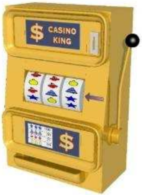
Nonstochastic Bandits and Partial Monitoring

Nicolò Cesa-Bianchi

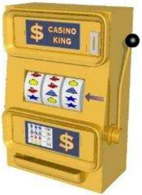
Università degli Studi di Milano

The bandit problem

[Robbins, 1952]



. . .



N slot machines

- Rewards $X_{i,1}, X_{i,2}, \ldots$ of machine i are i.i.d. random variables
- An allocation policy prescribes which machine I_t to play at time t based on the realization of $X_{I_1,1}, \ldots, X_{I_{t-1},t-1}$
- Want to play as often as possible the machine with largest reward expectation

$$\mu^* = \max_{i=1,\ldots,N} \mathbb{E}\, X_{i,1}$$

Finite-time regret

Definition (Regret after n plays)

$$\mu^* n - \sum_{t=1}^{n} \mathbb{E}\, X_{I_t,t}$$

Theorem (Lai and Robbins, 85)

There exist allocation policies satisfying

$$\mu^* n - \sum_{t=1}^{n} \mathbb{E}\, X_{I_t,t} \leqslant c\, N \ln n$$

uniformly over n

Horizon-dependent reward distributions

Fact

For each n, there are simple reward distributions such that the regret of any allocation policy is at least order of $\sqrt{nN}$

- Fix arbitrary policy A
- Assume $\{0, 1\}$-valued rewards are generated by fair coin flips
- Increase by $\sqrt{N/n}$ the expectation μ_k of a random machine k

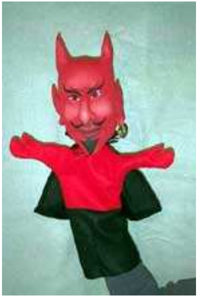
Proof sketch

- T_i = number of times i was chosen by A in the n plays
- Total reward of k increases by $n\sqrt{N/n} = \sqrt{nN}$
- $\mathbb{E}\, T_k$ increases by at most αn
- Total reward of A increases by at most $\alpha n\sqrt{N/n} = \alpha\sqrt{nN}$
- Regret is at least $(1-\alpha)\sqrt{nN}$

The nonstochastic bandit problem

[Auer, C-B, Freund, and Schapire, 2002]

What if probability is removed altogether?



Nonstochastic bandits

Bounded real rewards $x_{i,1}, x_{i,2}, \ldots$ are deterministically assigned to each machine i

- Analogies with repeated play of an unknown game [Baños, 1968; Megiddo, 1980]
- Allocation policies are allowed to randomize

0	1	0	0	7	9	9	8	9	0	0	1
5	7	9	6	0	0	2	2	0	0	0	1
0	2	0	1	0	1	0	0	8	9	8	7

Definition (Regret)

$$\max_{i=1,\dots,N}\left(\sum_{t=1}^{n} x_{i,t}\right) - \mathbb{E}\left[\sum_{t=1}^{n} x_{I_t,t}\right]$$

A nearly optimal randomized policy

- Reward estimates $\displaystyle \widehat{x}_{i,t} = \frac{x_{i,t}}{p_{i,t}} \mathbb{I}_{\{I_t = i\}}$
- Note

$$\mathbb{E}\left[\widehat{x}_{i,t} \,\middle|\, I_1, \dots, I_{t-1} \right] = \frac{x_{i,t}}{p_{i,t}} \times p_{i,t} + 0 \times (1 - p_{i,t}) = x_{i,t}$$

- Weights. At time t, machine i is assigned weight

$$w_{i,t-1} = \exp\left(\frac{\gamma}{N} \sum_{s=1}^{t-1} \widehat{x}_{i,s} \right)$$

- Randomization. At time t, machine i is played with prob.

$$(1-\gamma)\frac{w_{i,t-1}}{W_{t-1}} + \frac{\gamma}{N}$$

Regret bounds

$$G_n^* = \max_{i=1,\ldots,N} \sum_{t=1}^{n} x_{i,t} \qquad \text{and} \qquad \widehat{G}_n = \sum_{t=1}^{n} x_{I_t,t}$$

Theorem

$$G_n^* - \mathbb{E}\,\widehat{G}_n \quad \leqslant \quad 2\sqrt{2}\sqrt{nN \ln N}$$

- Lower bound was $\sqrt{nN}$
- Adaptive choice of γ avoids fixing the horizon n

Variance problem

- Variance of payoff estimates

$$\mathrm{VAR}\left[\widehat{x}_{i,t}\right] \approx \frac{1}{p_{i,t}^2} \times p_{i,t} \approx \frac{N}{\gamma} \approx \sqrt{\frac{nN}{\ln N}}$$

- Overall variance

$$\sum_{t=1}^{n} \mathrm{VAR}\left[\widehat{x}_{i,t}\right] \approx n^{3/2}$$

- Thus, with constant probability, the regret can be of the order of

$$\sqrt{\sum_{t=1}^{n} \mathrm{VAR}\left[\widehat{x}_{i,t}\right]} \approx n^{3/4}$$

Bounding the regret in probability

- Low-variance estimates

$$\widehat{x}_{i,t} = \frac{x_{i,t}}{p_{i,t}} \mathbb{I}_{\{I_t = i\}} + \frac{\beta}{p_{i,t}}$$

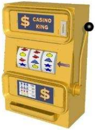
- Then, with high probability

$$\sum_{t=1}^{n} x_{i,t} \leqslant \sum_{t=1}^{n} \widehat{x}_{i,t} + \beta n N \qquad \text{for all } i = 1, \dots, N$$

