Prediction with Limited Advice and Multiarmed Bandits with Paid Observations

Yevgeny Seldin YEVGENY.SELDIN@GMAIL.COM
Queensland University of Technology and UC Berkeley

Peter Bartlett BARTLETT@EECS.BERKELEY.EDU
UC Berkeley and Queensland University of Technology

Koby Crammer KOBY@EE.TECHNION.AC.IL
The Technion

Yasin Abbasi-Yadkori YASIN.ABBASI@GMAIL.COM
Queensland University of Technology and UC Berkeley

Abstract

We study two problems of online learning under restricted information access. In the first problem, *prediction with limited advice*, we consider a game of prediction with expert advice, where on each round of the game we query the advice of a subset of M out of N experts. We present an algorithm that achieves $O\left(\sqrt{\frac{N}{M}T\ln N}\right)$ regret on T rounds of this game. The second problem, the *multiarmed bandit with paid observations*, is a variant of the adversarial N-armed bandit game, where on round t of the game we can observe the reward of any number of arms, but each observation has a cost c. We present an algorithm that achieves $O\left((cN\ln N)^{1/3}T^{2/3}+\sqrt{T\ln N}\right)$ regret on T rounds of this game in the worst case. Furthermore, we present a number of refinements that treat arm- and time-dependent observation costs and achieve lower regret under benign conditions. We present lower bounds that show that, apart from the logarithmic factors, the worst-case regret bounds cannot be improved.

1. Introduction

We study two problems of online learning under restricted information access. The first problem is a variation of the game of prediction with expert advice (see, for example, (Cesa-Bianchi & Lugosi, 2006)), which we call *prediction with limited advice*. In this game, the player has access to a set of N experts, but on each round of the game is allowed to query the advice of only M of the N experts. This game corresponds, for example, to a situation where each expert is a computationally-expensive function and there is a constraint on the response time. Because of the constraint, it may be possible to compute only a subset of M of the N functions. We provide an algorithm for this setting that achieves $O\left(\sqrt{\frac{N}{M}T\ln N}\right)$ regret on T rounds and a matching lower bound (up to logarithmic factors).

We note that there is a tight connection between prediction with limited advice and multiarmed bandits. In particular, if we ask for the advice of just one expert on every round (meaning that $M=1$), the problem of prediction with limited advice becomes equivalent to an N-armed bandit problem (we can treat each expert as an arm). Furthermore, if $M>1$ and we restrict the algorithm to follow the advice of one of the M experts (rather than playing some function of the advice) the problem is equivalent to an N-armed bandit where we play one arm and are allowed to observe the reward of $M-1$ additional arms. As M grows from 1 to N the game of prediction with limited advice interpolates between a limited-feedback game and a full-information game. Our regret bound provides an interpolation between the $O\left(\sqrt{T\ln N}\right)$ regret bound for full-information games and the $O\left(\sqrt{NT}\right)$ regret bound for bandit games (Cesa-Bianchi & Lugosi, 2006; Audibert & Bubeck, 2010).

The second question studied in this work considers a different type of restriction on information acquisition. We

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

define a variation of the adversarial N-armed bandit game, which we call the *multiarmed bandit with paid observations*. On each round of this game, the player pulls one arm and suffers the loss of that arm, but that loss is not necessarily observed. The player has the option to request to observe the loss of any number of arms, but the cost of each observation is c and it is added to the loss of the player. As a motivational example, we can think about a problem of signing annual contracts with service providers. For instance, imagine a medical insurance company that every year signs an annual contract with a hospital for some set of medical services. The insurance company can choose to order a follow-up survey of service quality in any number of hospitals from an independent inspection body, but each inspection will be associated with inspection cost c. The goal is, of course, to maximize service quality and minimize the cost of inspections. We derive an algorithm for this setting that ensures that the regret (that is, the observation costs plus the excess loss over the loss of the best fixed arm in hindsight) is $O\left((cN\ln N)^{1/3}T^{2/3}+\sqrt{T\ln N}\right)$. Note that we achieve a smooth transition between prediction with expert advice (which corresponds to zero observations cost, $c=0$, since when the cost is zero we can observe all arms for free) and the harder game with $c>0$. For $c>0$ sublinear regret is achieved by gradual decrease of exploration (the number of observations made), eventually getting into a regime where no observations are made on some rounds. We also provide a matching lower bound (up to logarithmic factors). Furthermore, we present a refined algorithm that handles arm- and time-dependent observations costs and reduces the regret under benign conditions.

