JMLR: Workshop and Conference Proceedings vol 35:1–16, 2014

Multiarmed Bandits With Limited Expert Advice

Satyen Kale[*] SATYEN@YAHOO-INC.COM
Yahoo! Labs New York

Abstract

We consider the problem of minimizing regret in the setting of advice-efficient multiarmed bandits with expert advice. We give an algorithm for the setting of K arms and N experts out of which we are allowed to query and use only M experts' advice in each round, which has a regret bound[1] of $\tilde{O}\left(\sqrt{\frac{\min\{K,M\}N}{M}T}\right)$ after T rounds. We also prove that any algorithm for this problem must have expected regret at least $\tilde{\Omega}\left(\sqrt{\frac{\min\{K,M\}N}{M}T}\right)$, thus showing that our upper bound is nearly tight. This solves the COLT 2013 open problem of Seldin et al. (2013).

1. Introduction

In many real world applications one is faced with the problem of choosing one of several actions: for example, in healthcare, a choice of treatment; in financial domains, a choice of investment. Typically in such scenarios one may utilize the advice of several domain experts to make an informed choice. Once an action is chosen, one obtains feedback for the action in terms of some loss (or reward), but no feedback for other actions is obtained. This is repeated over several rounds. Repeated decision-making in this context is modeled by the well-studied multiarmed bandits with expert advice problem (Auer et al., 2002). In this paper, we study an important practical consideration for this setting: frequently there are costs associated with obtaining useful advice, and budget constraints imply that only a few experts may be queried for advice. This constraint on the number of experts that can be queried in any round is modeled by the advice-efficient setting of the multiarmed bandits with expert advice problem, introduced by Seldin et al. (2013).

In this setting, in each round $t = 1, 2, \ldots, T$, the learner is required to pull one arm A_t from some set $\mathcal{A}$ of K arms. Simultaneously, an adversary sets losses $\ell_t(a) \in [0, 1]$ for each arm $a \in \mathcal{A}$, thus generating the loss vector $\ell_t \in \mathbb{R}^{\mathcal{A}}$. Assisting us in this task are N experts in the set $\mathcal{H}$. Each expert h can provide advice[2] on which arm to pull in the form of a probability distribution $\xi_t^h \in \mathbb{R}^{\mathcal{A}}$ on the set of arms. This advice gives the expert h an expected loss of $\xi_t^h \cdot \ell_t$ in round t. The catch is that we can only observe the advice of at most M experts of our choosing in each round. The goal is to choose subsets of M experts in each round to query the advice of, and using their advice play some arm $A_t \in \mathcal{A}$ (probabilistically, if desired) to minimize the expected regret with respect to the

[*] This work was done when the author was at IBM T. J. Watson Research Center.

1. Here, we use the $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ notation to suppress dependence on logarithmic factors in the problem parameters.
2. No assumptions are made on how this advice is chosen by the experts other in each round than that it is independent of the losses of the arms chosen by the adversary in that round.

loss of the best expert, where the regret is defined as:

$$\text{Regret}_T := \sum_{t=1}^{T} \ell_t(A_t) - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \xi_t^h \cdot \ell_t.$$

In this paper we give an algorithm whose expected regret is bounded by

$$\sqrt{\frac{2\min\{K, M\} N \log(N)}{M} T}$$

after T rounds, based on the Multiplicative Weights (MW) forecaster for prediction with expert advice (Littlestone and Warmuth, 1994). We can improve this upper bound using the PolyINF forecaster of Audibert and Bubeck (2010) to

$$4\sqrt{\frac{\min\{K, M\} N \log(\frac{8M}{\min\{K,M\}})}{M} T}.$$

This matches the regret of the best known algorithms for the special cases $M = 1$ and $M = N$, and interpolates between them for intermediate values of M. This solves the COLT 2013 open problem proposed by Seldin et al. (2013), and in fact gives a better regret bound than the bound conjectured in (Seldin et al., 2013), which was $O\left(\sqrt{\frac{KN\log(N)}{M}T}\right)$.

Furthermore, we also show that any algorithm for the problem must incur expected regret of $\Omega\left(\sqrt{\frac{\min\{K, \frac{M}{\log(K)}\}N}{M}T}\right)$ on some sequence of expert advice and arm losses, thus showing that our upper bound is nearly tight: the ratio between the upper and lower bounds is always bounded by $O(\max\{\sqrt{\log(K)}, \sqrt{\log(M/K)}\})$.

2. Preliminaries

For any event E, let $\mathbb{I}[E]$ be the indicator random variable set to 1 if E happens and 0 otherwise. In any round t of the algorithm, let $\Pr_t[\cdot]$ and $\mathbb{E}_t[\cdot]$ denote probability and expectation respectively conditioned on all the randomness defined up to round $t-1$. For two probability distributions $\mathbf{P}$ and $\mathbf{Q}$ defined on the same space let $\text{KL}(\mathbf{P} \parallel \mathbf{Q})$ and $d_{\text{TV}}(\mathbf{P}, \mathbf{Q})$ denote the KL-divergence and total variation distance between the two distributions respectively. Let $\|\cdot\|_p$ denote the p-norm for any $p \geq 1$.

Without loss of generality, we may assume that each expert suggests exactly one arm to play in any round; i.e. $\xi_t^h(a) = 1$ for exactly one arm $a \in \mathcal{A}$ and 0 for all other arms. Call such advice vectors "pure". To see this, for every expert h we can randomly round a general advice vector ξ_t^h to a pure vector by sampling some arm $a_h \sim \xi_t^h$ and constructing a new advice vector $\hat{\xi}_t^h$ by setting $\hat{\xi}_t^h(a_h) = 1$ and $\hat{\xi}_t^h(a) = 0$ for all $a \neq a_h$. Note that $\mathbb{E}[\hat{\xi}_t^h] = \xi_t^h$; thus for any expert h, following the randomly rounded advice $\hat{\xi}_t^h$ for $t = 1, 2, \ldots, T$ has the same expected cost as following the advice ξ_t^h. Since this randomized rounding trick can be applied to the advice (algorithmically for the observed advice, and conceptually for the unobserved advice), in the rest of the paper we assume that all advice vectors are pure vectors; this helps us in getting a tighter bound on the regret. Let a_t^h denote the action chosen by expert h at time t, so that the loss of the expert can be rewritten as $\xi_t^h \cdot \ell_t = \ell_t(a_t^h)$.

