

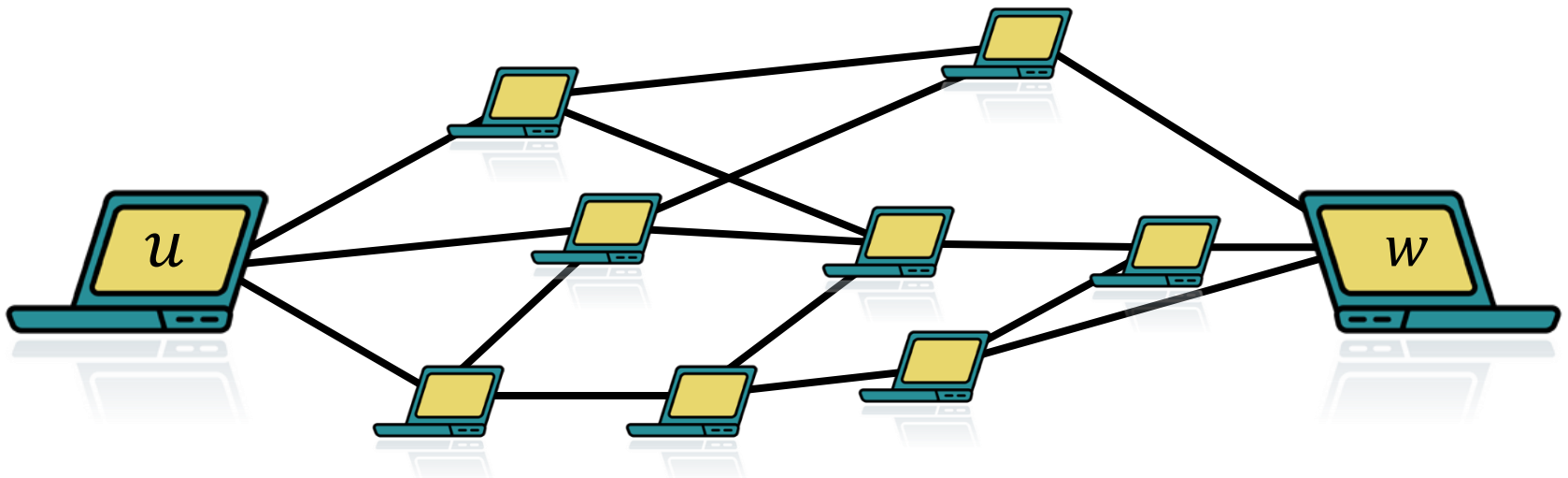
# First-order regret bounds for combinatorial semi-bandits