1.1. Related Work

Our work is not the first attempt to investigate what happens between full-information and limited-feedback games. Mannor & Shamir (2011) provided an alternative approach. Specifically, they considered an N-armed bandit game, where at each round there is a graph and the actions correspond to the nodes of this graph. When playing a node in the graph the player observes the reward of the node played and the rewards of all adjacent nodes in the graph. The work of Mannor and Shamir was further simplified, improved, and generalized by Alon et al. (2013). The main difference between this line of work and our work is that we allow complete freedom in the choice of observations to make (in prediction with limited advice the only restriction is the number of observations and in multiarmed bandits with paid observations there are no restrictions at all).

Other related work is that of Avner et al. (2012) on "decoupling exploration and exploitation". Avner et. al. studied a multiarmed bandit game, where on each round the player is allowed to play one arm and to observe the reward of one arm, but not necessarily the same arm that was played. Although coming from a different motivation, our work (especially Theorem 3) can be seen as a generalization of the work of Avner et. al. along two dimensions. First, we allow any number of observations on every round (including none) rather than making exactly one observation and, second, we take the cost of observations into account. Our work can also be seen as a generalization of label-efficient prediction (Cesa-Bianchi et al., 2005) and label-efficient bandits (Ottucsák & György, 2006; Audibert & Bubeck, 2010). In label-efficient prediction the player can choose to observe the loss of all arms or nothing and in label-efficient bandits the player can choose to observe the loss of the arm played or nothing and learning is done under a constraint on the total number of observations that can be made throughout the game. In our formulation the observed arm(s) does not have to be the one played and there may be any number of observations per round, which allows to improve the exploration strategy.

The effect of an information acquisition cost appears implicitly in locally non-observable partial monitoring games (Bartók et al., 2011; Foster & Rakhlin, 2012). Roughly speaking, local non-observability means that two desirable actions differ in their loss but are identical in their feedback, so that it is necessary to play a third action with higher loss in order to obtain information. Similarly to the results in locally non-observable partial monitoring games, our regret bound for the multiarmed bandit with paid observations scales as $T^{2/3}$. We note that casting multiarmed bandit with paid observations game as a partial monitoring game leads to an exponential increase of the size of the action set (since we have to consider all possible subsets of actions for the observation requests) and, as a result, suboptimal regret bounds.

Zolghadr et al. (2013) have recently introduced the *online probing* game. In online probing the learner has to predict labels of feature vectors when there is a cost for observing entries of the feature vectors and for observing the labels at the end of each prediction round. This game shares with our work the spirit of an online game with restricted information access. We believe that it will be possible to make mutual transfer of ideas in future work.

2. Main Results

In this section, we provide formal definitions of the games and present our main results. We start with prediction with limited advice in Section 2.1 and then present multiarmed bandits with paid observations in Section 2.2. More illuminating proofs are provided in Section 3, whereas more technical results are provided in the appendix.

2.1. Prediction with Limited Advice

The definition of this game is based on the setting of prediction with expert advice described in Cesa-Bianchi & Lugosi (2006). We denote the action space by $\mathcal{X}$, the outcome space by $\mathcal{Y}$, and the loss function by $\ell : \mathcal{X} \times \mathcal{Y} \to [0,1]$ (for our analysis there is no need to assume that the loss is convex in the first parameter). The number of experts is denoted by N and the experts are indexed by $h \in \{1,\ldots,N\}$. On each round t of the game, each expert h produces a piece of advice $\xi_t^h \in \mathcal{X}$, which is not necessarily observed by the player. The player gets a budget $1 \le M_t \le N$ and picks a subset $\mathcal{O}_t \subseteq \{1,\ldots,N\}$ of M_t experts and observes their advice. The player then plays an action $X_t \in \mathcal{X}$, the environment reveals an outcome $y_t \in \mathcal{Y}$, the player suffers a loss $L_t = \ell(X_t, y_t)$, and the experts suffer losses $\ell_t^h = \ell(\xi_t^h, y_t)$. The player observes the losses of all experts in $\mathcal{O}_t$, but gets no information on the losses of experts that are not in $\mathcal{O}_t$. We study the problem in a slightly restricted setting, where the player has to follow the advice of one of the experts (rather than playing some function of the experts advice).

Prediction with Limited Advice Game

For $t = 1, 2, \ldots$:

1. The algorithm gets M_t and plays $(H_t, \mathcal{O}_t)$, such that $\mathcal{O}_t \subseteq \{1,\ldots,N\}$ and $|\mathcal{O}_t| = M_t$ and $H_t \in \mathcal{O}_t$.
2. The environment reveals ℓ_t^h for $h \in \mathcal{O}_t$ and the algorithm suffers the loss $\ell_t^{H_t}$.

