Better algorithms for sleeping experts & bandits

Gergely Neu
INRIA, SequeL team

joint work with Michal Valko, to appear at NIPS 2014
Prediction with expert advice
Prediction with expert advice
Prediction with expert advice
Prediction with expert advice
Prediction with expert advice
Prediction with expert advice
Prediction with sleeping experts
Prediction with sleeping experts
Prediction with sleeping experts
Prediction with sleeping experts
Prediction with sleeping experts
Prediction with sleeping experts
Prediction with sleeping experts
Parameters: set of $N$ experts

In each round $t = 1, 2, \ldots, T$

- Environment chooses losses $\ell_{t,i} \in [0, 1]$ for all experts
- Environment chooses the set of available experts $S_t \in \{1, 2, \ldots, N\}$
- Learner picks distribution $p_t$ on available experts
- Learner suffers loss $p_t^T l_t$
Usual notion of regret:

\[ R_T = \sum_{t=1}^{T} p_t^T l_t - \min_{i \in \{1, \ldots, N\}} \sum_{t=1}^{T} \ell_{t,i} \]
Usual notion of regret:

\[ R_T = \sum_{t=1}^{T} p_t^T l_t - \min_{i \in \{1, \ldots, N\}} \sum_{t=1}^{T} \ell_{t,i} \]

This comparator is pointless!
Usual notion of regret:

\[
R_T = \sum_{t=1}^{T} p_t^T l_t - \min_{i \in \{1, \ldots, N\}} \sum_{t=1}^{T} \ell_{t,i}
\]

We should actually compete with policies of the form \( \pi: 2^{[N]} \rightarrow N \) such that \( \pi(S) \in S \)!
Regret definition

Regret against policy class $\Pi$:

$$R_T = \sum_{t=1}^{T} p_t^T l_t - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell_{t,\pi(S_t)}$$
## Previous results

<table>
<thead>
<tr>
<th></th>
<th>IID availability</th>
<th>Adversarial availability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IID losses</strong></td>
<td>(that’s kind of trivial)</td>
<td>Kleinberg et al. (2008): $R_T = \Theta(\sqrt{T N \log N})$</td>
</tr>
<tr>
<td><strong>Adversarial losses</strong></td>
<td>Kanade et al. (2009): $R_T = O(\sqrt{T \log N})$</td>
<td>Kleinberg et al. (2008): $R_T = \Theta(N \sqrt{T \log N})$</td>
</tr>
</tbody>
</table>
# Previous results

<table>
<thead>
<tr>
<th></th>
<th>IID availability</th>
<th>Adversarial availability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IID losses</strong></td>
<td>(that’s kind of trivial)</td>
<td>Kleinberg et al. (2008): $R_T = \Theta(\sqrt{TN \log N})$</td>
</tr>
<tr>
<td><strong>Adversarial losses</strong></td>
<td>Kanade et al. (2009): $R_T = O(\sqrt{T \log N})$</td>
<td>Kleinberg et al. (2008): $R_T = \Omega(\sqrt{T \log N})$</td>
</tr>
</tbody>
</table>
## Previous results

<table>
<thead>
<tr>
<th></th>
<th>IID availability</th>
<th>Adversarial availability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IID losses</strong></td>
<td>(that’s kind of trivial)</td>
<td>Kleinberg et al. (2008): $R_T = \Theta(\sqrt{T N \log N})$</td>
</tr>
<tr>
<td><strong>Adversarial losses</strong></td>
<td>Kanade et al. (2009): $R_T = \Theta(\sqrt{T N \log N})$</td>
<td>Hard</td>
</tr>
</tbody>
</table>

*Cheat*
Where’s the cheat?
Where’s the cheat?
Kanade et al. assume that $\ell_{t,i}$ is observed for all $i \in [N]$!
Kanade et al. assume that $\ell_{t,i}$ is observed for all $i \in [N]$!