Gergely Neu

INRIA Lille, SequeL team

→ Universitat Pompeu Fabra, Barcelona

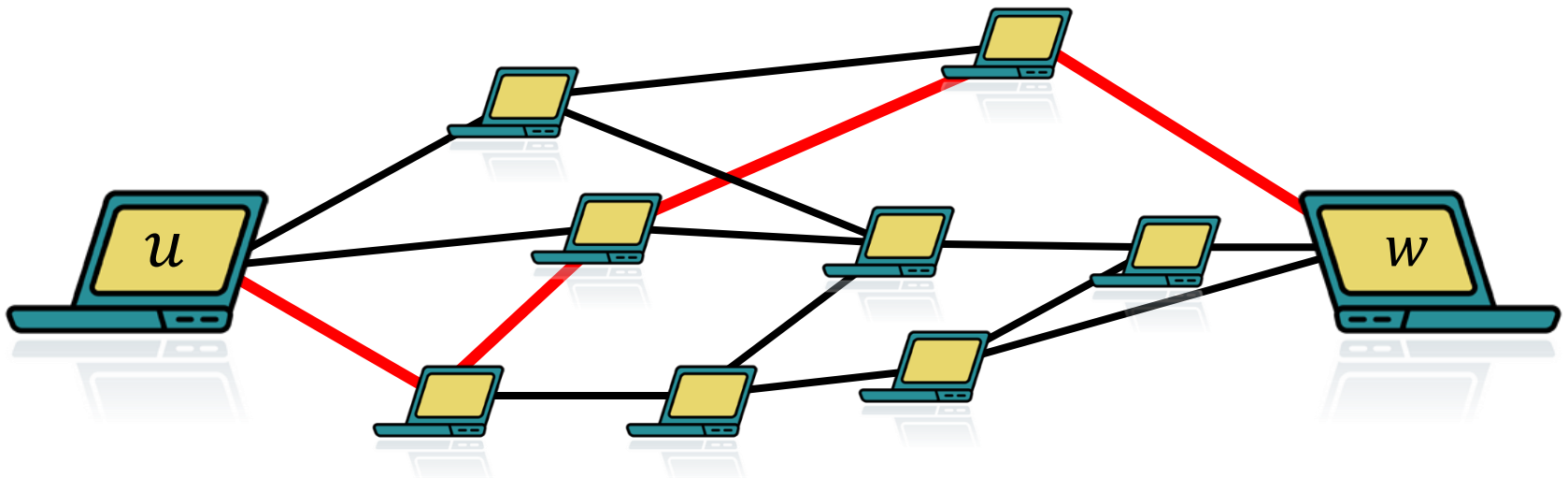
# Combinatorial semi-bandits



For every round  $t = 1, 2, \dots, T$

- learner picks an action  $V_t \in S \subseteq \{0, 1\}^d$
- Environment chooses loss vector  $\ell_t \in [0, 1]^d$
- Learner suffers loss  $V_t^\top \ell_t$
- Learner observes losses  $V_{t,i} \ell_{t,i}$

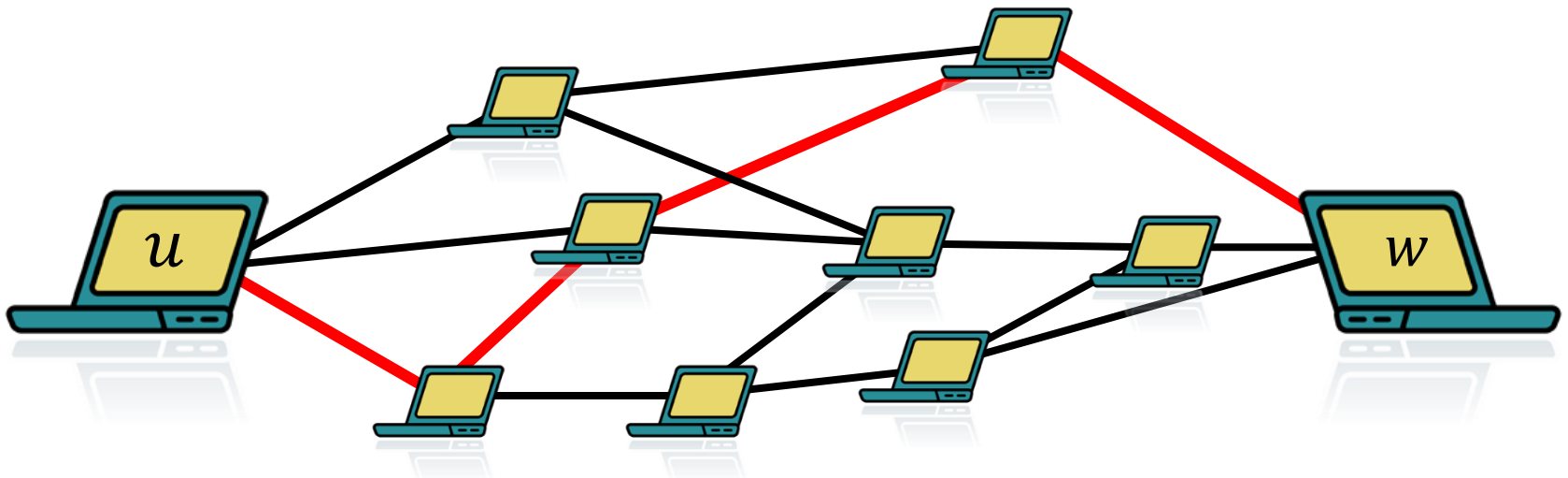
# Combinatorial semi-bandits



For every round  $t = 1, 2, \dots, T$

- learner picks an action  $V_t \in S \subseteq \{0, 1\}^d$
- Environment chooses loss vector  $\ell_t \in [0, 1]^d$
- Learner suffers loss  $V_t^\top \ell_t$
- Learner observes losses  $V_{t,i} \ell_{t,i}$

# Combinatorial semi-bandits



For every round  $t = 1, 2, \dots, T$

- learner picks an action  $V_t \in S \subseteq \{0, 1\}^d$
- Environment chooses loss vector  $\ell_t \in [0, 1]^d$
- Learner suffers loss  $V_t^\top \ell_t$
- Learner observes losses  $V_{t,i} \ell_{t,i}$