We emphasize that in this game the number of observations M_t is provided externally to the algorithm and it is assumed that $M_t \ge 1$. We evaluate the performance of algorithms by their regret defined as

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} L_t\right] - \min_h \left\{\sum_{t=1}^{T} \ell_t^h\right\}.$$

Our first result is an anytime (i.e., independent of the time horizon) algorithm for prediction with limited advice (see Algorithm 1) and a corresponding bound on its regret in Theorem 1. We note that for $M = N$ the algorithm recovers the exponentially weighted average forecaster algorithm for prediction with expert advice (Cesa-Bianchi & Lugosi, 2006) (with the restriction that we have to follow the advice of one expert rather than a linear combination of expert advice) and for $M = 1$ it recovers the EXP3 algorithm for multiarmed bandits (Auer et al., 2002a; Bubeck & Cesa-Bianchi, 2012). The proof of the theorem is provided in Section 3.

Theorem 1. *For any non-increasing positive sequence $\eta_1 \ge \eta_2 \ge \cdots > 0$ the expected regret of Algorithm 1 against an oblivious adversary satisfies:*

$$R_T \le \frac{N}{2}\sum_{t=1}^{T}\frac{\eta_t}{M_t} + \frac{\ln N}{\eta_T}.$$

In particular, for $\eta_t = \sqrt{\frac{\ln N}{N\sum_{s=1}^{t}\frac{1}{M_s}}}$ we have:

$$R_T \le 2\sqrt{N\left(\sum_{t=1}^{T}\frac{1}{M_t}\right)\ln N}.$$

If $M_t = M$ is constant and $\eta_t = \sqrt{\frac{M\ln N}{tN}}$ then:

$$R_T \le 2\sqrt{\frac{N}{M}T\ln N}.$$

Algorithm 1 Prediction with limited advice.

Remark: η_t is defined in Theorem 1.

Input: $M_1, M_2, \ldots$, such that $M_t \in \{1,\ldots,N\}$.

$\forall h$: $\hat{L}_0(h) = 0$.

for $t = 1, 2, \ldots$ **do**

Let

$$q_t(h) = \frac{e^{-\eta_t \hat{L}_{t-1}(h)}}{\sum_{h'} e^{-\eta_t \hat{L}_{t-1}(h')}}.$$

Draw one expert H_t according to q_t. Get advice $\xi_t^{H_t}$.

Sample $M_t - 1$ additional experts uniformly without replacement. Denote by $\mathcal{O}_t$ the set of sampled experts (we have $H_t \in \mathcal{O}_t$ and the cardinality of $\mathcal{O}_t$ is M_t) and let $\mathbb{1}_t^h = \mathbb{1}_{\{h\in\mathcal{O}_t\}}$.

Play $X_t = \xi_t^{H_t}$.

Observe outcome y_t and suffer loss $L_t = \ell(X_t, y_t)$.

$$\forall h : L_t^h = \frac{\ell_t^h}{q_t(h) + (1 - q_t(h))\frac{M_t - 1}{N-1}}\mathbb{1}_t^h. \quad (1)$$

$$\forall h : \hat{L}_t(h) = \sum_{s=1}^{t} L_s^h.$$

end for

The "price" that we pay for observing the advice of M rather than all N experts is the multiplicative $\sqrt{\frac{N}{M}}$ term. In Theorem 2 we provide a matching lower bound, showing that this price is inevitable without additional assumptions. For convenience, Theorem 2 is stated for N-armed bandits, where we are allowed to make arbitrary observations. Theorem 2 holds for any algorithm that makes MT observations throughout the game, no matter how they are distributed. In particular, the number of observations can

depend on the past observations and the number of observations on some rounds can be zero. We note that M in Theorem 2 does not have to be an integer, it is only the product MT that is assumed to be integer. The proof of the theorem is provided in the appendix.

Theorem 2. *For the N-armed bandit game with MT observed rewards and $T \geq \frac{3}{16}\frac{N}{M}$,*

$$\inf \sup R_T \geq 0.03\sqrt{\frac{N}{M}T},$$

where the infimum is over all playing strategies and the supremum is over all oblivious adversaries.

