V} p_t(i)\theta_t^i(a)$ and note that conditioned on $\mathcal{F}_{t-1}$, A_t is distributed according to ϕ_t. As p_t is $\mathcal{F}_{t-1}$-measurable, it is then easy to verify that $\mathbb{E}_t\widehat{y_t} = y_t$. Hence, Lemma 2 in [14] implies that

$$R_T \le \frac{N^{1-q}}{(1-q)\eta} + \frac{\eta}{2q}\sum_{t=1}^T \mathbb{E}\left[\sum_{i \in V} p_t(i)^{2-q}\,\widehat{y}_t(i)^2\right]. \tag{1}$$

For fixed $t \in [T]$ and $i \in V$, we have that

$$\mathbb{E}_t\left[\widehat{y}_t(i)^2\right] = \mathbb{E}_t\left[\frac{\theta_t^i(A_t)^2}{\phi_t(A_t)^2}\ell_t(A_t)^2\right] \le \mathbb{E}_t\left[\frac{\theta_t^i(A_t)^2}{\phi_t(A_t)^2}\right] = \mathbb{E}_t\left[\sum_{a \in \mathcal{A}}\frac{\theta_t^i(a)^2}{\phi_t(a)^2}\mathbb{I}\{a = A_t\}\right] = \sum_{a \in \mathcal{A}}\frac{\theta_t^i(a)^2}{\phi_t(a)} \tag{2}$$

where the inequality holds because $\ell_t(A_t) \in [0,1]$ and the final equality holds because $\mathbb{E}_t\,\mathbb{I}\{a = A_t\} = \mathbb{P}(a = A_t \mid \mathcal{F}_{t-1}) = \phi_t(a)$. Hence, it holds that

$$\begin{aligned}
\mathbb{E}_t\left[\sum_{i \in V} p_t(i)^{2-q}\,\widehat{y}_t(i)^2\right] &= \sum_{a \in \mathcal{A}}\frac{\sum_{i \in V} p_t(i)^{2-q}\theta_t^i(a)^2}{\phi_t(a)} \\
&\le \sum_{a \in \mathcal{A}}\frac{\sum_{i \in V} p_t(i)^{2-q}\theta_t^i(a)^{2-q}}{\phi_t(a)}\max_{i \in V}\theta_t^i(a)^q \\
&\le \sum_{a \in \mathcal{A}}\frac{\bigl(\sum_{i \in V} p_t(i)\theta_t^i(a)\bigr)^{2-q}}{\phi_t(a)}\max_{i \in V}\theta_t^i(a)^q \\
&= \sum_{a \in \mathcal{A}}\phi_t(a)\left(\frac{\max_{i \in V}\theta_t^i(a)}{\phi_t(a)}\right)^q \le \left(\sum_{a \in \mathcal{A}}\max_{i \in V}\theta_t^i(a)\right)^q \le (K \wedge N)^q\,,
\end{aligned}$$

where the second inequality follows from the superadditivity of x^{2-q} for $x \ge 0$ and $q \in (0,1)$, the third inequality follows from the concavity of x^q for $q \in (0,1)$ because of Jensen's inequality, and the last inequality holds since $\max_{i \in V}\theta_t^i(a) \le \min\bigl\{1, \sum_{i \in V}\theta_t^i(a)\bigr\}$. Substituting back into (1) yields that

$$R_T \le \frac{N^{1-q}}{(1-q)\eta} + \frac{\eta}{2q}(K \wedge N)^q T\,.$$

For brevity, let $\xi \coloneqq (K \wedge N)$. In a similar manner to the proof of Theorem 1 in [14], substituting the specified values of η and q allows us to conclude the proof:

$$\begin{aligned}
R_T &\leq \sqrt{\frac{2N^{1-q}\xi^q}{q(1-q)}T} \\
&= \sqrt{2T\exp\left(1+\frac{1}{2}\ln(\xi N) - \frac{1}{2}\sqrt{\ln{(N/\xi)}^2+4}\right)\left(2+\sqrt{\ln{(N/\xi)}^2+4}\right)} \\
&\leq \sqrt{2T\exp\left(1+\frac{1}{2}\ln(\xi N) - \frac{1}{2}\ln{(N/\xi)}\right)\left(2+\sqrt{\ln{(N/\xi)}^2+4}\right)} \\
&= \sqrt{2e\xi T\left(2+\sqrt{\ln{(N/\xi)}^2+4}\right)} \leq 2\sqrt{e\xi T\sqrt{\ln{(N/\xi)}^2+4}} \\
&\leq 2\sqrt{e\xi T\,(2+\ln(N/\xi))}\,.
\end{aligned}$$

□

4 An Improved Instance-Based Regret Bound

We now obtain a more refined regret bound whose form is analogous to the bound of Theorem 3.1 except that it depends on the similarity between the experts' recommendations at each round, replacing $K \wedge N$ with an effective number of experts. Before discussing the algorithm, we introduce some relevant quantities from [13]. For any round $t \in [T]$ and $\tau \in \Delta_N$, define

$$Q_t(\tau) \coloneqq \sum_{i\in V} \tau(i)\chi^2\big(\theta_t^i \,\big\|\, \textstyle\sum_{j\in V}\tau(j)\theta_t^j\big) = \displaystyle\sum_{a\in\mathcal{A}} \frac{\sum_{i\in V}\tau(i)\theta_t^i(a)^2}{\sum_{j\in V}\tau(j)\theta_t^j(a)} - 1\,,$$

where $\chi^2(p\,\|\,q) \coloneqq \sum_{a\in\mathcal{A}} q(a)\big(p(a)/q(a)-1\big)^2 = \sum_{a\in\mathcal{A}} p(a)^2/q(a) - 1$ is the chi-squared divergence between distributions $p, q \in \Delta_K$. Additionally, let

$$C_t \coloneqq \sup_{\tau\in\Delta_N} Q_t(\tau) \qquad \text{and} \qquad \overline{C}_T \coloneqq \frac{1}{T}\sum_{t=1}^{T} C_t$$

be the chi-squared capacity of the recommended distributions at round t and its average over the T rounds. As remarked before, C_t is never larger than $(K \wedge N) - 1$ and can be arbitrarily smaller depending on the agreement between the experts at round t. In particular, it vanishes when all recommendations are identical.

