Nonstochastic Bandits
and Partial Monitoring

Nicolò Cesa-Bianchi

Università degli Studi di Milano
The bandit problem

- 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 $X_{I_{t-1},1}, X_{I_{t-1},2}, \ldots, X_{I_{t-1},t-1}$.
- 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

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

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 $\mu_k$ 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 $\alpha n$
- 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
**Definition (Regret)**

\[
\max_{i=1,\ldots,N} \left( \sum_{t=1}^{n} x_{i,t} \right) - \mathbb{E} \left[ \sum_{t=1}^{n} x_{I_t,t} \right]
\]
A nearly optimal randomized policy

- **Reward estimates**
  \[ \hat{x}_{i,t} = \frac{x_{i,t}}{p_{i,t}} I\{I_t=i\} \]

- **Note**
  \[ \mathbb{E}\left[ \hat{x}_{i,t} \mid I_1, \ldots, I_{t-1} \right] = \frac{x_{i,t}}{p_{i,t}} \times p_{i,t} + 0 \times (1 - p_{i,t}) = x_{i,t} \]

- **Weights.** At time \( t \), machine \( i \) is assigned weight
  \[ w_{i,t-1} = \exp \left( \frac{\gamma}{N} \sum_{s=1}^{t-1} \hat{x}_{i,s} \right) \]

- **Randomization.** At time \( t \), machine \( i \) is played with prob.
  \[ (1 - \gamma) \frac{w_{i,t-1}}{W_{t-1}} + \frac{\gamma}{N} \]
Theorem

\[ G_n^* = \max_{i=1,\ldots,N} \sum_{t=1}^{n} x_{i,t} \quad \text{and} \quad \hat{G}_n = \sum_{t=1}^{n} x_{I_t,t} \]

\[ G_n^* - \mathbb{E} \hat{G}_n \leq 2 \sqrt{2 \sqrt{nN \ln N}} \]

- Lower bound was \( \sqrt{nN} \)
- Adaptive choice of \( \gamma \) avoids fixing the horizon \( n \)
Variance problem

- Variance of payoff estimates

\[ \text{VAR} [\hat{x}_{i,t}] \approx \frac{1}{p_{i,t}^2} \times p_{i,t} \approx \frac{N}{\gamma} \approx \sqrt{\frac{nN}{\ln N}} \]

- Overall variance

\[ \sum_{t=1}^{n} \text{VAR} [\hat{x}_{i,t}] \approx n^{3/2} \]

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

\[ \sqrt{\sum_{t=1}^{n} \text{VAR} [\hat{x}_{i,t}]} \approx n^{3/4} \]
Bounding the regret in probability

- **Low-variance estimates**

\[
\hat{x}_{i,t} = \frac{x_{i,t}}{p_{i,t}} \mathbb{I}_{\{I_t=i\}} + \frac{\beta}{p_{i,t}}
\]

- Then, with high probability

\[
\sum_{t=1}^{n} x_{i,t} \leq \sum_{t=1}^{n} \hat{x}_{i,t} + \beta nN \quad \text{for all } i = 1, \ldots, N
\]

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

\[
G_n^* - \hat{G}_n \leq \frac{11}{2} \sqrt{nN \ln \frac{N}{\delta}} + \frac{\ln N}{2} \quad \text{w.p. at least } 1 - \delta
\]
Competing against arbitrary policies

\[
\begin{array}{cccccccccc}
0 & 1 & 0 & 0 & 7 & 9 & 9 & 8 & 9 & 0 & 0 & 1 \\
5 & 7 & 9 & 6 & 0 & 0 & 2 & 2 & 0 & 0 & 0 & 1 \\
0 & 2 & 0 & 1 & 0 & 1 & 0 & 0 & 8 & 9 & 8 & 7
\end{array}
\]
Tracking regret

- **Regret against an arbitrary and unknown policy**
  \[(j_1, j_2, \ldots, j_n)\]
  \[
  \sum_{t=1}^{n} x_{j_t,t} - \mathbb{E} \left[ \sum_{t=1}^{n} x_{I_t,t} \right]
  \]

- **Weight sharing technique**
  \[
  w_{i,t} = w_{i,t-1} \exp \left( \frac{\gamma}{N} \hat{x}_{i,t} \right) + \frac{\alpha}{N} \sum_{j=1}^{N} w_{j,t-1}
  \]
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

\[\sqrt{S \, nN \ln N}\]
Payoffs are negative (losses) and come from a known **loss matrix** with entries in $[0, 1]$. 