2.2. Multiarmed Bandits with Paid Observations

In order to stress the relation with the game of prediction with limited advice, we use N to denote the number of arms in the multiarmed bandit and $h \in \{1, \ldots, N\}$ to index the arms. On every round of the game, the algorithm gets a vector of non-negative costs $c_t(1), \ldots, c_t(N)$ for observing the outcomes of the arms. The algorithm then plays one arm, denoted by H_t, and can request to observe the rewards of any subset of arms $\mathcal{O}_t \subseteq \{1, \ldots, N\}$. The cost of observations $\sum_{h \in \mathcal{O}_t} c_t(h)$ is added to the regret of the algorithm. The set of observed arms $\mathcal{O}_t$ can be empty. Even when $\mathcal{O}_t$ is not empty, H_t does not have to be in $\mathcal{O}_t$ (in other words, even when we make observations we are not obliged to observe the outcome of the arm that we played). Formally, the game proceeds as follows.

Multiarmed Bandits with Paid Observations Game

For $t = 1, 2, \ldots$:

1. The algorithm observes $c_t(1), \ldots, c_t(N)$.
2. The algorithm plays $(H_t, \mathcal{O}_t)$, such that $H_t \in \{1, \ldots, N\}$ and $\mathcal{O}_t \subseteq \{1, \ldots, N\}$.
3. The environment reveals ℓ_t^h for $h \in \mathcal{O}_t$ and the algorithm suffers the loss $\ell_t^{H_t} + \sum_{h \in \mathcal{O}_t} c_t(h)$, where $\ell_t^{H_t}$ is not necessarily observed.

We emphasize that in this game the number of observations is chosen by the algorithm and it may be equal to zero. In Algorithm 2 box we present an algorithm for this problem, which is analyzed in Theorem 3 (the proof is proved in Section 3). We use

$$R_T^c = \mathbb{E}\left[\sum_{t=1}^T L_t\right] + \mathbb{E}\left[\sum_{t=1}^T \sum_{h \in \mathcal{O}_t} c_t(h)\right] - \min_h \left\{\sum_{t=1}^T \ell_t^h\right\}$$

to denote cost-sensitive regret.

Theorem 3. *For any non-increasing positive sequence $\eta_1 \geq \eta_2 \geq \cdots > 0$ the regret of Algorithm 2 against an oblivious adversary satisfies:*

$$R_T^c \leq \sum_{t=1}^T \left(\frac{\eta_t}{2} + \sqrt{2\eta_t} \sum_{h=1}^N \sqrt{q_t(h)c_t(h)}\right) + \frac{\ln N}{\eta_T}.$$

In particular, if

$$\eta_t = \frac{1}{\left(\frac{\sqrt{\sum_{h=1}^N c_t(h)} + \sum_{s=1}^{t-1}\sum_{h=1}^N \sqrt{q_s(h)c_s(h)}}{\frac{\sqrt{2}}{3}\ln N}\right)^{2/3} + \sqrt{\frac{t}{\ln N}}}$$

is a non-increasing sequence we have:

$$R_T^c \leq (32 \ln N)^{1/3} \left(\sqrt{\sum_{h=1}^N c_T(h)} + \sum_{t=1}^{T-1}\sum_{h=1}^N \sqrt{q_t(h)c_t(h)}\right)^{2/3} + 2\sqrt{T \ln N}. \quad (2)$$

For

$$\eta_t = \frac{1}{\left(\frac{\sum_{s=1}^t \sqrt{\sum_{h=1}^N c_s(h)}}{\frac{\sqrt{2}}{3}\ln N}\right)^{2/3} + \sqrt{\frac{t}{\ln N}}}$$

(which is always a non-increasing sequence) we have:

$$R_T^c \leq (32 \ln N)^{1/3} \left(\sum_{t=1}^T \sqrt{\sum_{h=1}^N c_t(h)}\right)^{2/3} + 2\sqrt{T \ln N}. \quad (3)$$

Algorithm 2 Multiarmed Bandits with Paid Observations.

Remark: η_t is defined in Theorem 3.

$\forall h$: $\hat{L}_0(h) = 0$.

for $t = 1, 2, \ldots$ **do**

$$\forall h : q_t(h) = \frac{e^{-\eta_t \hat{L}_{t-1}(h)}}{\sum_{h'} e^{-\eta_t \hat{L}_{t-1}(h')}}.$$

Draw action H_t according to q_t and play it.

$$\forall h : p_t(h) = \min\left\{1, \sqrt{\frac{\eta_t q_t(h)}{2c_t(h)}}\right\}.$$

For each h query the loss of h with probability $p_t(h)$. Let $\mathbb{1}_t^h = 1$ if the loss of h was observed and $\mathbb{1}_t^h = 0$ otherwise.