The idea of Algorithm 2 is to tune Algorithm 1 as done in Theorem 3.1 but with $\overline{C}_T$ replacing $K \wedge N$. However, to avoid requiring prior knowledge of $\overline{C}_T$, we rely on a doubling trick to adapt to its value. In a given round t, we maintain a running instance of Algorithm 1 tuned with an estimate for $\overline{C}_T$. Let m_t be the round when the present execution of Algorithm 1 had started. If the current estimate is found to be smaller than $\frac{1}{2T}\sum_{s=m_t}^{t} Q_s(p_s)$, the algorithm is restarted and the estimate is (at least) doubled. This quantity we test against is a simple lower bound for $\overline{C}_T/2$ that can be constructed without computing the capacity at any round. As the value of $\overline{C}_T$ can be arbitrarily close to zero, the initial guess (which ideally should be a lower bound for $\overline{C}_T$) is left as a user-specified parameter (denoted by J) for the algorithm, and appears in the first (and more general) bound of Theorem 4.1. The second statement of the theorem shows that choosing $\ln(e^2N)/T$ as the initial guess suffices to obtain a bound of order $\sqrt{\sum_{t=1}^{T} C_t\big(1+\ln(N/\max\{\overline{C}_T, 1\})\big)}$, up to an additive $\ln N$ term. This simultaneously outperforms the $\sqrt{\sum_{t=1}^{T} C_t \ln N}$ bound of Eldowa et al. [13] and the $\sqrt{(K\wedge N)T\big(1+\ln(N/(K\wedge N))\big)}$ bound of Kale [16].

Algorithm 2 q-FTRL with the doubling trick for bandits with expert advice

```
input: $J \in (0, N]$
initialization: $r_1 \leftarrow \lceil \log_2 J \rceil - 1$, $m_1 \leftarrow 1$, $p_1(i) \leftarrow 1/N$ for all $i \in V$
define: For each integer $r \in (-\infty, \log_2 N]$,
    $$q_r \coloneqq \frac{1}{2}\left(1 + \frac{\ln(N/2^r)}{\sqrt{\ln(N/2^r)^2 + 4} + 2}\right)$$
    $$\eta_r \coloneqq \min\left\{\sqrt{\frac{q_r(N^{1-q_r} - 1)}{eT(1-q_r)\,(2^r)^{q_r}}},\ \frac{q_r}{1-q_r}\left(1 - e^{\frac{q_r - 1}{2 - q_r}}\right)\right\}$$
for $t = 1, \dots, T$ do
    receive expert advice $(\theta_t^i)_{i \in V}$
    draw expert $I_t \sim p_t$ and action $A_t \sim \theta_t^{I_t}$
    construct $\widehat{y}_t \in \mathbb{R}^N$ where $\widehat{y}_t(i) \coloneqq \frac{\theta_t^i(A_t)}{\sum_{j \in V} p_t(j)\theta_t^j(A_t)} \ell_t(A_t)$ for all $i \in V$
    if $\frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) > 2^{r_t+1}$ then
        $p_{t+1}(i) \leftarrow 1/N$ for all $i \in V$
        $r_{t+1} \leftarrow \lceil \log_2\left(\frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s)\right) \rceil - 1$, $m_{t+1} \leftarrow t + 1$
    else
        $p_{t+1} \leftarrow \arg\min_{p \in \Delta_N} \eta_{r_t} \left\langle \sum_{s=m_t}^{t} \widehat{y}_s, p \right\rangle + \psi_{q_{r_t}}(p)$
        $r_{t+1} \leftarrow r_t$, $m_{t+1} \leftarrow m_t$
    end if
end for
```

The proof (deferred to Appendix B) combines elements from the proof of Theorem 1 of Eldowa et al. [13] and the proof of Theorem 3 of Eldowa et al. [14], who adopt a similar algorithm to address online learning with time-varying feedback graphs. Compared to the latter work, we require a more refined analysis to account for the case when $\overline{C}_T < 1$. This refinement is achieved in part via the use of Lemma A.1, which also allows adapting the analysis of Eldowa et al. [13] to account for the fact that we use the q-Tsallis entropy as a regularizer in place of the Shannon entropy.

Theorem 4.1.

$$R_T \leq 38e\sqrt{(\overline{C}_T \vee J)T \ln\left(\frac{e^2 N}{\overline{C}_T \vee J \vee 1}\right)} + \log_2\left(\frac{\overline{C}_T}{J}\right)_+ + \frac{18e}{5}\log_2\left(\frac{4\left((JT \vee \overline{C}_T T) \wedge \ln(e^2 N)\right)}{JT}\right)_+ \ln\left(e^2 N\right) + 1\,.$$

In particular, setting $J = \ln(e^2 N)/T$ yields that

$$R_T \leq 38e\sqrt{\overline{C}_T T \ln\left(\frac{e^2 N}{\overline{C}_T \vee 1}\right)} + \log_2\left(\frac{\overline{C}_T T}{\ln(e^2 N)}\right)_+ + 46e\ln\left(e^2 N\right) + 1\,.$$

5 A Lower Bound for Restricted Advice via Feedback Graphs

In this section, we provide a novel lower bound on the minimax regret for a slightly harder formulation of the multi-armed bandit problem with expert advice. We consider a setting where the learner picks an expert I_t (possibly at random) at the beginning of each round $t \in [T]$ without observing any of the experts' recommendations beforehand. The action A_t to be executed is subsequently drawn by the environment from the chosen expert's distribution, i.e., $A_t \sim \theta_t^{I_t}$. Afterwards, the learner observes A_t, the incurred loss $\ell_t(A_t)$, and the advice θ_t^i only of experts $i \in V$ that have the drawn action A_t in their support, i.e., $\theta_t^i(A_t) > 0$. For experts outside this set, the learner can only infer that, by definition, $\theta_t^i(A_t) = 0$. We will refer to this variation of the problem as the multi-armed bandit with *restricted* expert advice (note that this differs from the limited expert advice model studied by Kale [16]). Observe that Algorithm 1 is still implementable in this scenario and guarantees a regret upper bound of order $\sqrt{\xi T\,(1 + \ln(N/\xi))}$ for $\xi \coloneqq K \wedge N$, as previously analyzed. Here we show that the regret of Algorithm 1 is the best regret we can hope for, up to constant factors, for any number K of actions and any number N of experts. While a $\Omega(\sqrt{NT})$ regret lower bound in the case $N \leq K$ is immediate (as mentioned before), the following theorem provides an $\Omega\big(\sqrt{KT\ln(N/K)}\big)$ lower bound when $N > K$, improving upon the $\Omega\big(\sqrt{KT(\ln N)/(\ln K)}\big)$ lower bound of Seldin and Lugosi [24].

