The non-stochastic multi-armed bandit problem

by Auer, Cesa-Bianchi, Freund, Schapire

Note Title 10/7/2008

Brief summary: Devavrat Shah.

Setup. The multi-armed bandit problem, but with adversarial reward sequences. Specifically, let there be K machine/arms that are played over time indexed by $t \in \mathbb{N}$. An adversary decides the rewards for machines across time. Let $x_i(t) \in [0,1]$ denote reward of machine i at time t; $\bar{x}(t) = (x_1(t), \ldots, x_K(t))$ be rewards for time t and $\mathbf{x} = (\bar{x}(t))_{t \in \mathbb{N}}$ be the reward sequence for entire horizon.

Strategy. A player chooses machine to play at each time given information about past rewards. Specifically, players strategy decides action A_t at time t based on information of past rewards, i.e.

$$A_t : ([0,1] \times \{1, \ldots, K\})^{t-1} \to \{1, \ldots, K\}$$

The reward of player with strategy $(A_t)_{t \in \mathbb{N}}$ upto time T is:

$$G_A(T) := \sum_{t=1}^{T} x_{A_t}(t).$$

"Goodness" of a strategy. The 'regret' or 'loss' of strategy $(A_t)_{t \in \mathbb{N}}$ with respect to a strategy that plays actions $J = (j_t)_{t \in \mathbb{N}}$ upto time T is:

$$G_J(T) - G_A(T) := \sum_{t=1}^{T} \left(x_{j_t}(t) - x_{A_t}(t) \right).$$

So, shall we define "goodness" of A as till time T as:

$$\max_J \quad G_J(T) - G_A(T) \quad ?$$

A simple example. Suppose an adversary chooses a sequence of rewards s.t. at each time exactly one machine has reward 1 and rest have 0. And suppose adversary chooses this "golden" machine ev. time unif. at random. Then, there is no hope for any strategy to make more than T/K reward on average. But, "oracle" strategy will make reward T. Thus, worst-case regret is $\Theta(T)$ and leads to trivial/non-interesting scenario.

Weak Regret. Recall that in Lai-Robbins setup, we wish to play the "best" machine. This suggests following as a reasonable criteria.

Let, $$G_{max}(T) = \max_{i=1}^{K} \left[\sum_{t=1}^{T} x_i(t) \right].$$

And, weak regret of strategy A is:

$$G_{max}(T) - G_A(T).$$

Regret vs. Complexity of strategy. The above definition compares regret with respect to a "fixed machine rule". More generally, a comparison strategy can be more complex. Natural notion of complexity from this perspective would be: "how often is the machine changed?" Let $H(J(T))$ be complexity of sequence $J(T) = (j_t)_{t=1}^{T}$ defined as

$$H(J(T)) = 1 + \#\{t : j_t \neq j_{t+1}, 1 \leq t < T\}.$$

Then, what we will call (not authors) m-weak regret:

$$G^m_{max}(T) - G_A(T) \quad \text{with}$$

$$G^m_{max}(T) = \max_{\substack{J(T): \\ H(J(T)) \le m}} G_{J(T)}(T).$$

Main Results.

Result 1. Weak regret

(a) Existence of A (randomized) with weak regret scaling as $\sqrt{TK\ln K}$ in expectation.

(b) Lower bound on weak regret for any scheme scaling as $\sqrt{KT}$ in expectation.

(c) The result of (a) holds with probability $1-\delta$, with modified weak regret bound scaling as: $\sqrt{KT\ln(KT/\delta)}$.

Result 2. m-weak regret.

There exists strategy with m-weak regret scaling as $m\sqrt{KT\ln(KT)} + 2e\sqrt{\frac{KT}{\ln(KT)}}$.

Result 3. n-Expert's advise.

In the presence of n experts (defined in terms of probabilistic strategies), the weak regret (define in terms of using best expert) scales as

$$\sqrt{KT\ln N}.$$

Next, we describe Result 1(a) in detail. Rest will be omitted. Because, all algorithms/strategis are variants of the one for 1(a) and lack of time.