$$\forall h : L_t^h = \frac{\ell_t^h}{p_t(h)}\mathbb{1}_t^h.$$

$$\forall h : \hat{L}_t(h) = \sum_{s=1}^t L_s^h.$$

end for

Finally, if the cost of observations is uniform over the arms and game rounds ($c_t(h) = c$ for all t and h) the first regret bound simplifies to

$$R_T^c \leq (32c \ln N)^{1/3} \left(\sqrt{N} + \sum_{t=1}^{T-1} \sum_{h=1}^{N} \sqrt{q_t(h)} \right)^{2/3} + 2\sqrt{T \ln N} \tag{4}$$

(in this case η_t is always a non-increasing sequence) and the second regret bound simplifies to

$$R_T^c \leq (32cN \ln N)^{1/3} T^{2/3} + 2\sqrt{T \ln N}. \tag{5}$$

The constant in the regret bounds satisfies $(32)^{1/3} < 3.2$. Note that by Jensen's inequality $\sum_{h=1}^{N} \sqrt{q_t(h) c_t(h)} \leq \sqrt{\sum_{h=1}^{N} c_t(h)}$, and so bound (2) (when it holds) is tighter than (3) and bound (4) is tighter than (5). If some arm h^* dominates all other arms, asymptotically the distribution q_t converges to a delta distribution on h^* and $\sum_{t=1}^{T} \sum_{h=1}^{N} \sqrt{q_t(h) c_t(h)}$ converges to $\sum_{t=1}^{T} \sqrt{c_t(h^*)}$ and for the uniform costs $\sum_{t=1}^{T} \sqrt{q_t(h)}$ converges to T. In such case bound (2) improves the dependence on the cost of observations of suboptimal arms and (4) improves the dependence on the number of arms (compared to (3) and (5), respectively).

In general, the first term of the regret bounds in Theorem 3 is dominating, unless the cost of observations is very small. When the cost of observations is very small the second term of the regret bounds dominates. For the uniform cost setting in Eq. (5) the domination switchover occurs at $c = \frac{2}{9N} \sqrt{\frac{\ln N}{T}}$. At the extreme of zero cost of observations the algorithm and the regret bound match the prediction with expert advice setting, where all arms are observed on all rounds.

Another interesting observation is the decrease rate in the number of observations made by the algorithm. For the uniform costs setting the number of observations per round decreases as $\Theta\left(\frac{(N \ln N)^{1/3}}{c^{2/3} t^{1/3}}\right)$. When the cost of observations is relatively small compared to the time horizon, the regime when the algorithm queries more than one observation per round has significant interest.

The last result of Theorem 3 is accompanied by a matching (up to logarithmic terms) lower bound. The proof of Theorem 4 is provided in Section 3.

Theorem 4. *In the N-armed bandit game with uniform cost of observations c*

$$\inf \sup R_T^c \geq \max \left\{ 0.19 \, (cN)^{1/3} T^{2/3}, \; 0.03\sqrt{T} \right\},$$

where the infimum is over all playing strategies and the supremum is over all oblivious adversaries.

We note that there is no contradiction between Theorem 4 and the data-dependent improvement achieved in Eq. (4), since Theorem 4 considers the worst case and the improvement is achieved under benign conditions.

3. Proofs

The analysis of both our algorithms is based on the following lemma, which represents an intermediate step in the analysis of EXP3 by Bubeck (2010).

Lemma 5. *For any N sequences of random variables $L_1^h, L_2^h, \ldots$ indexed by $h \in \{1, \ldots, N\}$, such that $L_t^h \geq 0$, and any non-increasing positive sequence $\eta_1, \eta_2, \ldots$, for $q_t(h) = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} L_s^h\right)}{\sum_{h'} \exp\left(-\eta_t \sum_{s=1}^{t-1} L_s^{h'}\right)}$ (assuming for $t = 1$ the sum in the exponent is zero) we have:*

$$\sum_{t=1}^{T} \sum_{h} q_t(h) L_t^h - \min_h \left(\sum_{t=1}^{T} L_t^h \right) \leq \sum_{t=1}^{T} \frac{\eta_t}{2} \sum_{h} q_t(h) \left(L_t^h\right)^2 + \frac{\ln N}{\eta_T}. \tag{6}$$

More precisely, we are using the following corollary, which follows by taking expectations of the two sides of (6) and using the fact that $\mathbb{E}[\min[\cdot]] \leq \min[\mathbb{E}[\cdot]]$. We decompose expectations of incremental sums into sums of conditional expectations and use $\mathbb{E}_t[\cdot]$ to denote expectations conditioned on observations up to round t.

Corollary 6. *Under the definitions of Lemma 5:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[\sum_{h} q_t(h) L_t^h\right]\right] - \min_h \left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[L_t^h\right]\right]\right) \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[\frac{\eta_t}{2} \sum_{h} q_t(h) \left(L_t^h\right)^2\right]\right] + \frac{\ln N}{\eta_T}. \tag{7}$$

We also use the following two technical lemmas. The proofs of the lemmas are provided in the appendix.

