JMLR: Workshop and Conference Proceedings 19 (2011) 817–819 24th Annual Conference on Learning Theory

A simple multi-armed bandit algorithm with optimal variation-bounded regret

Elad Hazan EHAZAN@IE.TECHNION.AC.IL
Technion
Haifa, Israel

Satyen Kale SKALE@YAHOO-INC.COM
Yahoo! Research
Santa Clara, CA 95054

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

We pose the question of whether it is possible to design a simple, linear-time algorithm for the basic multi-armed bandit problem in the adversarial setting which has a regret bound of $O(\sqrt{Q \log T})$, where Q is the total quadratic variation of all the arms.

We are interested in the fundamental multi-armed bandit (MAB) problem: iteratively at times $t = 1, 2, \ldots, T$ the decision maker has to choose (possibly randomly) one of n arms, i_t, and then receives the payoff of the arm, assumed to be in the range $[0, 1]$. The payoffs are constructed adversarially, as in (Auer et al., 2003), and we denote the payoff at time t for arm i by $f_t(i) \in [0, 1]$. The decision maker can only see her own payoff, and does not have access to the entire payoff vector f_t (otherwise it would have been the usual "experts" problem"). The goal is to minimize regret:

$$\text{Regret} = \max_{i \in [n]} \sum_{t=1}^{T} f_t(i) - \sum_{t=1}^{T} f_t(i_t),$$

where i_t is the arm chosen by the algorithm in round t. If the algorithm is randomized, then the aim is to minimize the expected regret over the internal randomization of the algorithm.

The EXP3 algorithm of Auer et al. (2003), is an efficient (linear in n time) algorithm that obtains regret of $O(\sqrt{nT \log T})$. This is optimal in T, n up to logarithmic factors.

But the quest for an optimal algorithm for this fundamental problem is not over. Cesa-Bianchi et al. (2007) conjectured that it should be possible to bound the regret of online learning algorithms by the *quadratic variation* in payoffs. For the MAB problem as defined above, we can define the quadratic variation by:

$$Q = \sum_{i=1}^{T} \|f_t - \mu\|^2,$$

where $\mu = \frac{1}{T} \sum_{t=1}^{T} f_t$ is the mean payoff. This is a natural parameter for measuring the difficulty of an online instance, as argued in Cesa-Bianchi et al. (2007), since it is related to the statistical properties of the underlying payoff sequences. Essentially, we may think

of the quadratic variability as the variance in our data, and the difficulty of learning should be proportional to how much the data deviates from the mean rather than the length of the prediction sequence. As a special case, Q can be a constant independent of the data sequence, in which case we would like our regret to remain constant independent of the number of iterations (this is motivated by financial applications, as in (Hazan and Kale, 2009b)).

Recently, online learning algorithms that bound the regret as a function of Q rather than T have been developed. For the online linear optimization setting this was obtained in (Hazan and Kale, 2008), and for the MAB setting, the following theorem was proven in (Hazan and Kale, 2009a):

Theorem 1 *There exists a polynomial-time MAB algorithm whose regret is bounded by* $O(n^2\sqrt{Q}\log T)$.

Since our payoffs are bounded by one, it holds that $Q \leq nT$, and hence as T grows large the above bound is superior to the EXP3 bound (for certain ranges of Q). However, the algorithm used to obtain this bound is rather complicated: it is based on self-concordant barrier functions as regularizers, which were introduced to learning theory in (Abernethy et al., 2008), and applies to a more general setting of bandit online linear optimization than MAB. This technology makes the algorithm poly-time, but not nearly linear time and simple as EXP3. More importantly, the above bound is sub-optimal in terms of n.

The open question is to design a simple, linear-time algorithm for MAB which has a regret bound of $O(\sqrt{Q\log T})$, hence improving upon EXP3.

We conjecture that such an algorithm exists, and it should not use any self-concordance technology. Rather, it should be basic, perhaps based on the multiplicative updates method, and bear resemblence to EXP3. We note that EXP3 itself has $\Omega(\sqrt{T})$ regret, since it mixes with the uniform distribution every iteration to enable sufficient exploration. Hence, the desired algorithm should be a little different from EXP3, incorporating just enough exploration proportional to the variation in the data.

One possible feature of the new algorithm is to use an unbiased estimator for the payoff vector f_t constructed by estimating the empirical mean and the deviation from the mean separately, as done in (Hazan and Kale, 2009a). An unbiased estimator for the mean can be constructed using the reservoir sampling ideas in (Hazan and Kale, 2009a). The deviation from the mean can be computed using importance weighted sampling as in EXP3.

References

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, 2008.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003. ISSN 0097-5397.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2-3):321–352, 2007. ISSN 0885-6125.

E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. In *COLT*, 2008.

E. Hazan and S. Kale. Better algorithms for benign bandits. In *SODA*, pages 38–47, 2009a.

E. Hazan and S. Kale. On stochastic and worst-case models for investing. In *NIPS*, 2009b.