A more realistic assumption:

Observe $\ell_{t,i}$ only for $i \in S_t \subseteq [N]$
Algorithm: Follow the Perturbed Leader

Initialization: let $L_{t,i} = 0$ for all $i \in [N]$ 

For all rounds $t = 1, 2, \ldots, T$:

- Observe $S_t \subseteq [N]$ 
- Draw perturbations $Z_{t,i} \sim \text{Exp}(\eta)$ for all $i \in S_t$ 
- Play expert 
  \[ I_t = \arg \min_{i \in S_t} \left(L_{t-1,i} - Z_{t,i}\right) \]
- Observe feedback and set for all $i \in N$ 
  \[ L_{t,i} = L_{t-1,i} + \ell_{t,i} \]
Algorithm: Follow the Perturbed Leader

Initialization: let $\hat{L}_{t,i} = 0$ for all $i \in [N]$

For all rounds $t = 1, 2, \ldots, T$:

• Observe $S_t \subseteq [N]$

• Draw perturbations $Z_{t,i} \sim \text{Exp}(\eta)$ for all $i \in S_t$

• Play expert

$$I_t = \arg \min_{i \in S_t} (\hat{L}_{t-1,i} - Z_{t,i})$$

• Observe feedback and set for all $i \in N$

$$\hat{L}_{t,i} = \hat{L}_{t-1,i} + \hat{\ell}_{t,i}$$
Loss estimation

- Assume IID availability:
  \[ S_t \sim Q \quad \forall t = 1,2,\ldots,T \]
- Then we can set \( q_i = P[i \in S_t] \) for all \( i \in [N] \)
- Losses can be estimated as
  \[
  \hat{\ell}_{t,i} = \begin{cases} 
  \frac{\ell_{t,i}}{q_i}, & \text{if } i \text{ is observed} \\
  0, & \text{otherwise}
  \end{cases}
  \]
Assume IID availability:

\[ S_t \sim Q \quad \forall t = 1, 2, \ldots, T \]

Then we can set \( q_i = P[i \in S_t] \) for all \( i \in [N] \)

Losses can be estimated as

\[ \hat{\ell}_{t,i} = \begin{cases} \frac{\ell_{t,i}}{q_i}, & \text{if } i \text{ is observed} \\ 0, & \text{otherwise} \end{cases} \]

Unbiased:

\[ E[\hat{\ell}_{t,i}] = \ell_{t,i} \]
Loss estimation

- Assume IID availability:
  \[ S_t \sim Q \quad \forall t = 1, 2, \ldots, T \]
- Then we can set \( q_i = P[i \in S_t] \) for all \( i \in [N] \)
- Losses can be estimated as
  \[ \ell_{t,i} = \begin{cases} \frac{\ell_{t,i}}{q_i} & \text{if } i \text{ is observed} \\ 0 & \text{otherwise} \end{cases} \]

Unbiased:
\[ E[\hat{\ell}_{t,i}] = \ell_{t,i} \]

But the \( q_i \)'s are unknown!!
Idea:
• Use $K$ samples to estimate $Q$!
• Compute estimates of $q_i$!
• Obtain low-bias reward estimates!
Loss estimation – the bad way

Idea:
• Use $K$ samples to estimate $Q$!
• Compute estimates of $q_i$!
• Obtain low-bias reward estimates!

Bad news:
• Regret becomes $O(T^{3/4})$
• Can fail horribly for large action sets
Loss estimation – the right way

- Observe that the downtime is a geometric RV!
Loss estimation – the right way

- Observe that the downtime is a geometric RV!

\[ t \]

Arm \( i \)
Observe that the downtime is a geometric RV!

Arm $i$

$t$  $t + 1$
Observe that the downtime is a geometric RV!

Arm $i$

$t$  $t+1$  $t+2$
Observe that the downtime is a geometric RV!

\[ t \quad t + 1 \quad t + 2 \quad \ldots \quad t + K \]
Observe that the downtime is a geometric RV!

\[ K_{t,i} \]

Arm \( i \)

\[ t \quad t + 1 \quad t + 2 \quad t + K \]
Loss estimation – the right way

- Observe that the downtime is a geometric RV!

\[ t \quad t + 1 \quad t + 2 \quad t + K \]

Arm \( i \)

\[ E_t[K_{t,i}] = \frac{1}{q_i} \]
Observe that the downtime is a geometric RV! 