Lemma 7. *For any probability distribution q on $\{1, \ldots, N\}$ and any $m \in [1, N]$:*

$$\sum_{h=1}^{N} \frac{q(h)(N-1)}{q(h)(N-m) + m - 1} \leq \frac{N}{m}. \tag{8}$$

Lemma 8. *For any sequence of non-negative numbers $a_1, a_2, \ldots$, such that $a_1 > 0$, and any power $\gamma \in (0, 1)$ we have:*

$$\sum_{t=1}^{T} \frac{a_t}{\left(\sum_{s=1}^{t} a_s\right)^{\gamma}} \leq \frac{1}{1-\gamma} \left(\sum_{t=1}^{T} a_t \right)^{1-\gamma}.$$

Lemma 8 is a generalization of Auer et al. (2002b, Lemma 3.5) from $\gamma = 1/2$ to arbitrary γ.

Now we are ready to present the proofs of the theorems.

3.1. Proof of Theorem 1

Proof. We study $\mathbb{E}_t\left[L_t^h\right]$, $\mathbb{E}_t\left[\sum_h q_t(h)L_t^h\right]$, and $\mathbb{E}_t\left[\sum_h q_t(h)\left(L_t^h\right)^2\right]$ for the case of our algorithm. We have:

$$\mathbb{E}_t\left[L_t^h\right] = \ell_t^h. \tag{9}$$

And we have:

$$\mathbb{E}_t\left[\sum_h q_t(h)L_t^h\right] = \mathbb{E}_t\left[\sum_h \ell_t^h \mathbb{1}_t^h\right] = \mathbb{E}_t\left[L_t\right]. \tag{10}$$

We also have:

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_h q_t(h)\left(L_t^h\right)^2\right] \\
&\quad= \mathbb{E}_t\left[\sum_h q_t(h)\left(\frac{\ell_t^h}{q_t(h)+(1-q_t(h))\frac{M_t-1}{N-1}}\mathbb{1}_t^h\right)^2\right] \\
&\quad= \mathbb{E}_t\left[\sum_h q_t(h)\frac{\left(\ell_t^h\right)^2}{\left(q_t(h)+(1-q_t(h))\frac{M_t-1}{N-1}\right)^2}\mathbb{1}_t^h\right] \\
&\quad\le \mathbb{E}_t\left[\sum_h q_t(h)\frac{1}{\left(q_t(h)+(1-q_t(h))\frac{M_t-1}{N-1}\right)^2}\mathbb{1}_t^h\right] \\
&\quad= \sum_h q_t(h)\frac{1}{\left(q_t(h)+(1-q_t(h))\frac{M_t-1}{N-1}\right)^2}\mathbb{E}_t\left[\mathbb{1}_t^h\right] \\
&\quad= \sum_h q_t(h)\frac{1}{q_t(h)+(1-q_t(h))\frac{M_t-1}{N-1}} \\
&\quad= \sum_h \frac{q_t(h)(N-1)}{q_t(h)(N-M_t)+M_t-1} \\
&\quad\le \frac{N}{M_t},
\end{aligned} \tag{11}$$

where the last inequality is by Lemma 7. By substituting (9), (10), and (11) into (7) we obtain for all h:

$$\mathbb{E}\left[\sum_{t=1}^T L_t\right] - \min_h\left(\sum_{t=1}^T \ell_t^h\right) \le \frac{N}{2}\sum_{t=1}^T \frac{\eta_t}{M_t} + \frac{\ln N}{\eta_T}.$$

This proves the first inequality in Theorem 1. The second inequality follows by the choice of η_t and Lemma 8 (for $\gamma = 1/2$) and the last inequality follows from the identity $\sum_{t=1}^T \frac{1}{M_t} = \frac{T}{M}$ when $M_t = M$ is constant. □