In what follows, we fix $N > K \geq 2$. We derive the lower bound relying on a reduction from the multi-armed bandit problem with feedback graphs [20, 3, 4, 5]. In this variant of the bandit problem, we assume there exists a graph $G = (V, E)$ over a finite set $V = [N]$ of actions from which the learner selects one action $J_t \in V$ at each round $t \in [T]$. Then, the learner observes the losses of the neighbours of J_t in G. For the construction of the lower bound, it suffices to assume that G is undirected and contains all self-loops, i.e., $(i, i) \in E$ for each $i \in V$. Consequently, the learner always observes the loss of the selected action and the graph G is strongly observable—see [4] for a classification of feedback graphs. We particularly focus on a specific family of graphs (also considered in the recent work of Chen et al. [11]) where the N vertices are partitioned into disjoint cliques with self-loops. Precisely, we let $M \coloneqq \lfloor K/2 \rfloor \geq 1$ be the number of disjoint cliques in G. For any $k \in [M]$, let C_k be the set of vertices of the k-th clique in G. Since each C_k is a clique with all self-loops, we have that $(i, j) \in E$ if and only if $i, j \in C_k$ for some $k \in [M]$, and thus $E = \bigcup_{k\in[M]}(C_k \times C_k)$. Additionally, for our purposes, we only consider the partition into cliques $C_k = \big\{i \in [N] : i \equiv k \mod M\big\}$ of roughly the same size $|C_k| \geq \lfloor N/M \rfloor \geq \lfloor 2N/K \rfloor \geq N/K$.

Hence, we will focus on the class of instances, denoted by Ξ_{FG}, of the multi-armed bandit problem with feedback graphs where the graph assumes the particular structure described above. In particular, any instance $\mathcal{I} \in \Xi_{\mathrm{FG}}$ is defined as a tuple $\mathcal{I} \coloneqq (T, G, \mathcal{L})$ containing the number T of rounds, the feedback graph $G = (V, E)$ over $V = [N]$ composed of the disjoint cliques $C_1, \ldots, C_M$ as defined above, and the sequence $\mathcal{L} \coloneqq (\ell_t)_{t\in[T]}$ of binary loss functions $\ell_t \colon V \longrightarrow \{0, 1\}$ over V. On the other hand, we let Ξ_{REA} be the class of instances for the multi-armed bandit problem with restricted expert advice, with N experts and K actions. An instance $\mathcal{I} \in \Xi_{\mathrm{REA}}$ is a tuple $\mathcal{I} \coloneqq \big(T, V, \mathcal{A}, \Theta, \mathcal{L}\big)$ containing the number T of rounds, the set $V = [N]$ of experts, the set $\mathcal{A} = [K]$ of actions, the sequence $\Theta \coloneqq (\theta_t^i)_{i\in V, t\in[T]}$ of expert advice where $\theta_t^i \in \Delta_K$, and the sequence $\mathcal{L} \coloneqq (\ell_t)_{t\in[T]}$ of loss functions $\ell_t \colon \mathcal{A} \longrightarrow \{0, 1\}$ over $\mathcal{A}$. The sought result is established by showing that the worst-case regret of any algorithm against a particular subset of instances in Ξ_{REA} is order-wise at least as large as the minimax regret on Ξ_{FG}, combined with a lower bound on the latter quantity by Chen et al. [11].

Theorem 5.1. *Let $\mathcal{B}$ be any possibly randomized algorithm for the multi-armed bandit problem with restricted expert advice for any number $K \geq 2$ of actions $\mathcal{A} = [K]$ and any number $N > K$ of experts $V = [N]$. Then, for a sufficiently large T, there exist a sequence $\ell_1, \ldots, \ell_T \colon \mathcal{A} \longrightarrow \{0, 1\}$ of binary loss functions and a sequence $(\theta_t^i)_{i\in V, t\in[T]}$ of expert advice such that the expected regret of $\mathcal{B}$ is $\Omega\big(\sqrt{KT\ln(N/K)}\big)$.*

Proof. We first describe a reduction from the multi-armed bandit problem with feedback graphs to the multi-armed bandit problem with restricted expert advice. We accomplish this by providing a mapping $\rho \colon \Xi_{\mathrm{FG}} \longrightarrow \Xi_{\mathrm{REA}}$ from the considered instance class Ξ_{FG} of the former problem to the instance class Ξ_{REA} of the latter.

Consider any instance $\mathcal{I} \coloneqq (T, G, \mathcal{L}) \in \Xi_{\mathrm{FG}}$ and recall that $G = (V, E)$ is a union of $M = \lfloor K/2 \rfloor$ disjoint cliques $C_1, \ldots, C_M$ over $V = [N]$. The mapped instance $\rho(\mathcal{I}) \coloneqq (T, V, \mathcal{A}, \Theta, \mathcal{L}') \in \Xi_{\mathrm{REA}}$ is defined over the same number of rounds T and an experts set corresponding to the actions V in the original instance $\mathcal{I}$, whose sequence of recommendations is provided by $\Theta = (\theta_t^i)_{i\in V, t\in[T]}$. We first observe that the cardinality of the new action set $\mathcal{A} = [K]$ does relate to the number of cliques M. In particular, considering the partition of experts given by the cliques in G, we also partition the actions (in the expert advice instance $\rho(\mathcal{I})$) by associating 2 actions to each clique. Precisely, for any $k \in [M]$, we associate actions $\mathcal{A}_k \coloneqq \{2k-1, 2k\}$ to C_k. If K is even, this partitions the entire set of actions $\mathcal{A}$, while it leaves out action K otherwise. We can ignore the latter case and assume K is even without loss of generality, since we can otherwise leave action K outside of the support of any expert advice $\theta_t^i \in \Delta_K$ in the following construction (thus becoming a spurious action).