For any time period t and any set $U \subseteq \mathcal{H}$, define the "active set of arms" to be the set of all arms recommended by experts in U, i.e.

$$\mathcal{A}_t^U = \{a \in \mathcal{A} : \exists h \in U \text{ s.t. } a_t^h = a\}.$$

Note that since we are allowed to query at most M experts in any round, if U is the queried set of experts in round t, then $|\mathcal{A}_t^U| \leq \min\{K, M\}$; this leads to $\min\{K, M\}$ factor in the regret bound. Define $K' := \min\{K, M\}$, the effective number of arms.

Throughout the paper we also assume that $N \geq 2$ and $K \geq 2$: in the remaining cases we trivially get 0 regret.

3. Algorithm

The algorithm, dubbed LEXP, works as follows. Assume for simplicity[3] that M divides N, and in the beginning, partition the N experts into $R := \frac{N}{M}$ groups of size M arbitrarily. Run an algorithm for prediction with expert advice (such as Multiplicative Weights (MW) forecaster of Littlestone and Warmuth (1994), or the PolyINF forecaster of Audibert and Bubeck (2010)) on all the experts. In each round, this base expert learning algorithm computes a distribution over the experts. Then LEXP samples an expert from this distribution, and chooses the group of experts it belongs to to query for advice, thus ensuring that at most M experts are queried in any round. It then plays the action recommended by the chosen expert, and observes its loss. It then constructs unbiased loss estimators for all experts using the observed loss and queried advice and passes these to the base expert learning algorithm, which updates its distribution. The loss estimators are non-zero only for experts in the chosen group; thus they can be computed for all experts and the algorithm is well-defined. The pseudo-code follows.

4. Analysis

We first prove a number of utility lemmas. The first lemma shows that the loss estimators we construct are unbiased for all experts with positive probability in the distribution (and an underestimate in general):

Lemma 1 *For all rounds t and all experts h, we have $\mathbb{E}_t[Y_t^h] \leq \ell_t(a_t^h)$ with equality holding if $q_t(h) > 0$.*[4] *Thus, $\mathbb{E}_t[q_t(h)Y_t^h] = q_t(h)\ell_t(a_t^h)$, and unconditionally, $\mathbb{E}[Y_t^h] \leq \ell_t(a_t^h)$.*

Proof Let $h \in B_i$. For clarity, let $a = a_t^h$. If $\Pr_t[i, a] > 0$, then by the definition of the loss estimator in (1), we have

$$\mathbb{E}_t[Y_t^h] \;=\; \mathbb{E}_t[\hat{\ell}_t^i(a)] \;=\; \mathbb{E}_t\left[\ell_t(a)\frac{\mathbb{I}[I_t = i, A_t = a]}{\Pr_t[i, a]}\right] \;=\; \ell_t(a)\frac{\Pr_t[i, a]}{\Pr_t[i, a]} \;=\; \ell_t(a).$$

If $\Pr_t[i, a] = 0$, then $\hat{\ell}_t^i(a) = 0$, and so $\mathbb{E}_t[\hat{\ell}_t^i(a)] = 0 \leq \ell_t(a)$. Thus in either case, $\mathbb{E}_t[Y_t^h] \leq \ell_t(a)$, and $\mathbb{E}_t[q_t(h)Y_t^h] = q_t(h)\ell_t(a_t^h)$. Finally, note that if $q_t(h) > 0$, then $\Pr_t[i, a] > 0$, so equality holds. ■

3. The regret bounds only change by a small constant factor if M doesn't divide N.
4. It is easy to see that both the MW and PolyINF forecasters always have positive probability on all experts, so if we use one of these two expert learning algorithms, then all the inequalities in this lemma are actually equalities.

Algorithm 1 Multiarmed Bandits with Limited Expert Advice Algorithm (LEXP).

1: Partition the N experts into $R = N/M$ groups of M experts each arbitrarily. Call the groups $B_1, B_2, \ldots, B_R$, and define $\mathcal{R} := \{1, 2, \ldots, R\}$.
2: Run an algorithm for prediction with expert advice (such as MW or PolyINF) on all the experts.
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Let q_t be the distribution over experts generated by the base expert learning algoithm. Sample an expert $H_t \sim q_t$, set I_t to be the index of the group to which H_t belongs.
5: Query the advice of all experts in B_{I_t}.
6: Play $A_t = a_t^{H_t}$, and observe its loss $\ell_t(A_t)$.
7: For every group B_i and every arm $a \in \mathcal{A}$, define the loss estimator given by

$$\hat{\ell}_t^i(a) \;:=\; \begin{cases} \ell_t^i(a)\frac{\mathbb{I}[I_t=i, A_t=a]}{\Pr_t[i,a]} & \text{if } \Pr_t[i,a] > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\Pr_t[i,a] \;=\; \sum_{h \in B_i} q_t(h)\xi_t^h(a)$ is the probability of the event $\{I_t = i, A_t = a\}$, conditioned on all the randomness up to round $t-1$.
8: For all experts $h \in B_i$, define the loss estimator $Y_t^h := \hat{\ell}_t^i(a_t^h)$, and pass them to the base expert learning algorithm.
9: **end for**

The next lemma says that the algorithm's expected loss in each round is the same as that of the base expert learning algorithm:

Lemma 2 *For all rounds t we have $\mathbb{E}[\ell_t(A_t)] = \mathbb{E}[\sum_{h\in\mathcal{H}} q_t(h)Y_t^h]$.*

Proof We have

$$\mathbb{E}_t[\ell_t(A_t)] \;=\; \mathbb{E}_t[\ell_t(a_t^{H_t})] \;=\; \sum_{h\in\mathcal{H}} q_t(h)\ell_t(a_t^h) \;=\; \mathbb{E}_t[\sum_{h\in\mathcal{H}} q_t(h)Y_t^h],$$

by Lemma 1. Taking expectation over all the randomness up to time $t-1$, the proof is complete. ■

The next lemma gives a bound on the variance of the estimated losses. We state this in slightly more general terms than necessary to unify the analysis of the algorithms using the MW or PolyINF forecasters as the expert learning algorithm.

Lemma 3 *Fix any $\alpha \in [1,2]$. For all rounds t we have*

$$\mathbb{E}[\sum_{h\in\mathcal{H}} (q_t(h))^\alpha (Y_t^h)^2] \;\le\; (RK')^{2-\alpha}.$$

Proof Let

$$S \;:=\; \{(i,a) \in \mathcal{R}\times\mathcal{A} \mid \Pr_t[i,a] > 0\}$$

be the set of all (group index, action) pairs that have positive probability in round t. Since in round t, the algorithm only plays arms in $\mathcal{A}_t^{B_{I_t}}$, and for any group B_i, the set of active arms in round t, $\mathcal{A}_t^{B_i}$, has size at most K', we conclude that $|S| \le RK'$.