Decision set:

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

# Regret

- Goal: minimize **regret**

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

- Minimax regret is

$$\hat{R}_T = \Theta(\sqrt{mdT})$$

- Best efficient algorithm (FPL) gives

$$\hat{R}_T = O(m\sqrt{dT \log(d)})$$

# Regret

- Goal: minimize **regret**

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

- Minimax regret is

$$\hat{R}_T = \Theta(\sqrt{mdT})$$

- Best efficient algorithm (EPL) gives

Can we do better?

# First-order bounds

- A well-known improvement:



where  $L_T^* = \min_{v \in S} v^\top (\sum_{t=1}^T \ell_t)$

# First-order bounds

- A well-known improvement:



where  $L_T^* = \min_{v \in S} v^\top (\sum_{t=1}^T \ell_t)$

- Many examples for full information
- A handful of results for bandits:
  - › Stoltz (2005):  $d\sqrt{L_T^*}$
  - › Allenberg et al. (2006):  $\sqrt{dL_T^*}$
  - › Rakhlin and Sridharan (2013):  $d^{3/2}\sqrt{L_T^*}$



# First-order bounds

- A well-known improvement:



where  $L_T^* = \min_{v \in S} v^\top (\sum_{t=1}^T \ell_t)$

- Many examples for full information
- A handful of results for bandits:

› Stoltz (2005):  $d \sqrt{L_T^*}$

› None of these generalize efficiently  
› to combinatorial settings!

$\sqrt{L_T^*}$

# This paper

- Algorithm: FPL-TRIX

Follow the  
perturbed leader

Implicit exploration

Truncated exponential  
perturbations

# This paper

- Algorithm: FPL-TRIX

Follow the  
perturbed leader

Implicit exploration

Truncated exponential  
perturbations

These allow proving

$$\mathbf{E} \left[ \sum_{v \in S} p_t(v) (v^\top \hat{\ell}_t)^2 \right] \leq mdL_T^* + \tilde{O}(1/\eta)$$

instead of the usual

$$\mathbf{E} \left[ \sum_{v \in S} p_t(v) (v^\top \hat{\ell}_t)^2 \right] \leq mdT$$

# This paper

- Algorithm: FPL-TRIX

Follow the  
perturbed leader

Implicit exploration

Truncated exponential  
perturbations

These allow proving

$$\mathbf{E} \left[ \sum_{v \in S} p_t(v) (v^\top \hat{\ell}_t)^2 \right] \leq mdL_T^* + \tilde{O}(1/\eta)$$

Main  
result

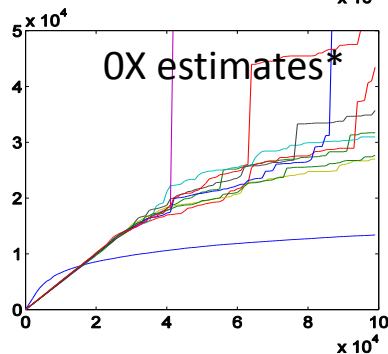
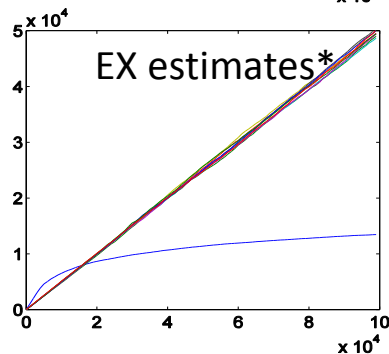
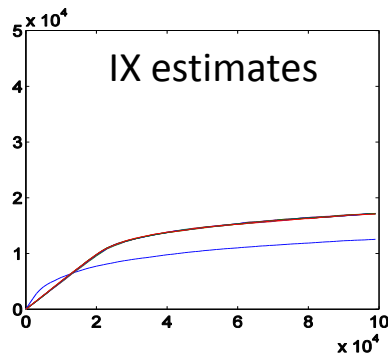
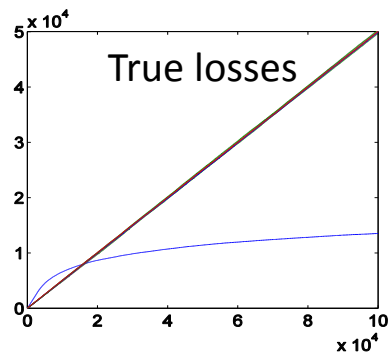
With the right tuning, FPL-TRIX guarantees

$$\hat{R}_T = O \left( m \sqrt{dL_T^* \log(d/m)} \right)$$

+ a better understanding of first-order bounds

# THANKS!

See you at the poster!