<table>
<thead>
<tr>
<th>outcomes</th>
<th>1</th>
<th>$\cdots$</th>
<th>$M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\ell(1,1)$</td>
<td>$\cdots$</td>
<td>$\ell(1,M)$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\ell(I_t,y_t)$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$N$</td>
<td>$\ell(N,1)$</td>
<td>$\cdots$</td>
<td>$\ell(N,M)$</td>
</tr>
</tbody>
</table>
After drawing $I_t$ the forecaster observes $y_t$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>$\ldots$</th>
<th>$y_t$</th>
<th>$\ldots$</th>
<th>$M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\ell(1,1)$</td>
<td>$\ell(1,y_t)$</td>
<td>$\ell(1,M)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>$\ell(N,1)$</td>
<td>$\ell(N,y_t)$</td>
<td>$\ell(N,M)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Regret: $\sqrt{n \ln N}$
After drawing $I_t$ the forecaster observes $\ell(I_t, y_t)$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>$\ell(1, 1)$</th>
<th>$\ell(1, M)$</th>
<th>$\ell(I_t, y_t)$</th>
<th>$\ell(N, 1)$</th>
<th>$\ell(N, M)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>$\ell(1, M)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\ell(I_t, y_t)$</td>
<td>$\vdots$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>$\ell(N, 1)$</td>
<td></td>
<td>$\ell(N, M)$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Regret: $\sqrt{nN \ln N}$
After drawing $I_t$ the forecaster observes $h(I_t, y_t)$

Loss matrix $L$  

<table>
<thead>
<tr>
<th>1</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell(1,1)$</td>
<td>$\ell(1,M)$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\ell(I_t,y_t)$</td>
</tr>
<tr>
<td>$\ell(N,1)$</td>
<td>$\ell(N,M)$</td>
</tr>
</tbody>
</table>

Feedback matrix $H$  

<table>
<thead>
<tr>
<th>1</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(1,1)$</td>
<td>$h(1,M)$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$h(I_t,y_t)$</td>
</tr>
<tr>
<td>$h(N,1)$</td>
<td>$h(N,M)$</td>
</tr>
</tbody>
</table>

In the bandit case, $H \equiv L$
The revealing action game (apple tasting)

[Helmbold, Littlestone, and Long, 2000]

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

H

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>c</td>
</tr>
</tbody>
</table>
Dynamic pricing

- Forecaster’s action $I_t \in \{1, 2, \ldots, N\}$ is the price at which a product sold online is offered to $t$-th customer.

- Adversary’s action $y_t \in \{1, 2, \ldots, N\}$ is maximum price at which $t$-th customer is willing to buy the product.

- Loss matrix arbitrary.

- Feedback matrix

$$h(I_t, y_t) = \begin{cases} 
\text{SOLD} & \text{if } I_t \leq y_t \\
\text{NOT SOLD} & \text{otherwise}
\end{cases}$$
Controlling the regret

- Sufficient (and almost necessary) condition

\[ L = KH \quad \text{for some matrix } K \]

- Define

\[ \hat{\ell}(i, y_t) = \frac{k(i, I_t) h(I_t, y_t)}{p_{I_t,t}} \]

- Since \( L = KH \)

\[ \mathbb{E}\left[ \hat{\ell}(i, y_t) \mid I_1, \ldots, I_{t-1} \right] = \sum_{j=1}^{N} \frac{k(i, j) h(j, y_t)}{p_{j,t}} \times p_{j,t} = \ell(i, y_t) \]
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 \(L = KH\) for some \(K\)
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 $y_1, \ldots, y_n$
In any partial monitoring problem,

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