JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–6

Open Problem: Adversarial Multiarmed Bandits with Limited Advice

Yevgeny Seldin YEVGENY.SELDIN@GMAIL.COM
Mathematical Sciences School, Queensland University of Technology, Brisbane, QLD, Australia
Department of Electical Engineering and Computer Sciences, UC Berkeley, CA, USA
Department of Computer Science, University College London, UK

Koby Crammer KOBY@EE.TECHNION.AC.IL
Department of Electrical Engineering, The Technion, Haifa, Israel

Peter Bartlett BARTLETT@EECS.BERKELEY.EDU
Dept. of Electical Engineering and Computer Sciences, Dept. of Statistics, UC Berkeley, CA, USA
Mathematical Sciences School, Queensland University of Technology, Brisbane, QLD, Australia

Abstract

Adversarial multiarmed bandits with expert advice is one of the fundamental problems in studying the exploration-exploitation trade-off. It is known that if we observe the advice of all experts on every round we can achieve $O\left(\sqrt{KT\ln N}\right)$ regret, where K is the number of arms, T is the number of game rounds, and N is the number of experts. It is also known that if we observe the advice of just one expert on every round, we can achieve regret of order $O\left(\sqrt{NT}\right)$. Our open problem is what can be achieved by asking M experts on every round, where $1 < M < N$.

Keywords: Adversarial Multiarmed Bandits with Expert Advice, EXP4

1. Introduction

Adversarial multiarmed bandits with expert advice is one of the fundamental problems in studying the exploration-exploitation trade-off (Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006; Bubeck and Cesa-Bianchi, 2012). The main use of this model is in problems, where we do not make statistic assumptions on the data generating process. The model was applied to real-world problems, such as online advertizing and news article recommendation (Beygelzimer et al., 2011; Li et al., 2010).

The adversarial multiarmed bandits with expert advice problem can be described as a game with T rounds. On each round t of the game there are K options (arms) indexed by $a \in \{1, \ldots, K\}$. Arm a on round t yields reward $r_t(a)$. It is assumed that the sequences of rewards are written down before the game starts, but not revealed to the player. On each round of the game the player observes advice of N experts in a form of a distribution ξ_t^h on $\{1, \ldots, K\}$, where $h \in \{1, \ldots, N\}$ indexes the experts. The player makes a choice of an arm A_t and observes and accumulates reward $r_t(A_t)$. The rewards of other arms are not observed. The reward of expert h on round t is defined as $r_t(h) \equiv \sum_a \xi_t^h(a) r_t(a)$. The goal of the player is to minimize the regret defined as $\max_h \left(\sum_{t=1}^T r_t(h)\right) - \sum_{t=1}^T r_t(A_t)$.

2. Open Problem Formulation and Motivation

In certain situations it may be overly expensive to query advice of all experts on all rounds of the game. For example, if experts are doctors giving advice on patient treatment options, it may be too expensive to ask for advice of all available doctors for each patient. The restriction does not have to be monetary, in some situations it may be computational or other constraints.

If we observe advice of all experts we can run the EXP4 algorithm of Auer et al. (2002) and achieve $O\left(\sqrt{KT\ln N}\right)$ regret. (See also Beygelzimer et al. (2011) for a high-probability version and Bubeck and Cesa-Bianchi (2012) for a simplified derivation.) If we observe the advice of just one expert of our choice, we can achieve $O\left(\sqrt{NT\ln N}\right)$ regret by running the EXP3 algorithm of Auer et al. (2002) for adversarial multiarmed bandits, where we consider each expert as an arm. (Audibert and Bubeck (2010) improve the result by $\sqrt{\ln N}$ factor.) Our question is what happens in between. Specifically, if we ask for advice of $1 < M < N$ experts on every round, what order of regret can be achieved? We call this setting multiarmed bandits with limited advice. Based on the result for the full information setting described below we conjecture that it should be possible to achieve $O\left(\sqrt{\frac{N}{M}KT\ln N}\right)$ regret. Both upper and lower bounds would be of interest.

