An Efficient Algorithm for Learning with Semi-Bandit Feedback

Gergely Neu
gergely.neu@gmail.com
INRIA Lille – Nord Europe

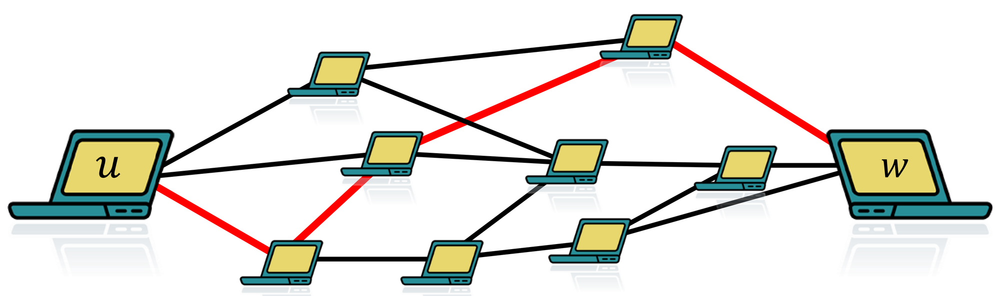
Gábor Bartók
bartok@inf.ethz.ch
ETH Zürich

The learning problem

- For each time step $t = 1,2,\dots,T$
  - Learner chooses action $V_t \in S \subseteq \{0,1\}^d$
  - Adversary selects loss vector $\ell_t \in [0,1]^d$
  - Learner suffers loss $V_t^\top \ell_t$
  - Learner observes feedback based on V_t and ℓ_t

Decision set:
$S = \{v_i\}_{i=1}^N \subseteq \{0,1\}^d$
$\|v_i\|_1 \leq m$

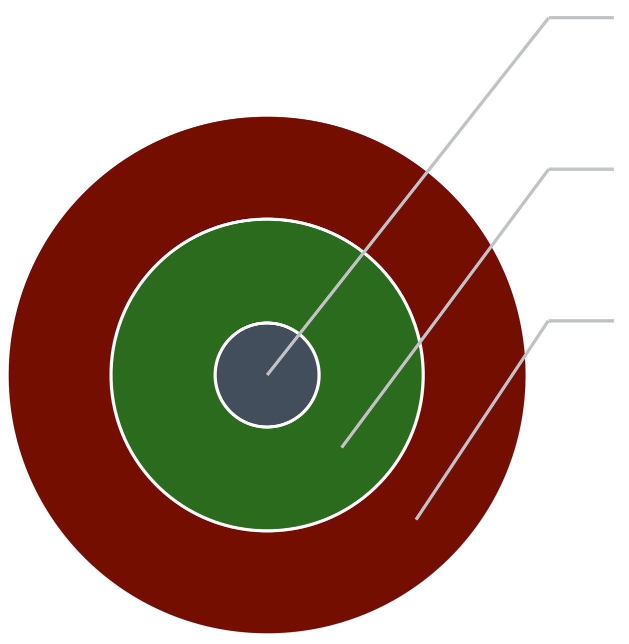
E.g: sequential routing



Goal: minimize (expected) *regret*

$$R_T = \max_{v \in S} \mathbf{E}\left[\sum_{t=1}^{T} (V_t - v)^\top \ell_t\right]$$

… under various feedback assumptions:



Follow the perturbed leader

Parameter: learning rate $\eta > 0$, $L_0 = 0$
For each time step $t = 1,2,\dots,T$
- Draw perturbation vector Z_t with $Z_{t,i} \sim \mathrm{Exp}(\eta)$ i.i.d. for all $i \in \{1,2,\dots,d\}$
- Choose $V_t = \arg\min_{v \in S} v^\top (L_{t-1,i} -$

FPL is efficient whenever the optimization $\min_{v \in S} v^\top \ell$ can be solved efficiently

BUT

"FPL is suboptimal by far!"
"FPL doesn't work with bandit feedback!"

The loss estimation problem

Traditional approach

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{\mathbf{E}_t[V_{t,i}]} V_{t,i}$$

… but how do we compute $\mathbf{E}_t[V_{t,i}]$ for FPL?
Poland (2005):
- assume they are given by an oracle
- use $O(T^2)$ samples

We need to estimate $1/q_{t,i}$, not $q_{t,i}$!
Observation: biased coin with $\mathbf{P}[\mathrm{H}] = q$

T T … T H

Expected number of tosses until first H: $1/q$

Loss estimation by Geometric Resampling

- Draw $V_t \sim \boldsymbol{p}_t$
- Observe $\{V_{t,i}\ell_{t,i}\}$
- Draw $V_t'(1), V_t'(2), \dots \sim \boldsymbol{p}_t$
- Let $K_{t,i} = \min\{k: V_t'(k) = 1\}$
- Let $\hat{\ell}_{t,i} = \ell_{t,i} K_{t,i} V_{t,i}$

Unbiased since
- $\mathbf{E}_t[V_{t,i}] = q_{t,i}$
- $\mathbf{E}_t[K_{t,i}] = 1/q_{t,i}$

"Low" variance:
- $\mathbf{E}_t\left[\sum_{i=1}^d q_{t,i}\hat{\ell}_{t,i}^2\right] \leq 1/q_{t,i}$

Results

Semi-bandit

Theorem:
$R_T^{FPL+GR} \leq 2m\sqrt{2dT(\log d + 1)}$

Full info

Theorem:
$R_T^{FPL} \leq 2m^{3/2}\sqrt{T(\log d + 1)}$

$\sqrt{d} \to \sqrt{m}$

But where does the sampling hurt?
- Had we known the $q_{t,i}$'s, we could do $2 \to \sqrt{2}$
- How much samples do we need?
  - Expectation: d
  - Worst-case: ∞

Stop sampling after M steps!
Additional regret: $\frac{dT}{eM}$

So what did we achieve?

	Full info	Semi-bandit	Full bandit	Efficient
EWA/EXP3	$m^{3/2}\sqrt{T\log(d/m)}$	$m\sqrt{dT\log(d/m)}$	$m^{3/2}\sqrt{dT\log(d/m)}$	sometimes
Mirror descent	$m\sqrt{T\log(d/m)}$	$\sqrt{mdT}$	???	sometimes
FPL	$m^{3/2}\sqrt{T\log d}$	$m\sqrt{dT\log d}$	???	always

Computational complexity
- $f(S) \triangleq$ Time to solve optimization on S
  - Shortest paths: $f(S) = O(d)$
  - Spanning trees: $f(S) = O(d\log d)$
  - Perfect matchings: $f(S) = O(md^2)$

Total running time:
- Expectation: $dTf(S)$
- Worst-case: $\sqrt{d}T^{3/2}f(S)/m$

Conclusion
- This is arguably the most efficient method for learning with semi-bandit feedback
- As a side result, we have proved that FPL is at least as good as EXP3

Future work

- Proving high probability bounds
  - Actually, not that difficult since the variance is not much higher…
  - … but we don't know how to compute upper confidence bounds efficiently
- Extending results to linear bandits with full bandit feedback
  - Use geometric expansion of the matrix inverse needed there?
- Can we strengthen guarantees for FPL even more?