Minimax Games with Bandits

Jacob Abernethy[*]
Division of Computer Science
UC Berkeley
jake@cs.berkeley.edu

Manfred K. Warmuth[†]
Department of Computer Science
UC Santa Cruz
manfred@cse.ucsc.edu

One of the earliest online learning games, now commonly known as the *hedge setting* [Freund and Schapire, 1997], goes as follows. On round t, a Learner chooses a distribution $\mathbf{w}_t$ over a set of n actions, an Adversary reveals $\boldsymbol{\ell}_t \in [0,1]^n$, a vector of losses for each action, and the Learner suffers $\mathbf{w}_t \cdot \boldsymbol{\ell}_t = \sum_{i=1}^n w_{t,i}\ell_{t,i}$. Freund and Schapire [1997] showed that a very simple strategy of *exponentially weighting* the actions according to their cumulative losses provides a near-optimal guarantee. That is, by setting

$$w_{t,i} \propto \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{s,i}\right) \tag{1}$$

and using an analysis of the related *Weighted Majority* algorithm [Littlestone and Warmuth, 1994]), it can be shown that with an appropriately chosen "learning rate" η, the loss of the Learner is at most

$$\text{Loss}_{\text{Alg}} = \sum_{t=1}^{T} \mathbf{w}_t \cdot \boldsymbol{\ell}_t \leq k + \sqrt{2k \ln n} + \ln n. \tag{2}$$

Here k is a bound on the loss of the best expert, which is assumed to be known in advance to the Learner for tuning η. In other words, the performance of the Learner will be not much worse than the performance of the best action.

More recently, the same repeated game was analyzed from a *minimax* perspective by [Abernethy et al., 2008b] with the slight restriction that the losses must be binary. Somewhat surprisingly, the optimal strategy for the Learner as well as its worst cast loss both have a natural and elegant interpretation and can be efficiently estimated by a randomized algorithm. Working with the same repeated game, define the *state* to be the vector $\mathbf{s} \in \mathbb{N}^n$, where s_i is the number of losses of expert i. Given the assumption that the best expert suffers at most k losses, we simply define any state $\mathbf{s}$ with $s_i > k$ for all i to be a "terminal node" of the game.

We can describe the optimal behavior of this game as follows. Imagine a randomized adversary that on each round simply assigns a loss of 1 to a single expert, chosen uniformly at random, and let this random sequence of experts be $i_1, i_2, i_3, \ldots \in [n]$. If our current state were $\mathbf{s}$, this random process would take us along a random sequence of states, $S_0 = \mathbf{s}, S_1 = S_0 + \mathbf{e}_{i_1}, S_2 = S_1 + \mathbf{e}_{i_2}$, etc. Somewhat surprisingly, this seemingly arbitrary adversarial strategy can be used to describe the optimal learner's strategy. Indeed, the minimax-optimal weight the player assigns to expert j at state $\mathbf{s}$ is exactly the probability that j is the last "surviving expert" when losses are assigned according to the random process described above. That is, the optimal weight for expert j is

$$w_j^*(\mathbf{s}) = Pr(\exists\, t \text{ s.t. } S_{t,i} > k \;\forall\, i \neq j \text{ and } S_{t,j} = k).$$

In addition, we can express the worst case loss $L(\mathbf{s})$ that the optimal Learner will suffer starting from $\mathbf{s}$ as $\frac{1}{n}$ times the expected number of rounds required to have all experts cross the k threshold. That is,

$$L(\mathbf{s}) = \frac{1}{n}\,\mathbb{E}(\min\{t \;:\; S_{t,i} > k \;\forall\, i\}).$$

Note that since the bound (2) is close to optimal [Freund and Schapire, 1997, Vovk, 98], the exponential weights (1) (when η is tuned based on k) must somehow approximate the above optimal weights.

This minimax characterization of the prediction game is useful, but unfortunately only applies in the *full-information* setting discussed so far. There is a second core problem, the so-called *bandit setting*, that is the focus of our proposed open problem. In this setting the Learner plays at round t an action i with probability $w_{t,i}$ and then is only revealed the loss $\ell_{t,i}$ of action i (The losses of the other actions at round t remain hidden). This problem, which was thoroughly studied in the seminal work of Auer et al. [2002], can be solved as well using an exponential weighting scheme. Of course, since only a small part of the vector $\boldsymbol{\ell}_t$ is revealed, the weights are computed using an *estimate* of the vector $\boldsymbol{\ell}_t$, with a cost for the resulting variance. The resulting bound is of the form

$$\text{Loss}_{\text{Alg}} = k + O(\sqrt{kn \log n}).$$

In addition, information-theoretic techniques are used to give a lower bound of the form $\text{Loss}_{\text{Alg}} - k = \Omega(\sqrt{kn})$, and hence the algorithm is essentially tight.

[*] Supported by a Yahoo! PhD Fellowship, DARPA grant FA8750-05-2-0249, and NSF grant DMS-0707060.

[†] Supported by NSF grant IIS 0325363.

This brings us to the proposed open problem. Despite the typical hardness of optimally solving general multi-round games, many such online decision games [Takimoto and Warmuth, 2000, Abernethy et al., 2008a,b, 2007] have been analyzed completely, with efficient solutions in each case. Yet, to our knowledge, no such minimax solution has been given for any decision game in the *bandit* setting. This would be particularly enlightening, as many bandit learning problems remain somewhat elusive. Several currently known bounds, for example, are not known to be tight.

Ideally, the proposed game would have a simple characterization as given in Abernethy et al. [2008b] for the hedge setting. We conclude by highlighting a number of important questions that must be addressed:

- What is a natural "sufficient statistic" of the state of the game? In the full-information setting, the vector s recorded all relevant information, yet in the bandit setting our observations are more complex; indeed, they even depend on the previous choices we made.

- What is a natural "stopping criterion" for the game? In the hedge setting we used a bound k on the loss of the best action, and it may be reasonable to keep this restriction. Yet perhaps a bound on the time of the game would give an easier analysis.

- In the hedge setting game, the optimal strategy can be characterized using the notion of a "random playout", where we imagine an adversary uniformly assigning losses to the experts. Does the minimax strategy in the bandit setting admit a similar characterization?

References

J. Abernethy, J. Langford, and M. K. Warmuth. Continuous experts and the Binning algorithm. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT06)*, pages 544–558. Springer, June 2007.

J. Abernethy, P. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2008a.

J. Abernethy, M. K. Warmuth, and J. Yellin. When random play is optimal against an adversary. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT 08)*, pages 437–445, July 2008b.

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

Y. Freund and R. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*, 55:119–139, 1997.

N. Littlestone and M. K. Warmuth. The Weighted Majority algorithm. *Inform. Comput.*, 108(2):212–261, 1994. Preliminary version in FOCS 89.

E. Takimoto and M. Warmuth. The minimax strategy for gaussian density estimation. In *Proc. 13th Annu. Conference on Comput. Learning Theory*, pages 100–106. Morgan Kaufmann, San Francisco, 2000.

V. Vovk. A game of prediction with expert advice. *J. Computer System Sci.*, 56(2):153–173, April 98.