Second, we focus on the construction of the loss sequence $\mathcal{L}' \coloneqq (\ell'_1, \dots, \ell'_T)$. For any $t \in [T]$, we define $\ell'_t \in \{0,1\}^{\mathcal{A}}$ as

$$\ell'_t(2k-1) \coloneqq 0 \qquad \text{and} \qquad \ell'_t(2k) \coloneqq 1 \qquad \forall k \in [M]\cdot$$

Finally, we define the sequence of expert advice $(\theta_t^i)_{i\in V, t\in[T]}$ depending on the sequence of losses $\mathcal{L}$ of the starting instance $\mathcal{I}$. For any $t \in [T]$, any $k \in [M]$, and any $i \in C_k$, we define $\theta_t^i \in \Delta_K$ as

$$\theta_t^i \coloneqq \begin{cases} \delta_{2k-1} & \text{if } \ell_t(i) = 0 \\ \delta_{2k} & \text{if } \ell_t(i) = 1 \end{cases},$$

where $\delta_j \in \Delta_K$ is the Dirac delta at $j \in \mathcal{A}$. This ensures that the loss of expert i at round t, given by $y_t(i) = \sum_{a\in\mathcal{A}} \theta_t^i(a)\ell'_t(a)$ coincides with $\ell_t(i)$, the loss of action i in the original feedback graphs instance at the same round. Moreover, the knowledge of $\ell_t(i)$ suffices to infer θ_t^i.

At this point, given our instance mapping ρ and our algorithm $\mathcal{B}$, we design an algorithm $\mathcal{B}_\rho$ for the class Ξ_{FG}. Consider any instance $\mathcal{I} \in \Xi_{\text{FG}}$. Over the interaction period, the algorithm $\mathcal{B}_\rho$, without requiring prior knowledge of $\mathcal{I}$, maintains a running realization of $\mathcal{B}$ on instance $\rho(\mathcal{I})$. At any round $t \in [T]$, let I_t be the expert selected by algorithm $\mathcal{B}$ in $\rho(\mathcal{I})$, and let $k_t \in [M]$ be the index of the clique I_t belongs to, i.e., $I_t \in C_{k_t}$. Algorithm $\mathcal{B}_\rho$, interacting with the instance $\mathcal{I}$, executes action $J_t = I_t$ provided by $\mathcal{B}$ and observes the losses $(\ell_t(i))_{i\in C_{k_t}}$. Then, thanks to the design of the mapping ρ, $\mathcal{B}_\rho$ can construct and provide $\mathcal{B}$ the feedback it requires and which complies with instance $\rho(\mathcal{I})$. Namely, it determines that $A_t = 2k_t - 1$ if $\ell_t(J_t) = 0$ or else that $A_t = 2k_t$, then passes A_t, its loss $\ell'_t(A_t)$ (trivially determined), and the restricted advice $(\theta_t^i)_{i\in C_{k_t}}$ to $\mathcal{B}$. The last of which is a super-set of the recommended distributions having positive support on A_t since A_t is never picked by experts outside C_{k_t} by construction.

Now, let

$$R^{\mathcal{B}}(\mathcal{I}') \coloneqq \mathbb{E}\left[\sum_{t=1}^T \ell'_t(A_t)\right] - \min_{i\in V} \sum_{t=1}^T \sum_{a\in\mathcal{A}} \theta_t^i(a)\ell'_t(a) = \mathbb{E}\left[\sum_{t=1}^T y_t(I_t)\right] - \min_{i\in V} \sum_{t=1}^T \sum_{a\in\mathcal{A}} \theta_t^i(a)\ell'_t(a)$$

be the expected regret of algorithm $\mathcal{B}$ on some instance $\mathcal{I}' = \left(T, V, \mathcal{A}, (\theta_t^i)_{i\in V, t\in[T]}, (\ell'_t)_{t\in[T]}\right) \in \Xi_{\text{REA}}$. Similarly, let

$$R^{\mathcal{B}_\rho}(\mathcal{I}) \coloneqq \mathbb{E}\left[\sum_{t=1}^T \ell_t(J_t)\right] - \min_{i\in V} \sum_{t=1}^T \ell_t(i)$$

be the expected regret of algorithm $\mathcal{B}_\rho$ on some instance $\mathcal{I} = \left(T, G, (\ell_t)_{t\in[T]}\right) \in \Xi_{\text{FG}}$. Since $J_t = I_t$, we have that $y_t(I_t) = \ell_t(J_t)$ via the properties of ρ laid out before. Hence, we can conclude that $R^{\mathcal{B}}(\rho(\mathcal{I})) = R^{\mathcal{B}_\rho}(\mathcal{I})$ for any instance $\mathcal{I} \in \Xi_{\text{FG}}$. Define $\rho(\Xi_{\text{FG}}) \coloneqq \left\{\rho(\mathcal{I}) : \mathcal{I} \in \Xi_{\text{FG}}\right\} \subseteq \Xi_{\text{REA}}$ as the subclass of instances in Ξ_{REA} obtained from Ξ_{FG} via ρ. Then, it holds that

$$\sup_{\mathcal{I}\in\Xi_{\text{REA}}} R^{\mathcal{B}}(\mathcal{I}) \geq \sup_{\mathcal{I}\in\rho(\Xi_{\text{FG}})} R^{\mathcal{B}}(\mathcal{I}) = \sup_{\mathcal{I}\in\Xi_{\text{FG}}} R^{\mathcal{B}}(\rho(\mathcal{I})) = \sup_{\mathcal{I}\in\Xi_{\text{FG}}} R^{\mathcal{B}_\rho}(\mathcal{I}) .$$

On the other hand, Lemma E.1 in [11] implies that

$$\sup_{\mathcal{I}\in\Xi_{\mathrm{FG}}} R_T^{\mathcal{B}_\rho}(\mathcal{I}) = \Omega\left(\sqrt{T\sum_{k\in[M]}\ln(1+|C_k|)}\right) = \Omega\big(\sqrt{KT\ln(N/K)}\big)$$

for sufficiently large T since $\sum_{k\in[M]}\ln(1+|C_k|) \geq M\ln(N/M) \geq K\ln(2N/K)/4$, thus concluding the proof. □

6 Conclusion

As the lower bound of Theorem 5.1 was proved for a harder formulation of the problem, it remains to be shown whether the same impossibility result holds for the more standard setup. We conjecture it should be possible to prove such a lower bound. If it indeed holds, this would imply that the minimax regret in the two variants is of the same order; that is, as far as we are only concerned with the worst-case regret, the standard feedback setup would be shown to be essentially as hard as the restricted one. In fact, in a concurrent work, Ito [15] proved a lower bound of the same order for a less restricted setting by directly adapting the techniques of Chen et al. [11]. In [15], the recommendations of *all* the experts are revealed to the learner at the end of each round; nonetheless, the learner must still choose an expert to follow before receiving the recommendations.

