A Lower Bound for Multi-Armed Bandits with Expert Advice

Yevgeny Seldin SELDIN@DI.KU.DK
Department of Computer Science
University of Copenhagen

Gábor Lugosi GABOR.LUGOSI@GMAIL.COM
ICREA and Department of Economics
Pompeu Fabra University

Abstract

In this note we derive a lower bound for the regret in the adversarial multi-armed bandit problem with expert advice, showing that the regret is $\Omega\left(\sqrt{KT\frac{\ln N}{\ln K}}\right)$, where K is the number of arms, T is the number of game rounds, and N is the number of experts. The $\sqrt{\ln N}$ factor in the lower bound solves a problem that was open since the work of Auer et al. (2002) and posed explicitly in Stoltz (2005).

1. Introduction

The adversarial multi-armed bandit problem with expert advice, introduced in Auer et al. (2002), is one of the fundamental models in online learning. It combines prediction with expert advice (Vovk, 1990; Littlestone and Warmuth, 1994) with the limited feedback framework of multi-armed bandits (Auer et al., 2002). In this model, a repeated game is played between a decision maker and an adversary. At every round of the game, the decision maker (or forecaster) has a selection of K possible actions (the so-called "arms") and access to the "advice" of N "experts", where the "experts" can be seen as (potentially randomized) algorithms for action selection. The decision maker selects one action and, simultaneously, the adversary assigns a (bounded) loss to each arm. The forecaster suffers the loss assigned to the selected arm and each "expert" suffers the loss of the arm of its choice. The loss of the arm selected by the forecaster is revealed to the forecaster, but the losses of other arms remain unobserved. (Note that this implies that the forecaster does not observe the losses of experts which did not select the same arm as the forecaster.) The performance of the forecaster is measured by the *regret*, which is defined as the difference between the cumulative loss of the forecaster and that of the best expert in hindsight.

Auer et al. (2002) proposed a randomized algorithm (EXP4) for the adversarial multi-armed bandit problem with expert advice and showed that in expectation it achieves a regret of order $O\left(\sqrt{KT\ln N}\right)$ after T rounds. It was also shown that with a slight modification of the algorithm the same order of regret can be guaranteed with high probability (Beygelzimer et al., 2011). The algorithm and its optimized modifications have been extensively applied to advertizing and personalized recommendations (Beygelzimer et al., 2011; Agarwal et al., 2014). While several improvements of the algorithm and the regret bound have been proposed for special cases (McMahan and Streeter, 2009; Seldin et al., 2011), in the general (worst case) scenario the regret guarantee of EXP4 remains the best known to date.

On the side of lower bounds, a well-known lower bound for the problem of prediction with expert advice under full information is $\Omega\left(\sqrt{T \ln N}\right)$ (Cesa-Bianchi et al., 1997) and a standard lower bound for the expected regret in adversarial multi-armed bandit problems (without expert advice) is $\Omega\left(\sqrt{KT}\right)$ (Auer et al., 2002). Both lower bounds are matched up to constants by upper bounds (Vovk, 1990; Littlestone and Warmuth, 1994; Audibert and Bubeck, 2010; Bubeck and Cesa-Bianchi, 2012). Since multi-armed bandits with expert advice generalize both problems, the lower bounds for the two problems apply to them and, therefore, the expected regret is at least $\Omega\left(\max\left\{\sqrt{T \ln N}, \sqrt{KT}\right\}\right)$. However, the necessity of the product $\sqrt{K \ln N}$ in the lower bound (that would match the upper bound) has remained an open question since the model was introduced by Auer et al..

We answer this question affirmatively, showing that the regret is at least $\Omega\left(\sqrt{KT\frac{\ln N}{\ln K}}\right)$. The lower bound is proved by constructing several independent instances of the multi-armed bandit game and linking them through expert advice. Note that for $N = K$ multi-armed bandit with expert advice can be reduced to a multi-armed bandit (by considering experts as arms) and that in this case the new regret lower bound matches the lower and upper bounds for multi-armed bandits. Therefore, we conjecture that the lower bound has the right asymptotic regret rate for the problem and that the $\sqrt{\frac{1}{\ln K}}$ factor in the lower bound cannot be improved.