3. Known Related Results

Seldin et al. (2013) derived an algorithm for prediction with limited advice (the full information counterpart of the open problem) with regret guarantee of $O\left(\sqrt{\frac{N}{M}T\ln N}\right)$. This result nicely interpolates between $O\left(\sqrt{T\ln N}\right)$ regret bound when observing advice of all experts and $O\left(\sqrt{NT}\right)$ regret bound when observing advice of just one expert (in the full information setting). On each round of the game the algorithm of Seldin et al. queries the advice of one expert according the distribution corresponding to the weights of exponentially weighted forecaster. The algorithm follows the advice of the sampled expert and then queries the advice of $M-1$ additional experts sampled uniformly at random. At the end of the round the algorithm updates the rewards of all experts using importance-weighting.

Seldin et al. also derived a matching (up to logarithmic factors) lower bound $\Omega\left(\sqrt{\frac{N}{M}T}\right)$ for prediction with limited advice. Obviously, the lower bound also holds for the harder multiarmed bandit with limited advice setting. (The interesting question in the bandit case is to introduce $\sqrt{K}$ into the lower bound.)

We note that the algorithm of Seldin et al. cannot be extended to the bandit case, since when we follow the advice of one expert the importance-weighted estimates of rewards of other experts have very large variance, even if we slightly smooth the advice of the sampled expert. (This is because ξ_t^h may be very different from $\xi_t^{h'}$ for $h \neq h'$.) In the appendix we describe another attempt to derive an algorithm for the open problem, where the playing strategy (distribution over arms played) in each round is based on the advice of all experts sampled on that round. However, this approach can give, in the best case,

$O\left(\sqrt{(N-M+1)KT\ln N}\right)$ regret bound, which has a bit disappointing dependence on M (even though it would provide a continuous interpolation between asking one expert and asking all experts).

The problem of prediction with limited advice is related to label-efficient prediction (Cesa-Bianchi and Lugosi, 2006; Audibert and Bubeck, 2010). In label-efficient prediction all experts are queried on a subset of game rounds and in prediction with limited advice a subset of experts is queried on all game rounds. We note that the formulation of prediction with limited advice is substantially different from learning with partially observed attributes (Cesa-Bianchi et al., 2011), but possibly some tools could be transferred between the settings.

Acknowledgements

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.

References

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

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

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

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

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

Nicolò Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Efficient learning with partially observed attributes. *Journal of Machine Learning Research*, 2011.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2010.

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. Technical report, http://arxiv.org/abs/1304.3708, 2013.

$\forall h$: $\hat{L}_0(h) = 0$.
for $i = 1, 2, \ldots$ **do**

Let

$$q_t(h) = \frac{e^{-\eta_t \hat{L}_{t-1}(h)}}{\sum_h e^{-\eta_t \hat{L}_{t-1}(h)}}.$$

Sample M experts without replacement, such that the probability of sampling expert h is $\tilde{q}_t(h)$. ($\tilde{q}_i(h)$ is specified in the analysis of the algorithm.) Let $\mathbb{1}_t^h = 1$ if expert h was sampled and $\mathbb{1}_t^h = 0$ otherwise.

Get advice vectors ξ_t^h for the experts sampled.

Let

$$p_t(a) = \frac{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)} \xi_t^h(a) \mathbb{1}_t^h}{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)} \mathbb{1}_t^h}.$$

Draw action A_t according to p_t and receive reward $R_t \in [0, 1]$.

$$\forall a: \; L_t^a = \frac{1 - R_t}{p_t(a)} \mathbb{1}_{\{A_t = a\}}.$$

$$\forall h: Y_t^h = \xi_t^h(A_t) L_t^{A_t} \frac{1}{\tilde{q}_t(h)} \mathbb{1}_t^h.$$

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

end

Algorithm 1: A general algorithm for multiarmed bandits with limited advice.

Appendix A. An Attempt to Solve the Problem

In this appendix we analyze a general algorithm that uses the advice of all M experts sampled to generate a prediction strategy for round t. We show that it is not trivial to control bias and variance simultaneously. See Algorithm 1 box for the description of the algorithm.

Analysis

The analysis is based on the following lemma, which follows from the analysis of EXP3 by Bubeck and Cesa-Bianchi (2012).