Acknowledgments

NCB, KE, and EE acknowledge the financial support from the MUR PRIN grant 2022EKNE5K (Learning in Markets and Society), the FAIR (Future Artificial Intelligence Research) project, and the EU Horizon RIA under grant agreement 101120237, project ELIAS (European Lighthouse of AI for Sustainability).

A Auxiliary Results

Lemma A.1. *Let $q \in (0,1)$, $b > 0$, $c > 1$, and $(y_t)_{t=1}^T$ be a sequence of loss vectors in $\mathbb{R}^N$ satisfying $y_t(i) \geq -b$ for all $t \in [T]$ and $i \in [N]$. Let $(p_t)_{t=1}^{T+1}$ be the predictions of FTRL with decision set Δ_N and the q-Tsallis regularizer ψ_q over this sequence of losses; that is, $p_1 = \arg\min_{p\in\Delta_N}\psi_q(p)$, and for $t \in [T]$,*

$$p_{t+1} = \operatorname*{arg\,min}_{p\in\Delta_N} \eta\sum_{s=1}^{t}\langle y_s, p\rangle + \psi_q(p)\,,$$

assuming the learning rate η satisfies $0 < \eta \leq \frac{q}{(1-q)b}\left(1 - c^{\frac{q-1}{2-q}}\right)$. Then for any $u \in \Delta_N$,

$$\sum_{t=1}^{T}\langle p_t - u, y_t\rangle \leq \frac{N^{1-q}-1}{(1-q)\eta} + \frac{\eta c}{2q}\sum_{t=1}^{T}\sum_{i=1}^{N} p_t(i)^{2-q}\, y_t(i)^2\,.$$

Proof. Let $p'_{t+1} \coloneqq \arg\min_{p\in\mathbb{R}^N_{\geq 0}}\langle p, y_t\rangle + D_{\psi_q}(p, p_t)$, where $D_{\psi_q}(\cdot,\cdot)$ denotes the Bregman divergence based on ψ_q. Via Lemma 7.14 in [22] we have that

$$\begin{aligned}\sum_{t=1}^{T}\langle p_t - u, y_t\rangle &\leq \frac{\psi_q(u)-\psi_q(p_1)}{\eta} + \frac{\eta}{2q}\sum_{t=1}^{T}\sum_{i=1}^{N} z_t(i)^{2-q}\, y_t(i)^2 \\ &\leq \frac{N^{1-q}-1}{(1-q)\eta} + \frac{\eta}{2q}\sum_{t=1}^{T}\sum_{i=1}^{N} z_t(i)^{2-q}\, y_t(i)^2\,,\end{aligned}$$

where z_t lies on the line segment between p_t and p'_{t+1}. A simple derivation shows that

$$p'_{t+1}(i) = p_t(i)\left(\frac{1}{1+\eta\frac{1-q}{q}y_t(i)p_t(i)^{1-q}}\right)^{\frac{1}{1-q}},$$

for each $i \in [N]$. On the other hand, it holds that

$$\eta\frac{1-q}{q}y_t(i)p_t(i)^{1-q} \geq -\eta\frac{1-q}{q}bp_t(i)^{1-q} \geq -\eta\frac{1-q}{q}b \geq c^{\frac{q-1}{2-q}} - 1\,,$$

where the first inequality uses that $y_t(i) \geq -b$ (and that $p_t(i), \eta > 0$), the second uses that $p_t(i) \leq 1$, and the third uses that $\eta \leq \frac{q}{(1-q)b}\left(1 - c^{\frac{q-1}{2-q}}\right)$. This entails that $p'_{t+1}(i) \leq c^{\frac{1}{2-q}}p_t(i)$, which implies that $z_t(i) \leq c^{\frac{1}{2-q}}p_t(i)$ concluding the proof. □

B Proof of Theorem 4.1

Theorem 4.1.

$$R_T \leq 38e\sqrt{(\overline{C}_T \vee J)T\ln\left(\frac{e^2N}{\overline{C}_T \vee J \vee 1}\right)} + \log_2\left(\frac{\overline{C}_T}{J}\right)_+ + \frac{18e}{5}\log_2\left(\frac{4\left(\left(JT \vee \overline{C}_T T\right) \wedge \ln(e^2N)\right)}{JT}\right)_+ \ln\!\left(e^2N\right) + 1\,.$$

In particular, setting $J = \ln(e^2N)/T$ yields that

$$R_T \leq 38e\sqrt{\overline{C}_T T\ln\left(\frac{e^2N}{\overline{C}_T \vee 1}\right)} + \log_2\left(\frac{\overline{C}_T T}{\ln(e^2N)}\right)_+ + 46e\ln\!\left(e^2N\right) + 1\,.$$

Proof. For brevity, we define $U \coloneqq \overline{C}_T \vee J$. Let $s \coloneqq \lceil\log_2 J\rceil - 1$ and $n \coloneqq \lceil\log_2 U\rceil - 1$, the latter of which is the largest value that r_t can take, since for any round t,

$$\frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \leq \frac{1}{T}\sum_{s=1}^{T} Q_s(p_s) \leq \frac{1}{T}\sum_{s=1}^{T} C_s \leq 2^{n+1}\,.$$

For any $r \in \{s, \ldots, n\}$, let T_r be the index of the first round $t \in [T+1]$ such that $r_t \geq r$, or let $T_r \coloneqq T+1$ if no such round exists. Additionally, define $T_{n+1} \coloneqq T+2$. Note that $q_r \in [1/2, 1)$ for any $r \in \{s, \ldots, n\}$. Let $i^* \in \arg\min_{i\in V}\sum_{t=1}^{T} y_t(i)$. We start by decomposing the regret over the intervals corresponding to fixed values of $r_t \in \{s, \ldots, n\}$ and bounding the instantaneous regret at the last step of each but the last interval by 1:

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\leq \mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] + n - s \\
&\leq \mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] + \log_2\!\big(U/J\big) + 1\,. \qquad (3)
\end{aligned}$$

We may assume, without loss of generality, that $T_{r+1} \geq T_r + 2$ for any r; otherwise, the corresponding sum in the first term of the decomposition is empty.