3.2. Proof of Theorem 3

Proof. Similar to the previous proof, it is easy to verify that $\mathbb{E}_t\left[L_t^h\right] = \ell_t^h$, that $\mathbb{E}_t\left[\sum_{h=1}^N q_t(h)L_t^h\right] = \mathbb{E}_t\left[L_t\right]$, and that $\mathbb{E}_t\left[\sum_{h\in\mathcal{O}_t} c_t(h)\right] = \sum_{h=1}^N p_t(h)c_t(h)$. Furthermore, we have $\mathbb{E}_t\left[\left(L_t^h\right)^2\right] \le \frac{1}{p_t(h)}$ and thus $\mathbb{E}_t\left[\sum_{h=1}^N q_t(h)\left(L_t^h\right)^2\right] \le \sum_{h=1}^N \frac{q_t(h)}{p_t(h)}$. By substituting this into (7) and adding the cost of observations we obtain:

$$\begin{aligned}
R_T^c &= \mathbb{E}\left[\sum_{t=1}^T L_t\right] + \mathbb{E}\left[\sum_{t=1}^T\sum_{h=1}^N \mathbb{1}_t^h c_t(h)\right] - \min_h\left(\sum_{t=1}^T \ell_t^h\right) \\
&\le \sum_{t=1}^T\sum_{h=1}^N\left(\frac{\eta_t q_t(h)}{2p_t(h)} + c_t(h)p_t(h)\right) + \frac{\ln N}{\eta_T},
\end{aligned} \tag{12}$$

where η_t has to be a non-increasing sequence.

Our first goal is to minimize the instantaneous contributions $\sum_{h=1}^N \frac{\eta_t q_t(h)}{2p_t(h)} + c_t(h)p_t(h)$. This leads to the following optimization problem

$$\begin{aligned}
\min_{p_t}\ &\sum_{h=1}^N \frac{\eta_t q_t(h)}{2p_t(h)} + c_t(h)p_t(h) \\
\text{s.t. }\ &\forall h:\ 0 \le p_t(h) \le 1,
\end{aligned} \tag{13}$$

which is solved by $p_t^*(h) = \min\left\{1, \sqrt{\frac{\eta_t q_t(h)}{2c_t(h)}}\right\}$. We note that if $p_t^*(h) = 1$ it means that $\frac{\eta_t q_t(h)}{2c_t(h)} \ge 1$, which in turn means that $c_t(h) = \sqrt{c_t(h)}\sqrt{c_t(h)} \le \sqrt{\frac{1}{2}\eta_t q_t(h)c_t(h)}$. By substituting p_t^* into the minimization problem (13) we obtain:

$$\begin{aligned}
&\sum_{h=1}^N \frac{\eta_t q_t(h)}{2p_t^*(h)} + c_t(h)p_t(h) \\
&\qquad= \sum_{h=1}^N \mathbb{1}_{\{p_t^*(h)=1\}}\left(\frac{\eta_t q_t(h)}{2} + c_t(h)\right) \\
&\qquad\quad + \left(\mathbb{1}_{\{p_t^*(h)<1\}}\right)\sqrt{2\eta_t q_t(h)c_t(h)} \\
&\qquad\le \frac{\eta_t}{2} + \sqrt{2\eta_t}\sum_{h=1}^N \sqrt{q_t(h)c_t(h)}.
\end{aligned}$$

By substituting this result back into (12) we obtain the first claim of the theorem:

$$R_T^c \le \sum_{t=1}^T\left(\frac{\eta_t}{2} + \sqrt{2\eta_t}\sum_{h=1}^N \sqrt{q_t(h)c_t(h)}\right) + \frac{\ln N}{\eta_T}. \tag{14}$$

Now we have to tune the learning rate η_t. We note that by Jensen's inequality $\sum_{h=1}^N \sqrt{q_t(h)c_t(h)} =$

$\sum_{h=1}^{N} q_t(h)\sqrt{\frac{c_t(h)}{q_t(h)}} \leq \sqrt{\sum_{h=1}^{N} c_t(h)}$. We set:

$$\eta_t = \frac{1}{\left(\frac{\sqrt{\sum_{h=1}^{N} c_t(h)} + \sum_{s=1}^{t-1}\sum_{h=1}^{N}\sqrt{q_s(h)c_s(h)}}{\frac{\sqrt{2}}{3}\ln N}\right)^{2/3} + \sqrt{\frac{t}{\ln N}}},$$

where $\sqrt{\sum_{h=1}^{N} c_t(h)}$ should be seen as an upper bound on $\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)}$. For a moment assume that $\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)} > 0$ for all t. Then we have:

$$\begin{aligned}
&\sum_{t=1}^{T}\sqrt{2\eta_t}\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)} \\
&\leq \sum_{t=1}^{T}\frac{\left(\frac{4}{3}\ln N\right)^{1/3}\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)}}{\left(\sqrt{\sum_{h=1}^{N} c_t(h)} + \sum_{s=1}^{t-1}\sum_{h=1}^{N}\sqrt{q_s(h)c_s(h)}\right)^{1/3}} \\
&\leq \left(\frac{4}{3}\ln N\right)^{1/3}\sum_{t=1}^{T}\frac{\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)}}{\left(\sum_{s=1}^{t}\sum_{h=1}^{N}\sqrt{q_s(h)c_s(h)}\right)^{1/3}} \\
&\leq \left(\frac{9}{2}\ln N\right)^{1/3}\left(\sum_{t=1}^{T}\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)}\right)^{2/3},
\end{aligned}$$