The pair (I_t, A_t) computed by the algorithm is in S. Conditioning on the value of (I_t, A_t), we can upper bound $\sum_{h\in\mathcal{H}}(q_t(h))^\alpha (Y_t^h)^2$ as follows:

$$\begin{aligned}
\sum_{h\in\mathcal{H}}(q_t(h))^\alpha (Y_t^h)^2 &= \sum_{h\in B_{I_t}} (q_t(h))^\alpha (\hat{\ell}_t^{I_t}(a_t^h))^2 && (\because Y_t^h = 0 \text{ for all } h \notin B_{I_t}) \\
&= \sum_{h\in B_{I_t}} (q_t(h))^\alpha \left(\xi^h(A_t)\cdot \frac{\ell_t(A_t)}{\Pr_t[I_t, A_t]}\right)^2 && (\because \hat{\ell}_t^{I_t}(a) = 0 \text{ for all } a \neq A_t) \\
&\leq \sum_{h\in B_{I_t}} (q_t(h)\xi^h(A_t))^\alpha \left(\frac{1}{\Pr_t[I_t, A_t]}\right)^2 && (\because \xi^h(A_t), \ell_t(A_t) \in [0,1],\ \alpha \leq 2) \\
&\leq \left(\sum_{h\in B_{I_t}} q_t(h)\xi^h(A_t)\right)^\alpha \left(\frac{1}{\Pr_t[I_t, A_t]}\right)^2 && (\because \|\cdot\|_\alpha \leq \|\cdot\|_1 \text{ since } \alpha \geq 1) \\
&= \Pr_t[I_t, A_t]^{\alpha-2}, && (2)
\end{aligned}$$

since $\Pr_t[I_t, A_t] = \sum_{h\in B_{I_t}} q_t(h)\xi^h(A_t)$. Next, we have

$$\begin{aligned}
\mathbb{E}_t[\sum_{h\in\mathcal{H}}(q_t(h))^\alpha (Y_t^h)^2] &= \mathbb{E}_t[\mathbb{E}_t[\sum_{h\in\mathcal{H}}(q_t(h))^\alpha (Y_t^h)^2 \mid (I_t, A_t)] \\
&\leq \sum_{(I_t,A_t)\in S} \Pr_t[I_t, A_t]\cdot \Pr_t[I_t, A_t]^{\alpha-2} && \text{(By (2))} \\
&= \sum_{(I_t,A_t)\in S} \Pr_t[I_t, A_t]^{\alpha-1} \\
&\leq \left(\sum_{(I_t,A_t)\in S} \Pr_t[I_t, A_t]\right)^{\alpha-1} \cdot \left(\sum_{(I_t,A_t)\in S} 1\right)^{2-\alpha} \\
&= |S|^{2-\alpha} \\
&\leq (RK')^{2-\alpha}.
\end{aligned}$$

The penultimate inequality follows by applying Hölder's inequality to the pair of dual norms $\|\cdot\|_{\frac{1}{\alpha-1}}$ and $\|\cdot\|_{\frac{1}{2-\alpha}}$. Taking expectation over all the randomness up to time $t-1$, the proof is complete. ■

4.1. Analysis using the MW forecaster

The MW forecaster for prediction with expert advice takes one parameter, η. It starts with q_1 being the uniform distribution over all experts, and for any $t \geq 1$, constructs the distribution q_{t+1} using the following update rule:

$$q_{t+1}(h) \;:=\; q_t(h)\exp(-\eta Y_t^h)/Z_t,$$

where Z_t is the normalization constant required to make q_{t+1} a distribution, i.e. $\sum_{h\in\mathcal{H}} q_{t+1}(h) = 1$.

Theorem 1 *Set $\eta = \sqrt{\frac{M\log(N)}{K'NT}}$. Then the expected regret of the algorithm using the MW forecaster is bounded by $\sqrt{\frac{2K'N\log(N)}{M}}T$.*

Proof The MW forecaster guarantees (see (Arora et al., 2012)) that as long as $Y_t^h \geq 0$ for all t, h, we have for any expert $h^\star$

$$\sum_{t=1}^{T} \sum_{h \in \mathcal{H}} q_t(h) Y_t^h \leq \sum_{t=1}^{T} Y_t^{h^\star} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{h \in \mathcal{H}} q_t(h) (Y_t^h)^2 + \frac{\log N}{\eta}. \tag{3}$$

Now, we have for any expert $h^\star$

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\ell_t(A_t)] &= \sum_{t=1}^{T} \mathbb{E}[\sum_{h \in \mathcal{H}} q_t(h) Y_t^h] && \text{(By Lemma 2)} \\
&\leq \sum_{t=1}^{T} \mathbb{E}[Y_t^{h^\star}] + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}[\sum_{h \in \mathcal{H}} q_t(h) (Y_t^h)^2] + \frac{\log N}{\eta} && \text{(By (3))} \\
&\leq \sum_{t=1}^{T} \ell_t(a_t^{h^\star}) + \frac{\eta}{2} RK'T + \frac{\log N}{\eta} \\
&\qquad \text{(By Lemma 1 and Lemma 3 with } \alpha = 1) \\
&\leq \sum_{t=1}^{T} \ell_t(a_t^{h^\star}) + \sqrt{\frac{2K'N \log(N)}{M}} T,
\end{aligned}$$

using $\eta = \sqrt{\frac{2\log(N)}{RK'T}} = \sqrt{\frac{2M\log(N)}{K'NT}}$. ■

4.2. Analysis using the PolyINF forecaster

The PolyINF forecaster for prediction with expert advice takes two parameters, η and $c > 1$. It starts with q_1 being the uniform distribution over all experts, and for any $t \geq 1$, constructs the distribution q_{t+1} as follows:

$$q_{t+1}(h) = \frac{1}{[\eta(\sum_{\tau=1}^{t} Y_\tau^h + C_{t+1})]^c}$$

where C_{t+1} is a constant chosen so that q_{t+1} is a distribution, i.e. $\sum_{h \in \mathcal{H}} q_{t+1}(h) = 1$.

Theorem 2 *Set* $c = \log(\frac{8M}{K'})$ *and* $\eta = 2N^{\frac{1}{2c}}[c(RK')^{1-\frac{1}{c}}T]^{-\frac{1}{2}}$. *Then the expected regret of the algorithm using the PolyINF forecaster is bounded by* $4\sqrt{\frac{K'N\log(\frac{8M}{K'})}{M}}T$.