Arm $i$ 

Estimate losses as 

$$\hat{\ell}_{t,i} = \begin{cases} 
\ell_{t,i}K_{t,i}, & \text{if } i \text{ is observed} \\
0, & \text{otherwise} 
\end{cases}$$ 

$$E_t[K_{t,i}] = \frac{1}{q_i}$$
Theorem 1
Assuming IID expert availability, the expected regret of FPL fed with loss estimates \( \{\hat{\ell}_{t,i}\} \) satisfies

\[
R_T = O\left(\sqrt{TN \log N}\right)
\]
Main result

**Theorem 1**
Assuming IID expert availability, the expected regret of FPL fed with loss estimates \( \{\hat{\ell}_{t,i}\} \) satisfies

\[
R_T = O(\sqrt{TN \log N})
\]

- This is worse by a factor of \( \sqrt{N} \) than the bound of Kanade et al. (2009)…
Main result

Theorem 1
Assuming IID expert availability, the expected regret of FPL fed with loss estimates $\{\hat{l}_{t,i}\}$ satisfies

$$R_T = O\left(\sqrt{TN \log N}\right)$$

- This is worse by a factor of $\sqrt{N}$ than the bound of Kanade et al. (2009)...
- ...but we didn’t cheat!
Theorem 1: $R_T = O(\sqrt{TN \log N})$

Theorem 2
Assuming IID expert availability, no algorithm can achieve better regret than

$$R_T = \Omega(\sqrt{TN})$$
Assume that

- each expert \( i \in [N] \) is associated with a binary vector \( \mathbf{v}(i) \in \{0,1\}^d \)
- losses are described by a loss vector \( \mathbf{l}_t \in [0,1]^d \)
- loss of expert \( i \) in round \( t \) is given as \( \mathbf{v}(i)^\top \mathbf{l}_t \leq m \)
Extensions: large action spaces

Assume that
- each expert $i \in [N]$ is associated with a binary vector $\mathbf{v}(i) \in \{0,1\}^d$
- losses are described by a loss vector $\mathbf{l}_t \in [0,1]^d$
- loss of expert $i$ in round $t$ is given as $\mathbf{v}(i) \mathbf{l}_t \leq m$

Theorem 3
Assuming IID expert availability, the expected regret of the combinatorial extension of FPL is

$$R_T = O(m\sqrt{dT \log d})$$
So far: assume we observe $\ell_{t,i}$ for all $i \in S_t$

Now: assume we only observe the loss $\ell_{t,I_t}$

Using a simple extension of FPL, we prove
So far: assume we observe $\ell_{t,i}$ for all $i \in S_t$

Now: assume we only observe the loss $\ell_{t,I_t}$

Using a simple extension of FPL, we prove

**Theorem 4**
Assuming IID expert availability, the expected regret of the bandit extension of FPL satisfies

$$R_T = O(T^{2/3})$$
So far: assume we observe \( \ell_{t,i} \) for all \( i \in S_t \)

Now: assume we only observe the loss \( \ell_{t,I_t} \)

Using a simple extension of FPL, we prove

Theorem 4
Assuming IID expert availability, the expected regret of the bandit extension of FPL satisfies

\[ R_T = O(T^{2/3}) \]

Best previous result was \( O(T^{4/5}) \)
Experiments

sleeping bandits, 5 arms, varying availability, average over 20 runs

- BSFPL
- SleepingCat
- RandomGuess

Cumulative regret at time $T = 10000$ as a function of availability.
Experiments

sleeping semi-bandits, shortest path, 3 x 3 grid, average over 20 runs

- CombinatorialBSFPL
- CombinatorialSleepingCat
- CombinatorialRandomGuess

regret per step vs time t
Experiments

Sleeping semi-bandits, shortest path, 10 x 10 grid, average over 20 runs

- CombinatorialBSFPL
- CombinatorialSleepingCat
- CombinatorialRandomGuess

Regret per step

Time t
Future work

- Prove $R_T = O\left(\sqrt{T}\right)$ for sleeping bandits?
  - Problem: knowing the $q_i$’s is not enough
- Extend results to more complicated availability assumptions:
  - Markovian arms
  - Mortal arms
Thanks!