Let $\mathbf{e}_{i^*} \in \mathbb{R}^N$ be the indicator vector for i^* and define $\tilde{y}_t \in \mathbb{R}^N$ where $\tilde{y}_t(i) \coloneqq \widehat{y}_t(i) - \ell_t(A_t)$ for every $i \in V$. Similar to the proof of Theorem 3 in [14], we note that for each $r \in \{s, \dots, n\}$,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} \big(y_t(I_t) - y_t(i^*)\big)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\right\}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\right\}\langle p_t - \mathbf{e}_{i^*}, \widehat{y}_t\rangle\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\right\}\langle p_t - \mathbf{e}_{i^*}, \tilde{y}_t\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} \langle p_t - \mathbf{e}_{i^*}, \tilde{y}_t\rangle\right]
\end{aligned}
$$

where (a) follows since $\mathbb{E}_t\left[y_t(I_t)\right] = \sum_{i \in V} p_t(i) y_t(i)$, $\mathbb{E}_t\left[\widehat{y}_t\right] = y_t$, and the indicator at round t is measurable with respect to $\mathcal{F}_{t-1}$ (where $\mathcal{F}_{t-1}$ and $\mathbb{E}_t$ are defined in the same way as in the proof of Theorem 3.1); and (b) follows since $p_t, \mathbf{e}_{i^*} \in \Delta_N$ and $\widehat{y}_t(i) - \tilde{y}_t(i) = \ell_t(A_t)$ is identical for all $i \in V$. Similarly to the last argument, the fact that $\left\langle \tilde{y}_s - \widehat{y}_s, p - q \right\rangle = 0$ holds for any $p, q \in \Delta_N$ at any round s implies that p_{t+1} can be equivalently defined as $\arg\min_{p \in \Delta_N} \eta_{r_t} \left\langle \sum_{s=m_t}^{t} \tilde{y}_s, p \right\rangle + \psi_{q_{r_t}}(p)$. Hence, using that $\tilde{y}_t(i) \ge -1$, we can invoke Lemma A.1 (with $b = 1$ and $c = e$) to obtain that

$$
\sum_{t=T_r}^{T_{r+1}-2} \langle p_t - \mathbf{e}_{i^*}, \tilde{y}_t\rangle \le \frac{N^{1-q_r} - 1}{(1 - q_r)\eta_r} + \frac{e\eta_r}{2q_r} \sum_{t=T_r}^{T_{r+1}-2} \sum_{i \in V} p_t(i)^{2-q_r} \tilde{y}_t(i)^2 \,.
$$

For any round $t \in [T]$ and action $a \in \mathcal{A}$, recall the definition $\phi_t(a) \coloneqq \sum_{i \in V} p_t(i)\theta_t^i(a)$. Similar to (2) in the proof of Theorem 3.1, we have that

$$
\begin{aligned}
\mathbb{E}_t\left[\tilde{y}_t(i)^2\right] &= \mathbb{E}_t\left[\ell_t(A_t)^2 \frac{\big(\theta_t^i(A_t) - \phi_t(A_t)\big)^2}{\phi_t(A_t)^2}\right] \\
&\le \mathbb{E}_t\left[\frac{\big(\theta_t^i(A_t) - \phi_t(A_t)\big)^2}{\phi_t(A_t)^2}\right] \\
&= \sum_{a \in \mathcal{A}} \frac{\big(\theta_t^i(a) - \phi_t(a)\big)^2}{\phi_t(a)} = \sum_{a \in \mathcal{A}} \phi_t(a)\left(\frac{\theta_t^i(a)}{\phi_t(a)} - 1\right)^2 = \chi^2(\theta_t^i \,\|\, \phi_t) \,.
\end{aligned}
$$

Hence, for any round t and any $r \in \{s, \dots, n\}$, it holds that

$$
\begin{aligned}
\mathbb{E}_t\left[\sum_{i\in V} p_t(i)^{2-q_r}\tilde{y}_t(i)^2\right] &\le \sum_{i\in V} p_t(i)^{2-q_r}\chi^2(\theta_t^i \,\|\, \phi_t)\\
&= Q_t(p_t)\sum_{i\in V}\frac{p_t(i)\chi^2(\theta_t^i \,\|\, \phi_t)}{Q_t(p_t)}p_t(i)^{1-q_r}\\
&\le Q_t(p_t)\left(\sum_{i\in V}\frac{p_t(i)\chi^2(\theta_t^i \,\|\, \phi_t)}{Q_t(p_t)}p_t(i)\right)^{1-q_r}\\
&= Q_t(p_t)^{q_r}\left(\sum_{i\in V} p_t(i)^2\chi^2(\theta_t^i \,\|\, \phi_t)\right)^{1-q_r}\\
&= Q_t(p_t)^{q_r}\left(\sum_{i\in V} p_t(i)^2\sum_{a\in\mathcal{A}}\frac{\theta_t^i(a)^2}{\phi_t(a)} - \sum_{i\in V} p_t(i)^2\right)^{1-q_r}\\
&= Q_t(p_t)^{q_r}\left(\sum_{a\in\mathcal{A}}\frac{\sum_{i\in V} p_t(i)^2\theta_t^i(a)^2}{\sum_{j\in V} p_t(j)\theta_t^j(a)} - \sum_{i\in V} p_t(i)^2\right)^{1-q_r}\\
&\le Q_t(p_t)^{q_r}\left(\sum_{a\in\mathcal{A}}\sum_{i\in V} p_t(i)\theta_t^i(a) - \sum_{i\in V} p_t(i)^2\right)^{1-q_r}\\
&= Q_t(p_t)^{q_r}\left(1 - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \le Q_t(p_t)^{q_r}\,,
\end{aligned}
$$

where the second inequality follows from the definition of $Q_t(p_t)$ and the fact that x^{1-q_r} is concave in $x \ge 0$, and the third inequality uses the superadditivity of x^2 for non-negative real numbers and the non-negativity of the quantity in brackets. Let $T_{r:r+1} \coloneqq T_{r+1} - T_r - 1$, it then holds that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\sum_{i\in V} p_t(i)^{2-q_r}\tilde{y}_t(i)^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\right\}\sum_{i\in V} p_t(i)^{2-q_r}\tilde{y}_t(i)^2\right]\\
&\le \mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t)^{q_r}\right]\\
&\le \mathbb{E}\left[T_{r:r+1}\left(\frac{1}{T_{r:r+1}}\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t)\right)^{q_r}\right]\\
&\le \mathbb{E}\left[T_{r:r+1}\left(\frac{T}{T_{r:r+1}}2^{r+1}\right)^{q_r}\right] \le 2T\,(2^r)^{q_r}\,,
\end{aligned}
$$

where the second inequality uses the concavity of x^{q_r} in $x \ge 0$ and the third uses that $(1/T)\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t) \le 2^{r+1}$ since the algorithm is not reset in the interval $[T_r, T_{r+1} - 2]$. Overall, we have shown that

$$
\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] \le \frac{N^{1-q_r} - 1}{(1 - q_r)\eta_r} + \frac{e\eta_r}{q_r}\,(2^r)^{q_r}\,T\,.
$$