Proof Audibert et al. (2011) prove that for the PolyINF forecaster, as long as $Y_t^h \geq 0$ for all t, h, we have for any expert $h^\star$:

$$\sum_{t=1}^{T} \sum_{h \in \mathcal{H}} q_t(h) Y_t^h \leq \sum_{t=1}^{T} Y_t^{h^\star} + \frac{c\eta}{2} \sum_{t=1}^{T} \sum_{h \in \mathcal{H}} (q_t(h))^{1+\frac{1}{c}} (Y_t^h)^2 + \frac{cN^{\frac{1}{c}}}{\eta(c-1)}. \tag{4}$$

Now, we have for any expert $h^\star$

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\ell_t(A_t)] &= \sum_{t=1}^{T}\mathbb{E}[\sum_{h\in\mathcal{H}} q_t(h)Y_t^h] && \text{(By Lemma 2)}\\
&\le \sum_{t=1}^{T}\mathbb{E}[Y_t^{h^\star}] + \frac{c\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\sum_{h\in\mathcal{H}}(q_t(h))^{1+\frac{1}{c}}(Y_t^h)^2] + \frac{2N^{\frac{1}{c}}}{\eta} && \text{(By (4), using } c\ge 2)\\
&\le \sum_{t=1}^{T}\ell_t(a_t^{h^\star}) + \frac{c\eta}{2}(RK')^{1-\frac{1}{c}}T + \frac{2N^{\frac{1}{c}}}{\eta}\\
&\qquad \text{(By Lemma 1 and Lemma 3 with } \alpha = 1+\tfrac{1}{c})\\
&\le \sum_{t=1}^{T}\ell_t(a_t^{h^\star}) + 2\sqrt{cRK'\left(\tfrac{N}{RK'}\right)^{\frac{1}{c}}T},\\
&\qquad \text{(Using } \eta = 2N^{\frac{1}{2c}}[c(RK')^{1-\frac{1}{c}}T]^{-\frac{1}{2}})\\
&\le \sum_{t=1}^{T}\ell_t(a_t^{h^\star}) + 4\sqrt{\frac{K'N\log(\frac{8M}{K'})}{M}T},
\end{aligned}$$

using $c = \log(\frac{8M}{K'}) = \log(\frac{8N}{RK'})$. ■

4.3. Extension to Changing Number of Queried Experts

The algorithm and its analysis extends easily to the situation where the number of experts queried is not fixed but can change from round to round. Specifically, at time t, the learner is told the number M_t of experts that can be queried in that round.

In this setting, consider the following variant of the algorithm. In each round t, the experts are re-partitioned into as N/M_t groups[5] of size M_t. The rest of the algorithm stays the same: viz. an expert is chosen from the current probability distribution over the experts, and the group it belongs to is chosen for querying for expert advice. The update to the distribution and the loss estimators are the same as in Algorithm 1.

The analysis of Algorithm 1 relies on Lemmas 1, 2 and 3 all of which concern a specific round t, and the re-partitioning doesn't affect them. Thus, we easily obtain the following bound:

Theorem 3 *In the setting where in each round t the number M_t of experts that can be queried in that round is specified, the extension of Algorithm 1 which re-partitions the experts in each round into $R_t := N/M_t$ groups of size M_t, has the following regret bound. For every round t, let $K'_t = \min\{K, M_t\}$ and $t^* = \arg\max_{t=1}^{T} M_t$. If the MW forecaster is used with $\eta = \sqrt{\frac{\log(N)}{N\sum_{t=1}^{T}K'_t/M_t}}$, then the expected regret is bounded by $\sqrt{\sum_{t=1}^{T}\frac{2K'_tN\log(N)}{M_t}}$. If the PolyINF forecaster is used with $c = \log(\frac{8M_{t^*}}{K'_{t^*}})$ and $\eta = 2N^{\frac{1}{2c}}[c\sum_{t=1}^{T}(R_tK'_t)^{1-\frac{1}{c}}]^{-\frac{1}{2}}$, then the expected regret is bounded by $4\sqrt{\sum_{t=1}^{T}\frac{K'_tN\log(8M_{t^*}/K'_{t^*})}{M_t}}$.*

5. Again, here we assume M_t divides N for convenience.

5. Lower Bound

In this section, we show a lower bound on the regret of any algorithm for the multiarmed bandit with limited expert advice setting which shows that our upper bound is nearly tight. To describe the lower bound, consider the well-studied balls-into-bins process. Here M balls are tossed randomly into K bins. In each toss a bin is chosen uniformly at random from the K bins independently of other tosses. Define the function $f(K, M)$ to be the expected number of balls in the bin with the maximum number of balls. It is well-known (see, for example, Raab and Steger (1998)) that $f(K, M) = O(\max\{\log(K), \frac{M}{K}\})$.

With this definition, we can prove the following lower bound. Note that this lower bound doesn't follow from a similar lower bound in (Seldin et al., 2014) because in their setting the experts' losses can be all uncorrelated, whereas in our setting the experts' losses are necessarily correlated because there are only K arms.

Theorem 4 *For any algorithm for the multiarmed bandit with limited expert advice setting, there is a sequence of expert advice and losses for each arm so that the expected regret of the algorithm is at least* $\Omega\left(\sqrt{\frac{N}{f(K,M)}T}\right) = \Omega\left(\sqrt{\frac{\min\{K, \frac{M}{\log(K)}\}N}{M}T}\right)$.

Proof The lower bound is based on standard information theoretic arguments (see, e.g. (Auer et al., 2002)). Let $\mathbb{B}(p)$ be the Bernoulli distribution with parameter p, i.e. 1 is chosen with probability p and 0 with probability $1 - p$.

In the following, we assume the online algorithm is deterministic: the extension to randomized algorithms is easy by conditioning on the random seed of the algorithm, since the sequence of advice and losses we construct do not depend on the algorithm.