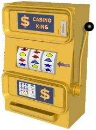
- Choosing $\beta \approx \sqrt{(\ln N)/(nN)}$

$$G_n^* - \widehat{G}_n \leqslant \frac{11}{2} \sqrt{nN \ln \frac{N}{\delta}} + \frac{\ln N}{2} \qquad \text{w.p. at least } 1 - \delta$$

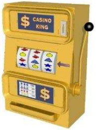
Competing against arbitrary policies



0 1 0 0 7 9 9 8 9 0 0 1



5 7 9 6 0 0 2 2 0 0 0 1



0 2 0 1 0 1 0 0 8 9 8 7

Tracking regret

- Regret against an arbitrary and unknown policy $(j_1, j_2, \dots, j_n)$

$$\sum_{t=1}^{n} x_{j_t,t} - \mathbb{E}\left[\sum_{t=1}^{n} x_{I_t,t}\right]$$

- Weight sharing technique

$$w_{i,t} = w_{i,t-1} \exp\left(\frac{\gamma}{N}\widehat{x}_{i,t}\right) + \frac{\alpha}{N}\sum_{j=1}^{N} w_{j,t-1}$$

A bound on the tracking regret

Definition (Complexity of a policy)

$(j_1, j_2, \ldots, j_n)$ is number of times the policy switches to a different machine

Theorem

For all fixed S, the regret of weight sharing against any policy of complexity bounded by S is at most

$$\sqrt{S\, nN \ln N}$$

Repeated games

Payoffs are negative (losses) and come from a known loss matrix with entries in $[0, 1]$

$$
\begin{array}{c|ccc|}
 & & \text{outcomes} & \\
 & 1 & \cdots & M \\
\hline
1 & \ell(1,1) & \cdots & \ell(1,M) \\
\vdots & \vdots & \ell(I_t, y_t) & \vdots \\
N & \ell(N,1) & \cdots & \ell(N,M) \\
\hline
\end{array}
$$

After drawing I_t the forecaster observes y_t

	1	$\cdots$	y_t	$\cdots$	M
1	$\ell(1,1)$		$\ell(1,y_t)$		$\ell(1,M)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
N	$\ell(N,1)$		$\ell(N,y_t)$		$\ell(N,M)$

Regret: $\sqrt{n \ln N}$

Nonstochastic bandits

After drawing I_t the forecaster observes $\ell(I_t, y_t)$

	1	$\cdots$		$\cdots$	M
1	$\ell(1,1)$				$\ell(1,M)$
$\vdots$	$\vdots$		$\ell(I_t,y_t)$		$\vdots$
N	$\ell(N,1)$				$\ell(N,M)$

Regret: $\sqrt{nN \ln N}$

Partial monitoring

[Rustichini, 1999]

After drawing I_t the forecaster observes $h(I_t, y_t)$

$$
\begin{array}{|ccc|}
1 & & M \\
\hline
\ell(1,1) & & \ell(1,M) \\
\vdots & \ell(I_t, y_t) & \vdots \\
\ell(N,1) & & \ell(N,M) \\
\hline
\end{array}
\qquad
\begin{array}{|ccc|}
1 & & M \\
\hline
h(1,1) & & h(1,M) \\
\vdots & h(I_t, y_t) & \vdots \\
h(N,1) & & h(N,M) \\
\hline
\end{array}
$$

Loss matrix L Feedback matrix H

In the bandit case, $H \equiv L$

The revealing action game (apple tasting)

[Helmbold, Littlestone, and Long, 2000]

	0	1
0	0	1
1	1	0
2	1	1

L

	0	1
0	a	a
1	a	a
2	b	c

H

Dynamic pricing

- Forecaster's action $I_t \in \{1, 2, \dots, N\}$ is the price at which a product sold online is offered to t-th customer
- Adversary's action $y_t \in \{1, 2, \dots, N\}$ is maximum price at which t-th customer is willing to buy the product
- Loss matrix arbitrary
- Feedback matrix

$$h(I_t, y_t) = \begin{cases} \textsc{sold} & \text{if } I_t \leqslant y_t \\ \textsc{not sold} & \text{otherwise} \end{cases}$$

Controlling the regret

- Sufficient (and almost necessary) condition

$$L = K\,H \qquad \text{for some matrix } K$$

- Define

$$\widehat{\ell}(i, y_t) = \frac{k(i, I_t)\,h(I_t, y_t)}{p_{I_t,t}}$$

- Since $L = K\,H$

$$\mathbb{E}\left[\widehat{\ell}(i, y_t)\,\middle|\, I_1, \ldots, I_{t-1}\right] = \sum_{j=1}^{N} \frac{k(i, j)\,h(j, y_t)}{p_{j,t}} \times p_{j,t} = \ell(i, y_t)$$

Upper bound

[C-B, Lugosi, and Stoltz, 2005]

Theorem

There exists a forecaster whose regret is with high probability at most

$$c(Nn)^{2/3}(\ln N)^{1/3}$$

for any partial monitoring game (L, H) *satisfying* $L = K\,H$ *for some* K

Lower bound

	0	1
0	0	1
1	1	0
2	1	1

L

	0	1
0	a	a
1	a	a
2	b	c

H

Theorem

In the revealing action game, if a forecaster plays the revealing action at most m times, then its regret is at least

$$c_1\, m + c_2\, \frac{n}{\sqrt{m}}$$

for some sequence $y_1, \ldots, y_n$

Gap theorem

In any partial monitoring problem,

- either the regret is $\Omega(n)$ for all forecasters
- or there exists a forecaster whose regret is $O\left(n^{2/3}\right)$