If $\sqrt{\frac{q_r(N^{1-q_r}-1)}{eT(1-q_r)(2^r)^{q_r}}} \le \frac{q_r}{1-q_r}\left(1-e^{\frac{q_r-1}{2-q_r}}\right)$, then substituting the values of η_r and q_r gives that

$$\begin{aligned}
\frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \frac{e\eta_r}{q_r}\,(2^r)^{q_r}\,T &= 2\sqrt{\frac{e(N^{1-q_r}-1)\,(2^r)^{q_r}\,T}{q_r(1-q_r)}} \\
&= 2\sqrt{\frac{N^{1-q_r}-1}{N^{1-q_r}}}\sqrt{\frac{eN^{1-q_r}\,(2^r)^{q_r}\,T}{q_r(1-q_r)}} \\
&\le 2e\sqrt{2}\sqrt{\frac{N^{1-q_r}-1}{N^{1-q_r}}}\sqrt{2^r\,(2+\ln(N2^{-r}))\,T} \\
&\le 2e\sqrt{2}\left(\sqrt{\frac{\ln N}{\ln(N2^{-r})}}\wedge 1\right)\sqrt{2^r\,(2+\ln(N2^{-r}))\,T} \\
&= 2e\sqrt{2}\sqrt{2^r\ln\big(e^2N(2^{-r}\wedge 1)\big)T}\,,
\end{aligned}$$

where the first inequality holds via the same arguments laid in the last passage of the proof of Theorem 3.1, and the second inequality holds since

$$\begin{aligned}
\frac{N^{1-q_r}-1}{N^{1-q_r}} &= 1-\exp\left(-\ln\big(N^{1-q_r}\big)\right) \\
&\le (1-q_r)\ln N \\
&= \frac{1}{2}\left(1-\frac{\ln(N/2^r)}{\sqrt{\ln(N/2^r)^2+4}+2}\right)\ln N \\
&= \frac{\ln N}{2\ln(N/2^r)}\left(2+\ln(N/2^r)-\sqrt{\ln(N/2^r)^2+4}\right) \le \frac{\ln N}{\ln(N/2^r)}\,,
\end{aligned}$$

where the first inequality follows from the fact that $1-e^{-x}\le x$. Otherwise, if $\sqrt{\frac{q_r(N^{1-q_r}-1)}{eT(1-q_r)(2^r)^{q_r}}} > \frac{q_r}{1-q_r}\left(1-e^{\frac{q_r-1}{2-q_r}}\right)$, then η_r takes the latter value and we obtain that

$$\begin{aligned}
\frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \frac{e\eta_r}{q_r}\,(2^r)^{q_r}\,T &\le \frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \eta_r\frac{N^{1-q_r}-1}{(1-q_r)}\left(\frac{1-q_r}{q_r\left(1-e^{\frac{q_r-1}{2-q_r}}\right)}\right)^2 \\
&= 2\frac{N^{1-q_r}-1}{q_r\left(1-e^{\frac{q_r-1}{2-q_r}}\right)} \\
&\le \frac{18\big(N^{1-q_r}-1\big)}{5q_r(1-q_r)} \\
&= \frac{18\,(2^r)^{-q_r}\,\big(N^{1-q_r}-1\big)\,(2^r)^{q_r}}{5q_r(1-q_r)} \\
&\le \frac{18e}{5}\,(2^r)^{1-q_r}\ln\big(e^2N(2^{-r}\wedge 1)\big) \\
&\le \frac{18e}{5}\left(1\vee\sqrt{2^r}\right)\ln\big(e^2N(2^{-r}\wedge 1)\big)\,,
\end{aligned}$$

where the last inequality holds since $q_r \geq 1/2$, and the second inequality holds since

$$1 - e^{\frac{q_r - 1}{2 - q_r}} \geq \frac{1 - q_r}{2 - q_r} - \frac{1}{2}\left(\frac{1 - q_r}{2 - q_r}\right)^2 = \frac{3 - q_r}{2(2 - q_r)^2}(1 - q_r) \geq \frac{5}{9}(1 - q_r)\ln\bigl(e^2 N(2^{-r} \wedge 1)\bigr),$$

where the first step uses that $e^{-x} \leq 1 - x + x^2/2$ for $x \geq 0$, and the last step uses again that $q_r \geq 1/2$. Hence, the results above yield that

$$\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\bigl(y_t(I_t) - y_t(i^*)\bigr)\right] \leq \max\left\{2e\sqrt{2}\sqrt{2^r T \ln\bigl(e^2 N(2^{-r} \wedge 1)\bigr)}\,,\ \frac{18e}{5}\bigl(1 \vee \sqrt{2^r}\bigr)\ln\bigl(e^2 N(2^{-r} \wedge 1)\bigr)\right\}. \tag{4}$$

Let $M \coloneqq \ln(e^2 N)/T$ and $m \coloneqq \log_2 M$, and note that $m \leq 0$ (and $M \leq 1$) by the assumption that $T \geq \ln(e^2 N)$. In the case when $n \leq 0$, we have that

$$\begin{aligned}
\mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\bigl(y_t(I_t) - y_t(i^*)\bigr)\right] &\leq \frac{18e}{5}\bigl((n \wedge \lfloor m \rfloor) - s + 1\bigr)_+ \ln\bigl(e^2 N\bigr) + 2e\sqrt{2}\sum_{r=n\wedge\lceil m \rceil}^{n}\sqrt{2^r T \ln\bigl(e^2 N\bigr)} \\
&\leq \frac{18e}{5}\log_2\bigl(4(U \wedge M)/J\bigr)_+ \ln\bigl(e^2 N\bigr) + 8e\sqrt{2UT\ln\bigl(e^2 N\bigr)}\,,
\end{aligned}$$