Result 1(a).

Strategy (Exp3).

Parameter: $\gamma \in (0,1]$

Variables: $w_i(t)$, $i = 1, \dots K$

- initially, $t = 1$

$$w_i(1) = 1, \quad i = 1, \dots K.$$

Update: for $t \geq 1$:

- set probabilities $p_i(t) = (1-\gamma)\dfrac{w_i(t)}{\sum_j w_j(t)} + \dfrac{\gamma}{K}$
- Draw action $A_t \equiv i_t$ at random with $\mathbb{P}(i_t = j) = p_j(t)$.
- Receive reward $x_{i_t}(t)$.
- For $j = 1, \dots, K$, set

$$\hat{x}_j(t) = \begin{cases} \dfrac{x_j(t)}{p_j(t)} & j = i_t \\ 0 & \text{o.w.} \end{cases} \Rightarrow \mathbb{E}\left[\hat{x}_j(t)\right] = x_j(t)$$

- update: $w_j(t+1) = w_j(t) \exp\left(\dfrac{\gamma \hat{x}_j(t)}{K}\right)$

$$\Rightarrow w_j(t+1) = \exp\left(\frac{\gamma}{K} \sum_{s \leq t} x_j(s) \cdot \mathbb{1}_{\{j = i_s\}}\right)$$

- $t \to t+1$ and repeat.

Theorem (3.1). For any $K > 0$, $\gamma \in (0,1]$, $T \geq 1$

$$G_{max}(T) - \mathbb{E}\left[G_{Exp3}(T)\right] \leq (e-1)\gamma\, G_{max}(T) + \frac{K \ln K}{\gamma}.$$

Best choice of γ:

$$\gamma^* = \min\left\{1, \sqrt{\frac{K \ln K}{(e-1)\, G_{max}}}\right\}$$

If $\gamma^* = 1$, it's a 'trivial' scenario. If not,

$$G_{max}(T) - \mathbb{E}\left[G_{Exp3}(T)\right] \leq 2\sqrt{K \ln K \cdot G_{max}^{(T)}(e-1)}$$

$$= O\left(\sqrt{TK \ln K}\right).$$

Proof of Theorem 3.1.

Some facts:

F1. $\hat{x}_i(t) := \frac{x_i(t)}{p_i(t)} \leq \frac{1}{p_i(t)} \leq \frac{K}{\gamma}$.

F2. $\mathbb{E}\left[\hat{x}_i(t)\right] = \mathbb{P}\left(i = i_t\right) \cdot \frac{x_i(t)}{p_i(t)} = x_i(t)$.

F3. $\sum_{i=1}^{K} p_i(t) \cdot \hat{x}_i(t) = \sum_{i=1}^{K} p_i(t) \cdot \mathbb{1}\{i = i_t\} \cdot \frac{x_i(t)}{p_i(t)}$

$= x_{i_t}(t)$.

F4. $\sum_{i=1}^{K} p_i(t) \cdot \hat{x}_i^2(t) = \sum_{i=1}^{K} p_i(t) \cdot \mathbb{1}\{i = i_t\} \cdot \frac{x_i^2(t)}{p_i^2(t)}$

$= \frac{1}{p_{i_t}(t)} \cdot x_{i_t}^2(t) = \hat{x}_{i_t}(t) \cdot x_{i_t}(t) \leq x_{i_t}(t)$.