Lemma 1 *For any N sequences of random variables $Y_1^h, Y_2^h, \ldots$ indexed by $h \in \{1, \ldots, N\}$, such that $Y_t^h \geq 0$, and any non-increasing sequence $\eta_1, \eta_2, \ldots$, such that $\eta_t \geq 0$, for $q_t(h) = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} Y_s^h\right)}{\sum_{h'} \exp\left(-\eta_t \sum_{s=1}^{t-1} Y_s^{h'}\right)}$ (assuming for $t = 1$ the sum in the exponent is zero), for any $h^\star$ we*

have:

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

We study $\sum_h q_t(h)Y_t^h$ and $\sum_h q_t(h)\left(Y_t^h\right)^2$ in the case of of Algorithm 1.

Study of $\sum_h q_t(h)Y_t^h$:

$$\sum_{h} q_t(h)Y_t^h = \sum_{h} q_t(h)\xi_t^h(A_t)L_t^{A_t}\frac{\mathbb{1}_t^h}{\tilde{q}_t(h)} = L_t^{A_t}\sum_{h} q_t(h)\xi_t^h(A_t)\frac{\mathbb{1}_t^h}{\tilde{q}_t(h)} = (1-R_t)\sum_{h}\frac{q_t(h)}{\tilde{q}_t(h)}\mathbb{1}_t^h. \tag{1}$$

Avoiding the Bias:

We remind that in the analysis of EXP4 $\sum_h q_t(h)Y_t^h = 1 - R_t$. One way to ensure that there is no bias in the estimation in (1) is to make sure that $\sum_h \frac{q_t(h)}{\tilde{q}_t(h)}\mathbb{1}_t^h = 1$ for any random draw of the hypotheses subset. This is achieved, for example, if $\tilde{q}_t(h) = 1$ for all h. Or, if $\tilde{q}_t(h) = q_t(h)$ for all h and we draw exactly one hypothesis. (The first choice corresponds to the EXP4 algorithm and the second choice corresponds to the case, where we sample just one expert on each round and play the EXP3 algorithm on the experts.) A more general way to eliminate the bias that combines the two approaches for a general M is described in Algorithm 2 box.

Bounding $\sum_h q_t(h)\left(Y_t^h\right)^2$ for sampling from $\tilde{q}_t(h)$:

$$\begin{aligned}
\sum_{h} q_t(h)\left(Y_t^h\right)^2 &= \sum_{h} q_t(h)\left(\xi_t^h(A_t)L_t^{A_t}\frac{\mathbb{1}_t^h}{\tilde{q}_t(h)}\right)^2 = \left(L_t^{A_t}\right)^2\sum_{h} q_t(h)\left(\xi_t^h(A_t)\frac{\mathbb{1}_t^h}{\tilde{q}_t(h)}\right)^2 \\
&\leq \frac{1}{\left(p_t(A_t)\right)^2}\sum_{h}\frac{q_t(h)}{\tilde{q}_t(h)^2}\xi_t^h(A_t)\mathbb{1}_t^h = \frac{1}{p_t(A_t)}\frac{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)^2}\xi_t^h(A_t)\mathbb{1}_t^h}{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)}\xi_t^h(A_t)\mathbb{1}_t^h}\sum_{h}\frac{q_t(h)}{\tilde{q}_t(h)}\mathbb{1}_t^h \\
&= \frac{1}{p_t(A_t)}\frac{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)^2}\xi_t^h(A_t)\mathbb{1}_t^h}{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)}\xi_t^h(A_t)\mathbb{1}_t^h} \leq \frac{1}{p_t(A_t)}\sum_{h}\frac{1}{\tilde{q}_t(h)}\mathbb{1}_t^h,
\end{aligned}$$

where in the last line we used the fact that $\sum_h \frac{q_t(h)}{\tilde{q}_t(h)}\mathbb{1}_t^h = 1$ and then lower bounded the sum in the denominator by its individual element.