where in the last inequality we used Lemma 8. Note that if for some t we have $\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)} = 0$ the corresponding game round makes no contribution to both sides of the inequality and so the result holds irrespective of the assumption that $\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)} > 0$ for all t. By substituting the selected η_t into (14) we obtain bound (2):

$$\begin{aligned}
&R_T^c \\
&\leq (32\ln N)^{1/3}\left(\sqrt{\sum_{h=1}^{N} c_{_T}(h)} + \sum_{t=1}^{T-1}\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)}\right)^{2/3} \\
&\quad + 2\sqrt{T\ln N}.
\end{aligned}$$

Inequality (3) follows in a similar way after applying $\sum_{h=1}^{N}\sqrt{q_t(h)c_t(h)} \leq \sqrt{\sum_{h=1}^{N} c_t(h)}$ in (14). □

3.3. Proof of Theorem 4

Proof. By Theorem 2 we know that without taking the cost of the queries into account the regret of any algorithm that makes MT observations is lower bounded as:

$$\inf\sup R_T \geq 0.03\sqrt{\frac{N}{M}T}.$$

Adding the cost of observations we have that the regret of any algorithm that makes MT observations is lower bounded by:

$$\begin{aligned}
R_T^c = R_T + cMT &\geq 0.03\sqrt{\frac{N}{M}T} + cMT \\
&\geq \max\left\{0.19c^{1/3}N^{1/3}T^{2/3}, 0.03\sqrt{T}\right\},
\end{aligned}$$

where the last inequality follows by the fact that the expression is minimized by $M = 0.03^{2/3}c^{-2/3}N^{1/3}T^{-1/3}$ and that M is upper bounded by N. □

4. Discussion

We defined the games of prediction with limited advice and multiarmed bandits with paid observations and provided algorithms and matching (up to logarithmic factors) upper and lower bounds for the two games. Our algorithm for multiarmed bandits with paid observations treats arm- and time-dependent observation costs and reduces the regret below the worst-case lower bound under benign conditions.

Our work opens multiple directions for future research. The multiarmed bandits with paid observations game can serve as a basic model for learning under restricted information access in more general reinforcement learning problems. At the same time, prediction with limited advice game poses interesting questions about learning under constraints on the resources, for example, whether it is possible to achieve sub-polynomial dependence on the number of experts with a sub-polynomial amount of advice, under some assumptions on the loss function and/or the experts class. (Due to the lower bounds we know that without additional assumptions this is impossible.) It would also be interesting to extend both games to continuous domains and to stochastic environments.

Acknowledgments

We would like to thank Claudio Gentile for pointing the reference to Auer et al. (2002b, Lemma 3.5) and Wouter Koolen for helpful discussions and comments on a preliminary draft. This research was supported by an Australian Research Council Australian Laureate Fellowship (FL110100281). We gratefully acknowledge the support of the NSF through grant CCF-1115788. This research was funded in part by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI).

References

Alon, Noga, Cesa-Bianchi, Nicolò, Gentile, Claudio, and Mansour, Yishay. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Audibert, Jean-Yves and Bubeck, Sébastien. Regret bounds and

minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11, 2010.

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1), 2002a.

Auer, Peter, Cesa-Bianchi, Nicolò, and Gentile, Claudio. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64, 2002b.

Avner, Orly, Mannor, Shie, and Shamir, Ohad. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Bartók, Gábor, Pál, Dávid, and Szepesvári, Csaba. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2011.

Bubeck, Sébastien. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille, 2010.

Bubeck, Sébastien and Cesa-Bianchi, Nicolò. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5, 2012.

Cesa-Bianchi, Nicolò and Lugosi, Gábor. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Cesa-Bianchi, Nicolò, Lugosi, Gábor, and Stoltz, Gilles. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51, 2005.

Foster, Dean P. and Rakhlin, Alexander. No internal regret via neighborhood watch. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Mannor, Shie and Shamir, Ohad. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Ottucsák, György and György, András. The combination of the label efficient and the multi-armed bandit problem in adversarial setting. Technical report, http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.126.1228, 2006.

Zolghadr, Navid, Bartók, Gábor, Greiner, Rissell, György, András, and Szepesvári, Csaba. Online learning with costly features and labels. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.