where the second inequality uses that

$$\sum_{r=\alpha}^{n}\bigl(\sqrt{2}\bigr)^r = \bigl(\sqrt{2}\bigr)^\alpha \sum_{r=0}^{n-\alpha}\bigl(\sqrt{2}\bigr)^r = \bigl(\sqrt{2}\bigr)^\alpha \frac{\bigl(\sqrt{2}\bigr)^{n-\alpha+1} - 1}{\sqrt{2} - 1} \leq \frac{\sqrt{2}}{\sqrt{2} - 1}\bigl(\sqrt{2}\bigr)^n \leq 4\sqrt{U}\,,$$

with $\alpha \coloneqq n \wedge \lceil m \rceil$. Otherwise, if $n > 0$, then

$$\begin{aligned}
\mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\bigl(y_t(I_t) - y_t(i^*)\bigr)\right] &\leq \frac{18e}{5}\log_2\,(4M/J)_+ \ln\bigl(e^2 N\bigr) + 8e\sqrt{2T\ln\bigl(e^2 N\bigr)} + \mathbb{E}\left[\sum_{r=s_+}^{n}\sum_{t=T_r}^{T_{r+1}-2}\bigl(y_t(I_t) - y_t(i^*)\bigr)\right] \\
&\leq \frac{18e}{5}\log_2\,(4M/J)_+ \ln\bigl(e^2 N\bigr) + 8e\sqrt{2T\ln\bigl(e^2 N\bigr)} + \frac{18e}{5}\sum_{r=0}^{n}\sqrt{2^r \ln\bigl(e^2 N 2^{-r}\bigr)T} \\
&\leq \frac{18e}{5}\log_2\,(4M/J)_+ \ln\bigl(e^2 N\bigr) + 8e\sqrt{2T\ln\bigl(e^2 N\bigr)} + 26e\sqrt{UT\ln\bigl(e^2 N/U\bigr)} \\
&\leq \frac{18e}{5}\log_2\,(4M/J)_+ \ln\bigl(e^2 N\bigr) + 38e\sqrt{UT\ln\bigl(e^2 N/U\bigr)}\,,
\end{aligned}$$

where the first inequality follows from the analysis of the first case with $n = 0$, the second inequality uses that $r \geq 0$ and the assumption that $T \geq \ln(e^2 N)$, the third inequality uses Lemma 4 in [14], and the fourth uses that $x\ln(e^2 N/x)$ is increasing in $[0, eN]$ and that $U \geq 2$ in this case. The theorem then follows by combining the bounds provided for the two cases with (3). □

References

[1] J. D. Abernethy, C. Lee, and A. Tewari. 2015. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc.

[2] Z. Allen-Zhu, S. Bubeck, and Y. Li. 2018. Make the minority great again: first-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 80. PMLR, 186–194.

[3] N. Alon, N. Cesa-Bianchi, C. Gentile, and Y. Mansour. 2013. From bandits to experts: a tale of domination and independence. *Advances in Neural Information Processing Systems*, 26.

[4] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. 2015. Online learning with feedback graphs: beyond bandits. In *Proceedings of The 28th Conference on Learning Theory* (Proceedings of Machine Learning Research). Vol. 40. PMLR, 23–35.

[5] N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir. 2017. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46, 6, 1785–1826.

[6] J. Audibert and S. Bubeck. 2009. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Conference on Learning Theory*.

[7] J.-Y. Audibert and S. Bubeck. 2010. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11, 94, 2785–2836.

[8] J.-Y. Audibert, S. Bubeck, and G. Lugosi. 2011. Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory* (Proceedings of Machine Learning Research). S. M. Kakade and U. von Luxburg, (Eds.) Vol. 19. PMLR, 107–132.

[9] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32, 1, 48–77.

[10] N. Cesa-Bianchi, P. Gaillard, C. Gentile, and S. Gerchinovitz. 2017. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Proceedings of the 2017 Conference on Learning Theory* (Proceedings of Machine Learning Research). Vol. 65. PMLR, 465–481.

[11] H. Chen, Y. He, and C. Zhang. 2024. On interpolating experts and multi-armed bandits. In *Proceedings of the 41st International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 235. PMLR, 6776–6802.

[12] A. Daniely and T. Helbertal. 2013. The price of bandit information in multiclass online classification. In *Proceedings of the 26th Annual Conference on Learning Theory* (Proceedings of Machine Learning Research). Vol. 30. PMLR, 93–104.

[13] K. Eldowa, N. Cesa-Bianchi, A. M. Metelli, and M. Restelli. 2024. Information capacity regret bounds for bandits with mediator feedback. *Journal of Machine Learning Research*, 25, 353, 1–36.

[14] K. Eldowa, E. Esposito, T. Cesari, and N. Cesa-Bianchi. 2023. On the minimax regret for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 46122–46133.

[15] S. Ito. 2024. On the minimax regret for contextual linear bandits and multi-armed bandits with expert advice. In *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 61793–61812.

[16] S. Kale. 2014. Multiarmed bandits with limited expert advice. In *Proceedings of The 27th Conference on Learning Theory* (Proceedings of Machine Learning Research). Vol. 35. PMLR, 107–122.

[17] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. 2010. Regret bounds for sleeping experts and bandits. *Machine learning*, 80, 2, 245–272.

[18] T. Lattimore and C. Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

[19] H. Luo, C.-Y. Wei, A. Agarwal, and J. Langford. 2018. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory* (Proceedings of Machine Learning Research). Vol. 75. PMLR, 1739–1776.

[20] S. Mannor and O. Shamir. 2011. From bandits to experts: on the value of side-observations. In *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc.

[21] H. B. McMahan and M. J. Streeter. 2009. Tighter bounds for multi-armed bandits with expert advice. In *Proceedings of the 22nd Conference on Learning Theory*.

[22] F. Orabona. 2023. A modern introduction to online learning. *arXiv preprint*, arXiv:1912.13213.

[23] Y. Seldin, K. Crammer, and P. Bartlett. 2013. Open problem: adversarial multiarmed bandits with limited advice. In *Proceedings of the 26th Annual Conference on Learning Theory* (Proceedings of Machine Learning Research). Vol. 30. PMLR, 1067–1072.

[24] Y. Seldin and G. Lugosi. 2016. A lower bound for multi-armed bandits with expert advice. In *The 13th European Workshop on Reinforcement Learning (EWRL)*.

Received 23 June 2024; accepted 14 April 2025