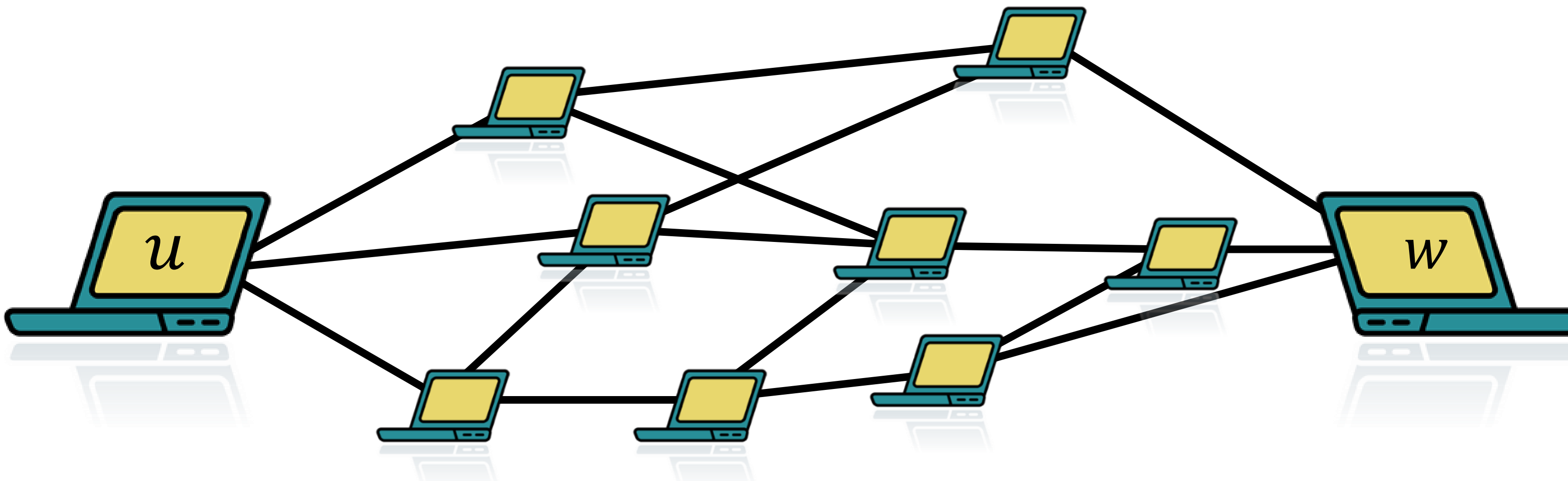
AN EFFICIENT ALGORITHM FOR LEARNING WITH SEMI-BANDIT FEEDBACK