Fix the parameter $\varepsilon := \frac{1}{16}\sqrt{\frac{N}{f(K,M)T}}$. The expert advice and the losses of the arms are generated randomly as follows.We define N probability distributions over advice and losses, $\mathbf{P}_h$ for all $h \in \mathcal{H}$. Fix an $h^\star \in \mathcal{H}$, and define $\mathbf{P}_{h^\star}$ as follows. In each round t, for all experts $h \in \mathcal{H}$, we set their advice to be a uniformly random arm in $\mathcal{A}$. Recall that the arm chosen by expert h in round t is a_t^h. Conditioned on the choice of the arm $a_t^{h^\star}$, the loss of arm $a_t^{h^\star}$ is chosen from $\mathbb{B}(\frac{1}{2} - \varepsilon)$, and the loss of all arms $a \neq a_t^{h^\star}$ from $\mathbb{B}(\frac{1}{2})$, independently. Unconditionally, the distribution of the loss of any arm a at any time t is $\mathbb{B}(p)$ where $p = \frac{1}{K} \cdot \left(\frac{1}{2} - \varepsilon\right) + \frac{K-1}{K} \cdot \frac{1}{2} = \frac{1}{2} - \frac{\varepsilon}{K}$. A similar calculation shows that for all experts $h \neq h^\star$, the distribution of the loss of their chosen arm is $\mathbb{B}(p)$ and thus has expectation p, and the expected loss of the arm chosen by $h^\star$ is $\frac{1}{2} - \varepsilon$. Thus the best expert is $h^\star$. Let $\mathbb{E}_{h^\star}$ denote expectation under $\mathbf{P}_{h^\star}$.

Consider another probability distribution $\mathbf{P}_0$ over advice and losses: in all rounds t, all experts choose their arms in $\mathcal{A}$ uniformly at random as before, and all arms have loss distributed as $\mathbb{B}(p)$. Let $\mathbb{E}_0$ denote the expectation of random variables under $\mathbf{P}_0$.

Before round 1, we choose an expert $h^\star \in \mathcal{H}$ uniformly at random, and advice and losses are then generated from $\mathbf{P}_{h^\star}$. In round t, let S_t denote the set of M experts chosen by the algorithm to query.

Lemma 4 below shows that if either of the events $[h^\star \notin S_t]$ or $[h^\star \in S_t,\ A_t \neq a_t^{h^\star}]$ happens, the algorithm suffers an expected regret of at least $\varepsilon/2$. Define the random variables

$$L_{h^\star} = \sum_{t=1}^{T} \mathbb{I}[h^\star \in S_t] \quad \text{and} \quad N_{h^\star} = \sum_{t=1}^{T} \mathbb{I}[h^\star \in S_t, A_t = a_t^{h^\star}].$$

Then to get a lower bound on the expected regret we need to upper bound $\mathbb{E}_{h^\star}[N_{h^\star}]$. To do this, we use arguments based on KL-divergence between the distributions $\mathbf{P}_{h^\star}$ and $\mathbf{P}_0$. Specifically, for all t, let

$$H_t = \langle (G_1, \ell_1(A_1)), (G_2, \ell_2(A_2)), \ldots, (G_t, \ell_t(A_t)) \rangle$$

denote the history up to time t; here, $G_\tau = \{(h, a_\tau^h) \mid h \in S_\tau\}$ is the set of pairs of experts and their advice for the experts queried at time τ. For convenience, we define $H_0 = \{\}$, the empty set. Note that since the algorithm is assumed to be deterministic, $N_{h^\star}$ is a deterministic function of the history H_T. Thus to upper bound $\mathbb{E}_{h^\star}[N_{h^\star}]$ we compute an upper bound on $\mathrm{KL}(\mathbf{P}_0(H_T) \parallel \mathbf{P}_{h^\star}(H_T))$. Lemma 5 below shows that

$$\mathrm{KL}(\mathbf{P}_0(H_T) \parallel \mathbf{P}_{h^\star}(H_T)) \;\le\; 6\varepsilon^2 \mathbb{E}_0[N_{h^\star}] + \frac{4\varepsilon^2}{K^2}\mathbb{E}_0[L_{h^\star}].$$

Thus, by Pinsker's inequality, we get

$$d_{\mathrm{TV}}(\mathbf{P}_0(H_T), \mathbf{P}_{h^\star}(H_T)) \;\le\; \sqrt{\tfrac{1}{2}\mathrm{KL}(\mathbf{P}_0(H_T) \parallel \mathbf{P}_{h^\star}(H_T))} \;\le\; \sqrt{3\varepsilon^2 \mathbb{E}_0[N_{h^\star}] + \frac{2\varepsilon^2}{K^2}\mathbb{E}_0[L_{h^\star}]}.$$

Since $N_{h^\star} \in [0, T]$, this implies that

$$\mathbb{E}_{h^\star}[N_{h^\star}] \;\le\; \mathbb{E}_0[N_{h^\star}] + T\sqrt{3\varepsilon^2 \mathbb{E}_0[N_{h^\star}] + \frac{2\varepsilon^2}{K^2}\mathbb{E}_0[L_{h^\star}]}.$$

By Jensen's inequality applied to the concave square root function, we get

$$\begin{aligned}
\frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_{h^\star}[N_{h^\star}] \;&\le\; \frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[N_{h^\star}] + T\sqrt{3\varepsilon^2\left[\frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[N_{h^\star}]\right] + \frac{2\varepsilon^2}{K^2}\left[\frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[L_{h^\star}]\right]} \\
&\le\; \frac{f(K,M)}{N}T + T\sqrt{3\varepsilon^2\frac{f(K,M)}{N}T + 2\varepsilon^2\frac{M}{K^2 N}T} \qquad (5) \\
&\le\; \frac{3T}{4} + 2\varepsilon T\sqrt{\frac{f(K,M)}{N}T}. \qquad (6)
\end{aligned}$$

Inequality (5) follows from Lemma 6 below using

$$\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[L_{h^\star}] = \sum_{t=1}^{T}\sum_{h^\star \in \mathcal{H}} \mathbf{P}_0[h^\star \in S_t] \;=\; \sum_{t=1}^{T} \mathbb{E}_0[|S_t|] \;\le\; MT$$

and

$$\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[N_{h^\star}] \;=\; \sum_{t=1}^{T}\sum_{h^\star \in \mathcal{H}} \mathbf{P}_0[h^\star \in S_t, A_t = a_t^{h^\star}] \;\le\; \sum_{t=1}^{T} \mathbb{E}_0[f(K, |S_t|)] \;\le\; f(K,M)T. \qquad (7)$$

To obtain inequality (6), we upper bound $2\varepsilon^2\frac{M}{K^2N}T$ by $\varepsilon^2\frac{f(K,M)}{N}T$ because $f(K,M)$ is at least the expected number of balls in each bin, which equals $\frac{M}{K}$, and so $f(K,M) \ge \frac{2M}{K^2}$ for $K \ge 2$. As for the $\frac{f(K,M)}{N}T$ term, we bound it using the fact that $f(K,M) \le f(2,N)$ for $K \ge 2$ (since f is clearly monotonically decreasing in the first argument and monotonically increasing in the second), and $f(2,N) \le \frac{N+\sqrt{N}}{2} \le \frac{3N}{4}$ for $N \ge 2$.