2. Problem Setting

For our derivation of the lower bound it is sufficient to consider deterministic experts and in order to keep things simple we consider a problem setting with deterministic experts. For discussion of alternative versions of the model we refer the reader to Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012). Note that since deterministic experts are a special case of randomized experts the lower bound also holds for the more general setting.

Let $K, N, T \geq 2$ be positive integers denoting the number of arms, number of experts, and the length of the game, respectively. At each round $t = 1, \ldots, T$, each expert $h \in \{1, \ldots, N\}$ selects an arm $e_{h,t} \in \{1, \ldots, K\}$. Upon observing the advice vector $(e_{1,t}, \ldots, e_{N,t})$, the forecaster selects an arm A_t in a possibly randomized manner. We consider the so-called *oblivious adversary* model in which the adversary cannot react to the choices of the forecaster. Formally, the losses x_t^a for $t \in \{1, \ldots, T\}$ and $a \in \{1, \ldots, K\}$ are arbitrary and fixed before the game starts. For the derivation of a lower bound we assume without loss of generality that the losses are binary, $x_t^a \in \{0, 1\}$.

The cumulative loss of the forecaster after T rounds is $\sum_{t=1}^T x_t^{A_t}$ and the regret of the forecaster after T rounds is the difference between its cumulative loss and the cumulative loss of the best expert in hindsight, that is, $\sum_{t=1}^T x_t^{A_t} - \min_{h \in \{1,\ldots,N\}} \sum_{t=1}^T x_t^{e_{h,t}}$. Note that the regret is a random variable due to the randomization of the forecaster. We study the expected regret $\mathbb{E}\left[\sum_{t=1}^T x_t^{A_t}\right] - \min_{h\in\{1,\dots,N\}} \sum_{t=1}^T x_t^{e_{h,t}}$. The *minimax regret* is

$$\inf_A \sup_{x,e} \left(\mathbb{E}\left[\sum_{t=1}^T x_t^{A_t}\right] - \min_{h\in\{1,\dots,N\}} \sum_{t=1}^T x_t^{e_{h,t}} \right) ,$$

where the supremum is taken over all possible loss assignments $x = \{x_t^a\}_{t\in\{1,\dots,T\},a\in\{1,\dots,K\}} \in \{0,1\}^{KT}$ and expert advice $e = \{e_{h,t}\}_{h\in\{1,\dots,N\},t\in\{1,\dots,T\}} \in \{1,\dots,K\}^{NT}$ and the infimum is over all randomized forecasters.

As it is usual in proving lower bounds for the minimax regret, the first steps are

$$\begin{aligned}
&\inf_A \sup_{x,e} \left(\mathbb{E}\left[\sum_{t=1}^T x_t^{A_t}\right] - \min_{h\in\{1,\dots,N\}} \sum_{t=1}^T x_t^{e_{h,t}} \right) \\
&\quad \geq \inf_A \sup_{e} \mathbb{E}\left[\sum_{t=1}^T X_t^{A_t} - \min_{h\in\{1,\dots,N\}} \sum_{t=1}^T X_t^{e_{h,t}} \right] \\
&\quad\quad \text{(where } X_t^a\text{-s are random loss assignments and the expectation is w.r.t } X_t^a\text{-s and } A_t\text{-s)} \\
&\quad \geq \inf_A \sup_{e} \left(\mathbb{E}\left[\sum_{t=1}^T X_t^{A_t}\right] - \min_{h\in\{1,\dots,N\}} \mathbb{E}\left[\sum_{t=1}^T X_t^{e_{h,t}}\right] \right) .
\end{aligned}$$

The right-hand side is sometimes called the *pseudo regret*.

3. Main Result

The main result of the paper is the following theorem.