Gergely Neu
INRIA Lille

Gábor Bartók
ETH Zürich

EXAMPLE: SEQUENTIAL ROUTING

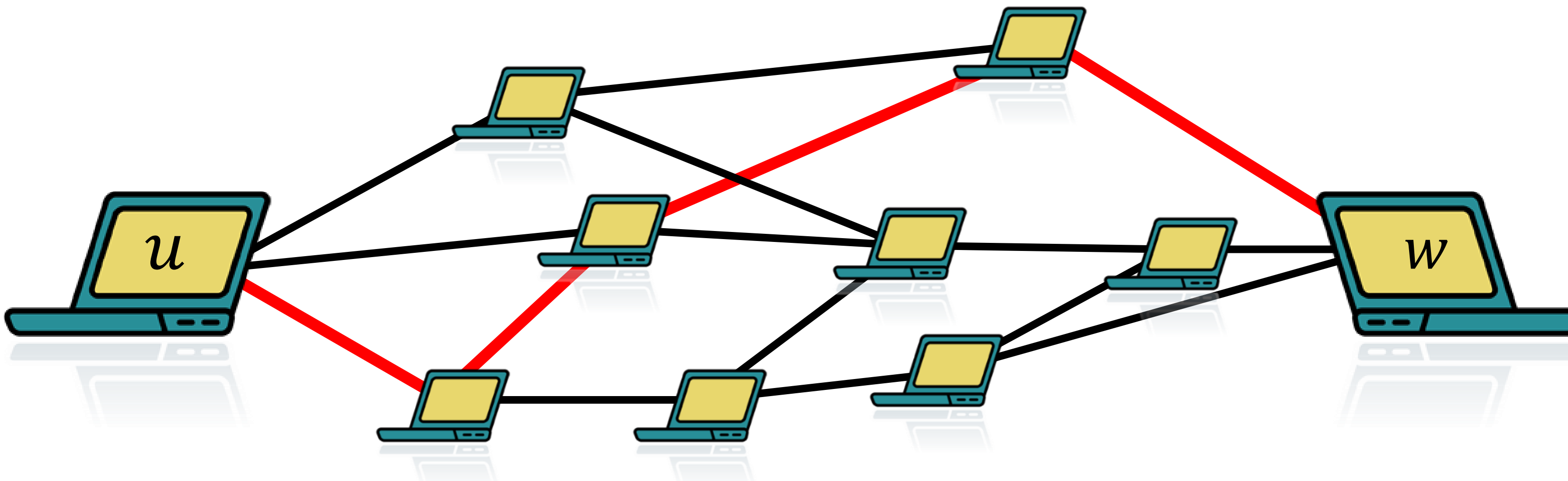


Decision set: set of all $u \to w$ paths

Delay on each edge can change arbitrarily over time

Goal: minimize total delay

EXAMPLE: SEQUENTIAL ROUTING



Decision set: set of all $u \to w$ paths

Delay on each edge can change arbitrarily over time

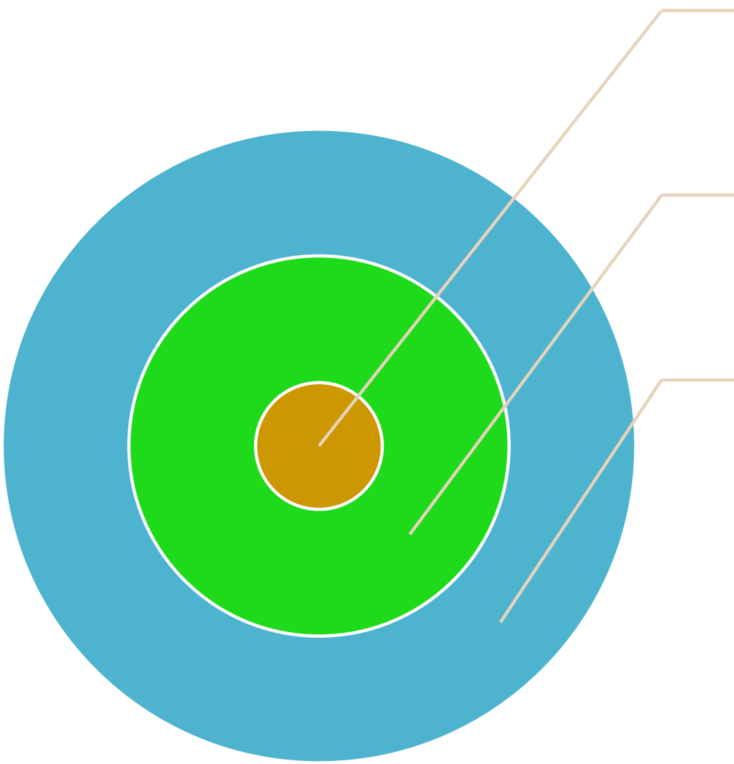
Goal: minimize total delay

ONLINE COMBINATORIAL OPTIMIZATION

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

FEEDBACK ASSUMPTIONS



Full bandit:
$V_t^\top \ell_t \in [0, m]$

Semi-bandit:
$\ell_{t,i}$ for all i s.t. $V_{t,i} = 1$

Full info:
$\ell_t \in [0,1]^d$

REGRET

Goal: minimize (expected) *regret*

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

FOLLOW THE PERTURBED LEADER (FPL)

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 $V_t = \arg\min_{v \in S} v^\top \left(L_{t-1,i} - Z_{t,i}\right)$
- Observe ℓ_t and let $L_t = L_{t-1} + \ell_t$

THE ADVANTAGE OF FPL

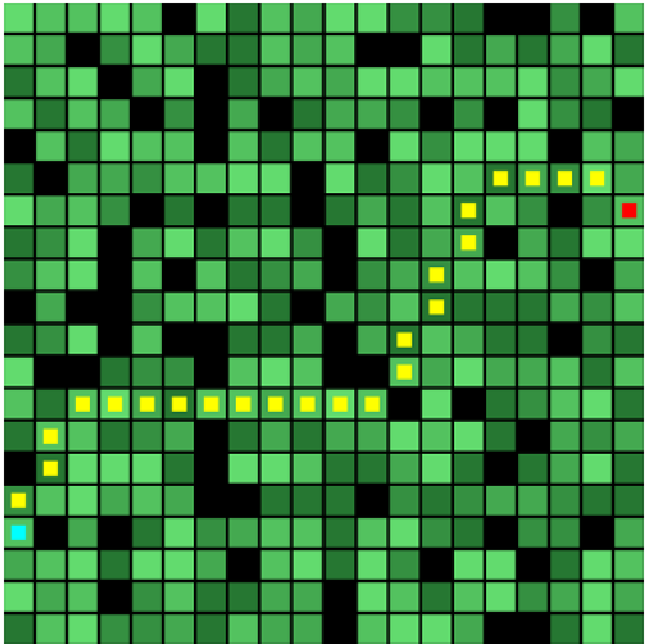
FPL is efficient whenever the optimization

$$\min_{v \in S} v^\top \ell$$

can be solved efficiently

Examples:

- Shortest paths



THE ADVANTAGE OF FPL

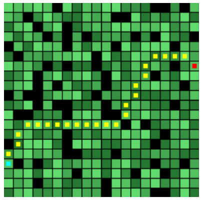
FPL is efficient whenever the optimization

$$\min_{v \in S} v^\top \ell$$

can be solved efficiently

Examples:

- Shortest paths
- Ranking



THE ADVANTAGE OF FPL

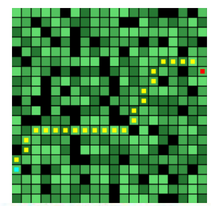
FPL is efficient whenever the optimization

$$\min_{v \in S} v^{\top} \ell$$

can be solved efficiently

Examples:

- Shortest paths
- Ranking
- Perfect matchings



THE ADVANTAGE OF FPL

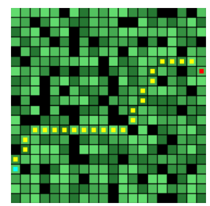
FPL is efficient whenever the optimization

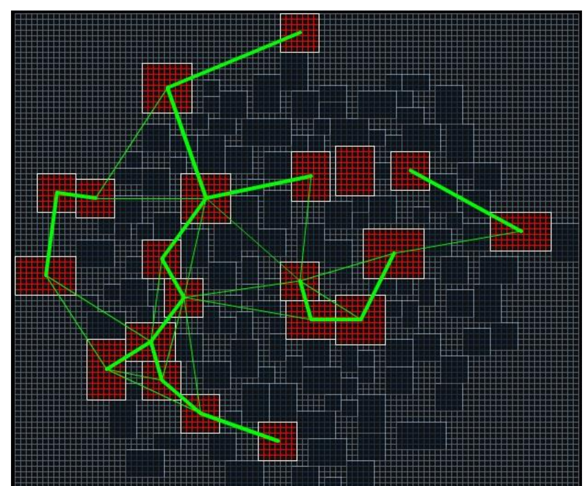
$$\min_{v \in S} v^\top \ell$$

can be solved efficiently

Examples:

- Shortest paths
- Ranking
- Perfect matchings
- Spanning trees
- etc.





“FPL only works for oblivious adversaries!”

“FPL is suboptimal by far!”

“FPL doesn’t work with bandit feedback!”

“PROBLEMS” WITH FPL

BEST KNOWN RESULTS

	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	$m\sqrt{dT\log d}$	???	???	always

BEST KNOWN RESULTS + OUR NEW RESULTS

	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

"FPL DOESN'T WORK WITH BANDIT FEEDBACK"

Q: how do we estimate unobserved losses?

A: use the estimates

$$\widehat{\ell}_{t,i} = \frac{\ell_{t,i}}{\mathbf{P}_t\left[V_{t,i} = 1\right]} V_{t,i}$$

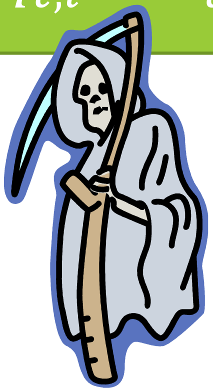
Unbiased since $\mathbf{E}_{\mathrm{t}}\left[V_{t,i}\right] = \mathbf{P}_t\left[V_{t,i} = 1\right]$...

... but how do we compute this?

“FPL DOESN’T WORK WITH BANDIT FEEDBACK”

$$V_t = \underset{v}{\operatorname{argmin}}\, v^\top\left(\hat{L}_{t-1} + Z_t\right)$$