Now, taking expectation over the choice of the expert $h^\star$, the expected regret of the algorithm is at least

$$\frac{1}{N}\sum_{h^\star\in\mathcal{H}}\frac{\varepsilon}{2}(T-\mathbb{E}_{h^\star}[N_{h^\star}]) \;\geq\; \frac{\varepsilon}{8}T-\varepsilon^2T\sqrt{\frac{f(K,M)}{N}T}$$
$$= \frac{1}{256}\sqrt{\frac{N}{f(K,M)}T} \;=\; \Omega\left(\sqrt{\frac{\min\{K,\frac{M}{\log(K)}\}N}{M}T}\right),$$

using the setting $\varepsilon=\frac{1}{16}\sqrt{\frac{N}{f(K,M)T}}$ and the fact that $f(K,M)=O(\max\{\log(K),\frac{M}{K}\})$. ■

Lemma 4 *Suppose $h^\star$ is the expert chosen in the beginning and advice and losses are then generated from $\mathbf{P}_{h^\star}$. Then in any round t, if either of the events $[h^\star\notin S_t]$ or $[h^\star\in S_t,\ A_t\neq a_t^{h^\star}]$ happens, the algorithm suffers an expected regret of at least $\varepsilon/2$.*

Proof First, recall that the expert $h^\star$ always incurs an expected loss of $\frac{1}{2}-\varepsilon$ in each round t.

Now if $h^\star\notin S_t$, then the losses of the arms are independent of the advice of the experts in S_t, and hence their distribution *conditioned on the advice of experts in S_t* is $\mathbb{B}(p)$. Thus, the distribution of the chosen arm A_t is also $\mathbb{B}(p)$, which implies that the algorithm suffers an expected regret of $p-(\frac{1}{2}-\varepsilon)=\varepsilon(1-1/K)\geq\varepsilon/2$.

If $h^\star\in S_t$ but $A_t\neq a_t^{h^\star}$, then the distribution of the loss of A_t, conditioned on the advice of the experts in S_t, is $\mathbb{B}(\frac{1}{2})$. This implies that the algorithm suffers an expected regret of $\frac{1}{2}-(\frac{1}{2}-\varepsilon)=\varepsilon\geq\varepsilon/2$. ■

Lemma 5 *We have*

$$KL(\mathbf{P}_0(H_T)\parallel\mathbf{P}_{h^\star}(H_T)) \;\leq\; 6\varepsilon^2\mathbb{E}_0[N_{h^\star}]+\frac{4\varepsilon^2}{K^2}\mathbb{E}_0[L_{h^\star}].$$

Proof We have

$$\mathrm{KL}(\mathbf{P}_0(H_T)\parallel\mathbf{P}_{h^\star}(H_T)) \;=\; \sum_{t=1}^{T}\mathrm{KL}(\mathbf{P}_0((G_t,\ell_t(A_t))|H_{t-1})\parallel\mathbf{P}_{h^\star}((G_t,\ell_t(A_t))|H_{t-1})) \tag{8}$$
$$= \sum_{t=1}^{T}[\mathrm{KL}(\mathbf{P}_0(\ell_t(A_t)|H_{t-1},G_t)\parallel\mathbf{P}_{h^\star}(\ell_t(A_t)|H_{t-1},G_t))$$
$$+\mathrm{KL}(\mathbf{P}_0(G_t|H_{t-1})\parallel\mathbf{P}_{h^\star}(G_t|H_{t-1}))] \tag{9}$$
$$= \sum_{t=1}^{T}\mathrm{KL}(\mathbf{P}_0(\ell_t(A_t)|H_{t-1},G_t)\parallel\mathbf{P}_{h^\star}(\ell_t(A_t)|H_{t-1},G_t)) \tag{10}$$
$$= \sum_{t=1}^{T}\mathbf{P}_0[h^\star\in S_t,A_t=a_t^{h^\star}]\mathrm{KL}(\mathbb{B}(p)\parallel\mathbb{B}(\tfrac{1}{2}-\varepsilon))$$
$$+\mathbf{P}_0[h^\star\in S_t,A_t\neq a_t^{h^\star}]\mathrm{KL}(\mathbb{B}(p)\parallel\mathbb{B}(\tfrac{1}{2}))$$
$$+\mathbf{P}_0[h^\star\notin S_t]\mathrm{KL}(\mathbb{B}(p)\parallel\mathbb{B}(p)) \tag{11}$$

$$\leq \sum_{t=1}^{T} \mathbf{P}_0[h^\star \in S_t, A_t = a_t^{h^\star}] \cdot 6\varepsilon^2 + \mathbf{P}_0[h^\star \in S_t, A_t \neq a_t^{h^\star}] \cdot \frac{4\varepsilon^2}{K^2} \tag{12}$$

$$\leq \sum_{t=1}^{T} 6\varepsilon^2 \mathbf{P}_0[h^\star \in S_t, A_t = a_t^{h^\star}] + \frac{4\varepsilon^2}{K^2} \mathbf{P}_0[h^\star \in S_t]$$

$$= 6\varepsilon^2 \mathbb{E}_0[N_{h^\star}] + \frac{4\varepsilon^2}{K^2} \mathbb{E}_0[L_{h^\star}].$$

Equalities (8) and (9) follow from the chain rule for relative entropy. Equality (10) follows because the distribution of G_t conditioned on H_{t-1} is identical in $\mathbf{P}_0$ and $\mathbf{P}_{h^\star}$. Equality (11) follows under $\mathbf{P}_0$, the loss of the chosen arm always follows $\mathbb{B}(p)$, and under $\mathbf{P}_{h^\star}$, if $h^\star \notin S_t$, then the loss of the chosen arm follows $\mathbb{B}(p)$, if $h^\star \in S_t$ and $A_t = a_t^{h^\star}$, then the loss of the chosen arm follows $\mathbb{B}(\frac{1}{2} - \varepsilon)$, and if $h^\star \in S_t$ and $A_t \neq a_t^{h^\star}$, then the loss of the chosen arm follows $\mathbb{B}(\frac{1}{2})$. Finally, inequality (12) follows using standard calculations for KL-divergence between Bernoulli random variables. ■

Recall that $f(K, M)$ is the expected number of balls in the bin with the maximum balls in a M-balls-into-K-bins process.

