Reducing Bandit problems to expert problems

Baruch Awerbuch

June 15, 2003

In this note, we consider the multi-armed bandit problem of [1] with K options for duration T. We and provide a simple modular reduction to the best experts problem [3, 4, 2].

Notations. Let $0 \leq \gamma(t, j) \leq 1$ be the cost of j's option at time t. Let us divide the time into T/τ phases $\mathcal{T}_i$ of length $\tau = K/\delta$. Consider the average cost of j's option in phase i

$$\beta(j, \mathcal{T}_i) = \mathrm{Exp}_{t \in \mathcal{T}_i} \gamma(t, j)$$

The average grade of each option is

$$\beta(j) = \mathrm{Exp}_{1 \leq i \leq T/\tau} \beta(j, \mathcal{T}_i)$$

and let the best option be $i^* = \mathsf{argmin}_i \beta(i)$.

The solution. With probability δ we sample one of the options at random, and with probability $1 - \delta$, we exploit a "black box" experts algorithm, which picks at each phase i option j with probability $\pi_i(j)$. The experts's distributed is fixed in each phase, and it is based on the feedback from *previous* phases. Note that we sampling each option at least constant number of times in a phase i; one of these samples, selected at random, will be the feedback for the experts algorithm for this option in phase i. The expectation of this feedback is exactly $\beta(j, \mathcal{T}_i)$. The total expected cost experienced in phase i is

$$\gamma_i = \sum_j \pi_i(j) \cdot \beta(j, \mathcal{T}_i)$$

Note that the experts algorithm runs for $\tilde{T} = T/\tau$ steps, and thus per [3, 4, 2] has an error of

$$\tilde{\epsilon} = \frac{\log K}{\epsilon \cdot \tilde{T}} + \epsilon = \frac{\log K}{\epsilon \cdot T/\tau} + \epsilon = \frac{\log K \cdot K}{\epsilon \cdot T \cdot \delta} + \epsilon$$

The error experience by this algorithm consists of the experts errors, and e a per-step error of δ that is attributed to sampling. The total error is

$$\tilde{\epsilon} + \delta = \frac{\log K \cdot K}{\epsilon \cdot T \cdot \delta} + \epsilon + \delta$$

This is optimized by choosing $\epsilon = \delta$ in which case we have error of $\dfrac{\log K \cdot K}{T \cdot \delta^2} + 2\delta$ which is minimized by selecting $\delta^3 = \dfrac{\log K \cdot K}{T}$ yielding average error of $\delta = \sqrt[3]{\dfrac{\log K \cdot K}{T}}$

Note that [1] accomplishes smaller error, namely $\sqrt[2]{\dfrac{\log K \cdot K}{T}}$.

References

[1] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. IEEE Computer Society Press, Los Alamitos, CA, 1995.

[2] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. pages 322–331, 1995.

[3] Adam Kalai and Santosh Vempala. Geometric algorithms for online optimization, 2003. unpublished manuscript.

[4] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994. A preliminary version appeared in FOCS 1989.