We remind that the quantity of interest in the analysis of EXP4 was $\mathbb{E}_t\left[\sum_h q_t(h)\left(Y_t^h\right)^2\right]$, where $\mathbb{E}_t[\cdot]$ denotes expectation conditioned on realizations of random variables up to round t. Taking this expectation in our case yields

$$\mathbb{E}_t\left[\frac{1}{p_t(A_t)}\sum_{h}\frac{1}{\tilde{q}_t(h)}\mathbb{1}_t^h\right] = \mathbb{E}_t\left[\sum_{h}\frac{1}{\tilde{q}_t(h)}\mathbb{1}_t^h\mathbb{E}\left[\frac{1}{p_t(A_t)}\middle|\{\mathbb{1}_t^h\}_h\right]\right] = KN.$$

Without loss of generality, assume that $q_t(h)$ are ordered in decreasing order $q_t(1) \geq q_t(2) \geq \cdots \geq q_t(N)$. Let $h_t^\dagger = \max\{h : Mq_t(h) \geq 1\}$.

For $h \leq h_t^\dagger$ define $\tilde{q}_t(h) = 1$. *% For $h \leq h_t^\dagger$ experts h are sampled w.p. 1*

Let $w_t^+ = \sum_{h=1}^{h_t^\dagger} q_t(h)$. *% w_t^+ is the total q_t-mass of the experts that are sampled w.p. 1*

$h = h_t^\dagger + 1$. *%The loop below adds more experts that after scaling q_t are sampled w.p. 1*

while $h \leq N$ *AND* $\frac{\left(M - h_t^\dagger\right) q_t(h)}{1 - w_t^+} \geq 1$ **do**

  $\tilde{q}_t(h) = 1$.

  $h_t^\dagger = h_t^\dagger + 1$.

  $w_t^+ = w_t^+ + q_t(h)$.

  $h = h + 1$.

end

For $h > h_t^\dagger$ define $\tilde{q}_t(h) = \frac{\left(M - h_t^\dagger\right) q_t(h)}{1 - w_t^+}$.

Algorithm 2: Algorithm for defining $\tilde{q}_t(h)$.

There is a bit tighter way to bound $\frac{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)^2} \xi_t^h(A_t) \mathbb{1}_t^h}{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)} \xi_t^h(A_t) \mathbb{1}_t^h}$. Assume without loss of generality that $q_t(1) \geq \cdots \geq q_t(N)$. Then for any $h_t^\ddagger \in \{1, \ldots, N\}$:

$$\sum_h q_t(h) \left(Y_t^h\right)^2 \leq \frac{1}{p_t(A_t)} \frac{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)^2} \xi_t^h(A_t) \mathbb{1}_t^h}{\sum_h \frac{q_t(h)}{\tilde{q}_t(h)} \xi_t^h(A_t) \mathbb{1}_t^h} \leq \frac{1}{p_t(A_t)} \left(\frac{1}{q_t(h_t^\ddagger)} + \sum_{h=h_t^\ddagger+1}^{N} \frac{1}{\tilde{q}_t(h)} \mathbb{1}_t^h \right).$$

With the above bound we obtain $\mathbb{E}_t \left[\sum_h q_t(h) \left(Y_t^h\right)^2 \right] \leq K \left(\frac{1}{\tilde{q}_t(h_t^\ddagger)} + N - h_t^\ddagger \right)$. Using this approach it seems possible to get a bound on $\mathbb{E}_t \left[\sum_h q_t(h) \left(Y_t^h\right)^2 \right]$ of order $K(N - M + 1)$ (we did not prove such bound, but we also could not find a counter example). This shows at least some minimal advantage of sampling more than one expert. Unfortunately, it does not seem possible to achieve higher benefit from M experts using this approach. Examples of "hard" distributions include $q_t(h) = 2^{-h}$ (where $h \in \{1, \ldots, N\}$) and $q_t(h) = (1-\varepsilon)/(M-1)$ for $h \in \{1, \ldots, M-1\}$ and $q_t(h) = \varepsilon/(N - M + 1)$ for $h \in \{M, \ldots, N\}$ and a small ε.

We note that it does not seem possible to apply smoothing to achieve sufficient reduction in the variance while keeping the bias under control.