$$q_{t,i} = \mathbf{P}_t[V_{t,i} = 1]$$

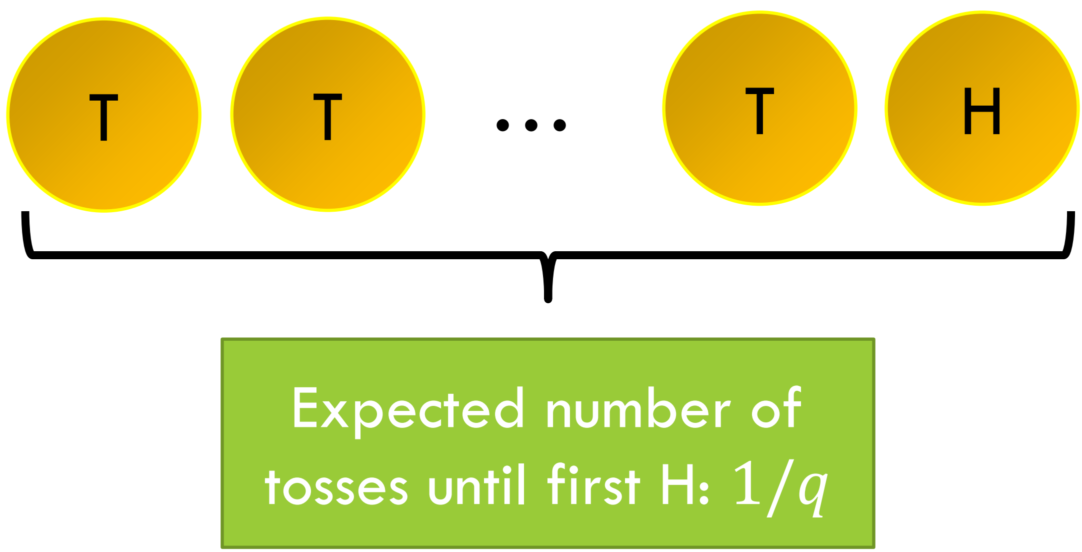


Poland (2005):
Estimate $q_{t,\cdot}$ by $O(T^2)$ samples

IDEA: GEOMETRIC RESAMPLING

Observe that we need to estimate $1/q_{t,i}$, not $q_{t,i}$!

Back to school: biased coin with $\mathbf{P}[\text{heads}] = q$



GEOMETRIC RESAMPLING FOR SEMI-BANDIT INFO

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

Unbiased since

- $\mathbf{E}_{\mathrm{t}}[V_{\boldsymbol{t},\boldsymbol{i}}] = q_{t,i}$
- $\mathbf{E}_{\mathrm{t}}[K_{t,i}] = 1/q_{t,i}$

REGRET GUARANTEES

Semi-bandit

Theorem:

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

Full info

Theorem:

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

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

FULL INFO PROOF SKETCH – STANDARD PART

For the analysis, introduce $\tilde{Z} \sim Z_1$

Introduce

$$\tilde{V}_t = \arg\min_{v \in S} v^\top (\hat{L}_t - \tilde{Z})$$

Notice that $\tilde{V}_t \sim V_{t+1}$ and the two are independent

Be-the-leader lemma: for any $v \in S$,

$$\mathbf{E}\left[\sum_{t=1}^{T} \left(\tilde{V}_t - v\right)^\top \ell_t\right] \leq \frac{m(\log d + 1)}{\eta}$$

FULL INFO PROOF SKETCH – NEW PART

Let $\tilde{p}_t(v) = \mathbf{P}[\tilde{V}_t = v]$

Show that

$$\tilde{p}_t(v) \geq \tilde{p}_{t-1}(v)(1 - \eta v^\top \ell_t),$$

and thus

$$\begin{aligned}\mathbf{E}[V_t^\top \ell_t] &\leq \mathbf{E}[\tilde{V}_t^\top \ell_t] + \eta \sum_{v \in S} \tilde{p}_{t-1}(v)(v^\top \ell_t)^2 \\ &\leq \mathbf{E}[\tilde{V}_t^\top \ell_t] + \eta m^2\end{aligned}$$

FULL INFO PROOF SKETCH – PUTTING IT TOGETHER

Eventually, we get

$$E\left[\sum_{t=1}^{T}(V_t - v)^{\top}\ell_t\right] \le \frac{m(\log d + 1)}{\eta} + \eta m^2 T$$

SEMI-BANDIT PROOF SKETCH

$$\frac{m(\log d + 1)}{\eta} \Rightarrow \frac{m(\log d + 1)}{\eta}$$

$$\eta m^2 \Rightarrow 2\eta m d$$

WHERE DOES THE SAMPLING HURT?

Had we known the $q_{t,i}$'s, we could do

$$2\eta m d \to \eta m d$$

How much samples do we need?

- Expectation: d ☺
- Worst-case: ∞ ☹

Stop sampling after M steps!

Additional regret: $\frac{dT}{eM}$

WHERE DOES THE SAMPLING HURT?

$M = \infty$

Theorem:

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

$M = O(\sqrt{dT}/m)$

Theorem:

$$R_T^{FPL+GR} \leq 3m\sqrt{2dT(\log d + 1)}$$

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: $\sqrt{d}T^{3/2}f(S)/m$

CONCLUSIONS & FUTURE WORK

Results

- Most efficient method for online learning with semi-bandit feedback
- Closed the gap between performance guarantees of expanded EXP3 and FPL

Open problems

- Full bandit feedback?
- Even stronger bounds for FPL?
- Is there an inherent computation/performance tradeoff in online learning?