Lemma 6 *For all t, we have*

$$\sum_{h^\star \in \mathcal{H}} \mathbf{P}_0[h^\star \in S_t] = \mathbb{E}_0[|S_t|] \quad \textit{and} \quad \sum_{h^\star \in \mathcal{H}} \mathbf{P}_0[h^\star \in S_t, A_t = a_t^{h^\star}] \leq \mathbb{E}_0[f(K, |S_t|)].$$

Proof First, we have

$$\sum_{h^\star \in \mathcal{H}} \mathbf{P}_0[h^\star \in S_t] = \mathbb{E}_0 \left[\sum_{h^\star \in \mathcal{H}} \mathbb{I}[h^\star \in S_t] \right] = \mathbb{E}_0[|S_t|].$$

Next, we have

$$\sum_{h^\star \in \mathcal{H}} \mathbf{P}_0[h^\star \in S_t, A_t = a_t^{h^\star}] = \mathbb{E}_0 \left[\sum_{h^\star \in \mathcal{H}} \mathbb{I}[h^\star \in S_t, A_t = a_t^{h^\star}] \right] = \mathbb{E}_0 \left[|\{h^\star \in S_t : \ A_t = a_t^{h^\star}\}| \right]$$

$$\leq \mathbb{E}_0 \left[\max_{a \in \mathcal{A}} \{ |\{h^\star \in S_t : \ a = a_t^{h^\star}\}| \} \right] = \mathbb{E}_0 \left[\mathbb{E}_0 \left[\max_{a \in \mathcal{A}} \{ |\{h^\star \in S_t : \ a = a_t^{h^\star}\}| \} \mid S_t \right] \right]$$

$$= \mathbb{E}_0[f(K, |S_t|)].$$

The penultimate equality follows because conditioning on the choice of S_t, the random variable $\max_{a \in \mathcal{A}} \{ |\{h^\star \in S_t : \ a = a_t^{h^\star}\}| \}$ is completely determined by the choice of the arms recommended by the experts $h^\star \in S_t$. Since these arms are chosen uniformly at random from $\mathcal{A}$ independently for each expert $h^\star \in S_t$, we can think of the $|S_t|$ experts in S_t as "balls" and the K arms in $\mathcal{A}$ as "bins" in a balls-into-bins process. Then the random variable of interest is exactly the number of balls in the bin with maximum number of balls. The expectation of this random variable is $f(K, |S_t|)$. ■

5.1. Extension to Global Limit on Queries

In certain situations a global limit on the number of queries made to experts over the entire run of the algorithm, rather than a per-round limit, is more natural. Then the analysis of Theorem 4 can be extended easily to give the following theorem (proved in Appendix A):

Theorem 5 *In the setting of the multiarmed bandits with limited expert advice problem where there is a global limit of MT queries to experts over the T rounds, for any algorithm, there is a sequence of expert advice and losses for each arm so that the expected regret of the algorithm is at least* $\Omega\left(\sqrt{\frac{\min\{K, \frac{M}{\log(K)}\}N}{M}T}\right)$.

This shows that the up to logarithmic factors, the optimal allocation of queries over the rounds is the uniform allocation of M queries per round.

5.2. Extension to Changing Number of Queried Experts

The lower bound also extends to the setting of Section 4.3 where in each round t, the learner is told the number of experts that can be queried, M_t. The analysis is basically the same with a few modifications to handle the changing number of experts to be queried. In Appendix A, we prove the following theorem:

Theorem 6 *For any algorithm working in the setting where the algorithm is told the number of experts M_t that can be queried in each round t, there is a sequence of expert advice and losses for each arm so that the expected regret of the algorithm is at least* $\Omega\left(\sqrt{\sum_{t=1}^{T}\frac{N}{f(K,M_t)}}\right) = \Omega\left(\sqrt{\sum_{t=1}^{T}\frac{\min\{K, \frac{M_t}{\log(K)}\}N}{M_t}}\right)$.

6. Conclusions

In this paper, we presented near-optimal algorithms for the multiarmed bandits with limited expert advice problem, solving the COLT 2013 open problem of Seldin et al. (2013). The upper bound uses a novel grouping idea combined with a standard experts learning algorithm, whereas the lower bound uses an information-theoretic approach and a connection to the classic ball-into-bins problem to get a nearly-tight dependence on the problem parameters. The binning strategy might be useful in other contexts such as settings where there may be non-uniform cost associated with the advice for each expert. An interesting open question is to close the sub-logarithmic gap between the upper and lower bounds.

Acknowledgments

The author thanks Elad Hazan, Dean Foster, Rob Schapire, and Yevgeny Seldin for discussions on this problem.

References

Sanjeev Arora, Elad Hazan, and Satyen Kale. The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. *Theory of Computing*, 8(1):121–164, 2012.

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

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. *Journal of Machine Learning Research - Proceedings Track*, 19:107–132, 2011.

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

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2): 212–261, 1994.

Martin Raab and Angelika Steger. "Balls into Bins" - A Simple and Tight Analysis. In *RANDOM*, pages 159–170, 1998.

Yevgeny Seldin, Koby Crammer, and Peter Bartlett. Open Problem: Adversarial Multiarmed Bandits with Limited Advice. In *COLT*, 2013.

Yevgeny Seldin, Peter L. Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *ICML*, 2014.

Appendix A. Proofs of Extensions to Lower Bounds

In this section, we provide missing proofs for extensions to lower bounds on the regret.

A.1. Global Limit on Queries: Proof of Theorem 5.

First, note that since $f(K,M) = O(\max\{\log(K), \frac{M}{K}\})$, we have that $f(K,M) \le g(K,M) := c(\log(K) + \frac{M}{K})$ for some constant c. Note that g is linear in its second argument M (as opposed to f) so it is easier to manipulate.