F5. $W_i(t) = \exp\left(\frac{\gamma}{K} \sum_{s \leq t} \hat{x}_i(s) \cdot \mathbb{1}\{i = i_s\}\right)$

Notation: $W_t := \sum_{i=1}^{K} W_i(t)$

$$\frac{W_{t+1}}{W_t} := \sum_{i=1}^{K} \frac{w_i(t+1)}{W_t} = \sum_{i=1}^{K} \frac{w_i(t)}{W_t} \cdot \exp\left(\frac{\gamma \hat{x}_i(t)}{K}\right)$$

$$\overset{(a)}{=} \sum_{i=1}^{K} \left[\frac{p_i(t) - \gamma/K}{1-\gamma}\right] \exp\left(\gamma \cdot \frac{\hat{x}_i(t)}{K}\right)$$

$$\overset{(b)}{\leq} \sum_{i=1}^{K} \left[\frac{p_i(t) - \gamma/K}{1-\gamma}\right] \cdot \left[1 + \frac{\gamma \hat{x}_i(t)}{K} + (e-2)\frac{\gamma^2}{K^2}\hat{x}_i^2(t)\right]$$

$$\leq \left[\sum_{i=1}^{K} \frac{(p_i(t) - \gamma/K)}{1-\gamma}\right] + \sum_{i=1}^{K} \frac{\gamma}{K} \cdot \frac{p_i(t)}{1-\gamma} \cdot \hat{x}_i(t) + \frac{\gamma^2(e-2)}{K^2(1-\gamma)} \cdot \left[\sum_{i=1}^{K} p_i(t)\, \hat{x}_i^2(t)\right]$$

$$\overset{(c)}{\leq} 1 + \frac{\gamma}{K(1-\gamma)} \cdot x_{i_K}(t) + \frac{(e-2)\gamma^2}{K^2(1-\gamma)}\left[\sum_{i=1}^{K} \hat{x}_i(t)\right] \quad - (1)$$

(a) . by def of $w_i(\cdot)$; (b) : $1 + e^x \leq 1 + x + (e-2)x^2$; $x \in [0,1]$

(c) F1-F4.

Using (1) and $\ln(1+x) \le x$ for $x > 0$, we have

$$\ln\left(\frac{W_{t+1}}{W_t}\right) \le \frac{\gamma}{K(1-\gamma)} x_{i_t}(t) + \frac{\gamma^2(e-2)}{k^2(1-\gamma)} \left[\sum_{j=1}^{K} \hat{x}_i(t)\right]$$

Therefore,

$$\ln\left(\frac{W_{T+1}}{W_1}\right) \le \frac{\gamma}{(1-\gamma)K}\left[\sum_{t=1}^{T} x_{i_t}(t)\right] + \frac{\gamma^2(e-2)}{k^2(1-\gamma)}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} \hat{x}_i(t)\right].$$

— (2)

And, for any j:

$$\ln\left(\frac{W_{T+1}}{W_1}\right) \ge \ln\left(\frac{W_j(T+1)}{W_1}\right) \ge \frac{\gamma}{k}\sum_{t=1}^{T} \hat{x}_j(t) - \ln K.$$

— (3)

(2) + (3) imply: use $G_{Exp3} \equiv \sum_{t=1}^{T} x_{i_t}(t)$.

$$G_{Exp3} \ge (1-\gamma)\sum_{t=1}^{T} \hat{x}_j(t) - \frac{K \ln K}{\gamma} - (e-2)\frac{\gamma}{k}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} \hat{x}_i(t)\right]$$

Taking expectation w.r.t. algorithm's randomization.

$$\mathbb{E}\left[G_{Exp3}\right] \geq (1-\gamma)\sum_{t=1}^{T} x_j(t) - \frac{K\ln K}{\gamma} - \frac{(e-2)\gamma}{K}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} x_i(t)\right]$$

Clearly, $\frac{1}{K}\sum_{t=1}^{T}\sum_{i=1}^{K} x_i(t) \leq G_{max}$; and choose $j = j^*$ that maximizes in G_{max}, to obtain

$$\mathbb{E}\left[G_{Exp3}\right] \geq (1-\gamma)\cdot G_{max} - \frac{K\ln K}{\gamma} - (e-2)\gamma\, G_{max}$$

Therefore:

$$G_{max} - \mathbb{E}\left[G_{Exp3}\right] \leq \frac{K\ln K}{\gamma} + (e-1)\gamma\, G_{max}.$$

#