Theorem 1 *Assume that $N = K^M$ for an integer M and that T is a multiple of M. Assume that $T/M > K/(4\ln(4/3))$. Then there exists a distribution for loss assignments, such that the pseudo regret satisfies*

$$\inf_A \sup_{e} \left(\mathbb{E}\left[\sum_{t=1}^T X_t^{A_t}\right] - \min_{h\in\{1,\dots,N\}} \mathbb{E}\left[\sum_{t=1}^T X_t^{e_{h,t}}\right] \right) \geq c\sqrt{KT\frac{\ln N}{\ln K}} ,$$

where inf *is an infimum over all possible forecasting strategies,* sup *is a supremum over all possible expert advice sequences, the expectations are with respect to both the random generation of losses and the internal randomization of the forecaster, and $c = \frac{\sqrt{2}-1}{\sqrt{32\ln(4/3)}} >$ 0.13.*

4. Proof of the Main Result

The proof is based on the following lower bound for multi-armed bandits due to Auer et al. (2002, Theorem 5.1). The formulation of the theorem is slightly modified based on Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012).

Theorem 2 (Auer et al. 2002) *Assume that $T > K/(4\ln(4/3))$. Then there exists a distribution for loss assignments*[1]*, such that*

$$\inf_A \left(\mathbb{E}\left[\sum_{t=1}^T X_t^{A_t} \right] - \min_{a \in \{1,\dots,K\}} \mathbb{E}\left[\sum_{t=1}^T X_t^a \right] \right) \geq c\sqrt{KT} \ ,$$

where inf *is an infimum over all forecasting strategies, the expectations are with respect to both the random generation of losses and the internal randomization of the forecaster, and* c *is defined in Theorem 1.*

Proof of Theorem 1. Split the time interval $1,\dots,T$ into M non-overlapping subintervals of length T/M. Let $m \in \{1,\dots,M\}$ index the subintervals. For each subinterval design a multi-armed bandit game according to the construction in the proof of Theorem 2, independently of other intervals. Design K^M sequences of expert advice, such that for every possible sequence of arms $a_1,\dots,a_M \in \{1,\dots,K\}^M$ there is an expert h that recommends the arms from the sequence throughout the corresponding subintervals. Due to independence of subintervals, for each subinterval by Theorem 2 we have

$$\inf_A \left(\mathbb{E}\left[\sum_{s=1}^{T/M} X_s^{A_s} \right] - \min_{a \in \{1,\dots,K\}} \mathbb{E}\left[\sum_{s=1}^{T/M} X_s^a \right] \right) \geq c\sqrt{KT/M} \ ,$$

where s indexes the game rounds within a subinterval. Since by our construction there is an expert that picks the best action within each subinterval, we obtain:

$$\begin{aligned}
&\inf_A \sup_e \left(\mathbb{E}\left[\sum_{t=1}^T X_t^{A_t} \right] - \min_{h \in \{1,\dots,N\}} \mathbb{E}\left[\sum_{t=1}^T X_t^{e_{h,t}} \right] \right) \\
&\qquad \geq \sum_{m=1}^M \inf_A \left(\mathbb{E}\left[\sum_{t=(m-1)(T/M)+1}^{m(T/M)} X_t^{A_t} \right] - \min_{a \in \{1,\dots,K\}} \mathbb{E}\left[\sum_{t=(m-1)(T/M)+1}^{m(T/M)} X_t^a \right] \right) \\
&\qquad \geq cM\sqrt{KT/M} \\
&\qquad = c\sqrt{KTM} \ .
\end{aligned}$$

By our construction, we have $N = K^M$ or, equivalently, $M = \frac{\ln N}{\ln K}$. ■

5. Discussion

We presented an $\Omega\left(\sqrt{KT\frac{\ln N}{\ln K}}\right)$ lower bound for the adversarial multi-armed bandit problem with expert advice. Our result leads to a new open question of whether the multiplicative gap of $\sqrt{\frac{1}{\ln K}}$ between the new lower bound and the upper bound can be closed. We

[1] See Auer et al. (2002); Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012) for an explicit construction.

conjecture that the presented lower bound has the right asymptotic rate and that closing the gap will require more advanced tools for the upper bound.

Acknowledgments

Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER

References

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

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, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3), 1997.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108, 1994.

H. Brendan McMahan and Matthew Streeter. Tighter bounds for multi-armed bandits with expert advice. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2009.

Yevgeny Seldin, Peter Auer, François Laviolette, John Shawe-Taylor, and Ronald Ortner. PAC-Bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Gilles Stoltz. *Incomplete Information and Internal Regret in Prediction of Individual Sequences.* PhD thesis, Université Paris-Sud, 2005.

Vladimir Vovk. Aggregating strategies. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1990.