We use the exact same construction of expert advice and losses as in the proof of Theorem 4, with the choice of $\varepsilon = \frac{1}{16}\sqrt{\frac{N}{g(K,M)T}}$. The only change that needs to be made to the proof is in inequality (7), which now becomes

$$\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[N_{h^\star}] \le \sum_{t=1}^{T} \mathbb{E}_0[f(K,|S_t|)] \le \sum_{t=1}^{T} \mathbb{E}_0[g(K,|S_t|)] = \mathbb{E}_0\left[\sum_{t=1}^{T} g(K,|S_t|)\right] \le g(K,M)T.$$

Since, as proved in the paragraph after inequality (7), we have $f(K,M) \le \frac{3N}{4}$ for all $M \le N$, we also have that

$$\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[N_{h^\star}] \le \sum_{t=1}^{T} \mathbb{E}_0[f(K,|S_t|)] \le \frac{3N}{4}T.$$

Using these two bounds, we can now derive the following analogue of inequality (6):

$$\frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_{h^\star}[N_{h^\star}] \le \frac{3T}{4} + 2\varepsilon T\sqrt{\frac{g(K,M)}{N}T}.$$

The rest of the analysis goes through just as before, and yields a regret lower bound of $\frac{1}{256}\sqrt{\frac{N}{g(K,M)}T} = \Omega\left(\sqrt{\frac{\min\{K, \frac{M}{\log(K)}\}N}{M}T}\right)$.

A.2. Changing Number of Queried Experts: Proof of Theorem 6.

We use the essentially the same construction as in the proof of Theorem 4 but with one important twist: the ε controlling the loss of the best expert changes in each round. Specifically, in round t, we set

$$\varepsilon_t = \frac{N/f(K,M_t)}{16\sqrt{\sum_{\tau=1}^{T} N/f(K,M_\tau)}}$$

and the loss of the arm $a_t^{h^\star}$ is chosen from $\mathbb{B}(\frac{1}{2} - \varepsilon_t)$. The losses of all other arms $a \ne a_t^{h^\star}$ are chosen from $\mathbb{B}(\frac{1}{2})$ as before.

We now turn to the analysis. First, we note that since Lemma 4 gives a lower bound on expected regret in specific rounds, summing over all rounds, we conclude that the expected regret of the algorithm is at least

$$\sum_{t=1}^{T} \frac{\varepsilon_t}{2}(1 - \mathbb{I}[h^\star \in S_t, A_t = a_t^{h^\star}]).$$

Thus, define the random variable

$$G_{h^\star} = \sum_{t=1}^{T} \frac{\varepsilon_t}{2} \mathbb{I}[h^\star \in S_t, A_t = a_t^{h^\star}].$$

To get a lower bound on regret, we need to upper bound this random variable. Next, because ε_t changes in different rounds, we need to consider slightly different random variables:

$$L'_{h^\star} = \sum_{t=1}^{T} \varepsilon_t^2 \mathbb{I}[h^\star \in S_t] \quad \text{and} \quad N'_{h^\star} = \sum_{t=1}^{T} \varepsilon_t^2 \mathbb{I}[h^\star \in S_t, A_t = a_t^{h^\star}].$$

With this definition, the statement of Lemma 5 extends easily to the following:

$$\mathrm{KL}(\mathbf{P}_0(H_T) \,\|\, \mathbf{P}_{h^\star}(H_T)) \;\le\; 6\mathbb{E}_0[N'_{h^\star}] + \frac{4}{K^2}\mathbb{E}_0[L'_{h^\star}].$$

Define $U = \sum_{t=1}^{T} \frac{\varepsilon_t}{2}$. Continuing the analysis as in the proof of Theorem 4, using Pinsker's inequality, the fact that $G_{h^\star} \in [0, U]$, and Jensen's inequality applied to the concave square root function to conclude that

$$\begin{aligned}
\frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_{h^\star}[G_{h^\star}] \;&\le\; \frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[G_{h^\star}] + U\sqrt{3\left[\frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[N'_{h^\star}]\right] + \frac{2}{K^2}\left[\frac{1}{N}\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[L'_{h^\star}]\right]} \\
&\le \sum_{t=1}^{T} \frac{\varepsilon_t}{2} \cdot \frac{f(K, M_t)}{N} + U\sqrt{\sum_{t=1}^{T} 3\varepsilon_t^2 \frac{f(K, M_t)}{N} + \sum_{t=1}^{T} 2\varepsilon_t^2 \frac{M}{K^2 N}} \qquad (13) \\
&\le \frac{3U}{4} + 2U\sqrt{\sum_{t=1}^{T} \varepsilon_t^2 \frac{f(K, M_t)}{N}}. \qquad (14)
\end{aligned}$$

Inequality (13) follows from Lemma 6 using the following bounds:

$$\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[L'_{h^\star}] = \sum_{t=1}^{T}\sum_{h^\star \in \mathcal{H}} \varepsilon_t^2 \mathbf{P}_0[h^\star \in S_t] \;=\; \sum_{t=1}^{T} \varepsilon_t^2 \mathbb{E}_0[|S_t|] \;\le\; \sum_{t=1}^{T} \varepsilon_t^2 M_t,$$

$$\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[N'_{h^\star}] \;=\; \sum_{t=1}^{T}\sum_{h^\star \in \mathcal{H}} \varepsilon_t^2 \mathbf{P}_0[h^\star \in S_t, A_t = a_t^{h^\star}] \;\le\; \sum_{t=1}^{T} \varepsilon_t^2 \mathbb{E}_0[f(K, |S_t|)] \;\le\; \sum_{t=1}^{T} \varepsilon_t^2 f(K, M_t),$$

and

$$\sum_{h^\star \in \mathcal{H}} \mathbb{E}_0[G_{h^\star}] \;=\; \sum_{t=1}^{T}\sum_{h^\star \in \mathcal{H}} \frac{\varepsilon_t}{2} \mathbf{P}_0[h^\star \in S_t, A_t = a_t^{h^\star}] \;\le\; \sum_{t=1}^{T} \frac{\varepsilon_t}{2} \mathbb{E}_0[f(K, |S_t|)] \;\le\; \sum_{t=1}^{T} \frac{\varepsilon_t}{2} f(K, M_t).$$

Inequality (14) follows from the bound $f(K, M_t) \ge \frac{2M_t}{K^2}$ for $K \ge 2$, and the bound $f(K, M_t) \le \frac{3N}{4}$ for $N \ge 2$. Finally, taking expectation over the choice of the expert $h^\star$, the expected regret of the

algorithm is at least

$$\frac{1}{N}\sum_{h^\star\in\mathcal{H}}(U-G_{h^\star}) \;\geq\; \frac{U}{4} - 2U\sqrt{\sum_{t=1}^{T}\varepsilon_t^2\frac{f(K,M_t)}{N}}$$
$$= \frac{1}{256}\sqrt{\sum_{t=1}^{T}\frac{N}{f(K,M_t)}} \;=\; \Omega\left(\sqrt{\sum_{t=1}^{T}\frac{\min\{K,\frac{M_t}{\log(K)}\}N}{M_t}}\right),$$

using the definition of ε_t. This gives us the required lower bound on the expected regret.