**Gergely Neu**  
INRIA Lille, Sequel  
= Universitat Pompeu Fabra, Barcelona

## First-order regret bounds for combinatorial semi-bands

**Combinatorial semi-bands**

- For each round  $t = 1, 2, \dots, T$ 
  - Environment chooses **decision set**  $S_t \subseteq S$
  - Learner chooses **action**  $V_t \in S_t \subseteq \{0,1\}^d$
  - Environment suffers **loss vector**  $\ell_t \in [0,1]^d$
  - Learner suffers **loss**  $V_t^\top \ell_t$
  - Learner observes **losses**  $V_{t'}^\top \ell_{t'}$
- Decision set:  $S = \{v \in \{0,1\}^d \mid |v|_1 \leq m\}$
- E.g.: sequential routing

• Goal: minimize **regret**  

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

• Minimax regret is  $\hat{R}_T = \Theta(\sqrt{m d T})$

• Best efficient algorithm (FPL) gives  $\hat{R}_T = O(m \sqrt{d T} \log(d))$

**Can we do better?**

**First-order bounds**

A well-known improvement:  
 $\sqrt{T} \rightarrow \sqrt{L_T}$

where  $L_T = \min_{v \in S} v^\top (\sum_{t=1}^T \ell_t)$

- Many examples for full feedback
- Stoltz (2005):  $d \sqrt{L_T}$
- Allenberg et al. (2006):  $\sqrt{d L_T}$
- Rakhlin and Sridharan (2013):  $d \sqrt{d L_T}$

None of these generalize efficiently to combinatorial settings!

**Algorithm: FPL-TRIX**

**Parameters:** non-decreasing sequences  $(\eta_t), (\gamma_t), (\beta_t)$

**Initialization:**  $L_0 = 0$

**For each round**  $t = 1, 2, \dots, T$

- Draw perturbation vector  $Z_t$  with  $Z_{t,i} \sim f(\cdot) \log(1/\beta_t)$
- Play action  $V_t = \arg \max_{v \in S} (v^\top (\eta_{t-1} - Z_t))$
- Compute  $\hat{\ell}_{t,i} = \frac{\ell_{t,i} V_{t,i}}{\eta_t |V_t|_1 + \gamma_t}$
- Let  $L_t = L_{t-1} + \hat{\ell}_t$

**Trick #1**

- Truncated perturbations  $Z_t$ 
  - $f(x) = e^{-x} \mathbf{1}_{\{x \in [0, \log(1/\beta_t)]\}}$
  - Suppresses suboptimal actions a.s.
- Follow the perturbed leader (FPL)

**Trick #2**

- Implicit exploration (IX)
- Provides "optimistic" bias
- Ensures that  $\hat{\ell}_{t,i}$  is bounded

**The key idea**

- A typical regret bound (Exp3, FPL, ...):  

$$\frac{C_1}{\eta} + \eta \cdot C_2 \sum_{i=1}^d \hat{\ell}_{t,i}$$
- where  $\eta > 0$  is a **learning rate**
- If  $\mathbb{E}[\hat{\ell}_{t,i}] = \ell_{t,i}$ , then this becomes  

$$\frac{C_1}{\eta} + \eta \cdot C_2 \cdot d \max_i L_{T,i}$$
- giving  $\tilde{O}(\sqrt{d \max_i L_{T,i}}) = \tilde{O}(\sqrt{d T})$

**Idea:** Introduce a bias in  $\hat{\ell}_{t,i}$  that ensures for all  $i$   

$$L_{T,i} \leq \min_{v \in S} v^\top L_T + \tilde{O}\left(\frac{1}{\eta}\right)$$

This allows proving  

$$\frac{C_1}{\eta} + \eta \cdot C_2 \cdot d L_T \rightarrow \tilde{O}(\sqrt{d L_T})$$
if  $\mathbb{E}[\min_{v \in S} v^\top L_T] \leq L_T$  also holds

**Main result**

With the right tuning, FPL-TRIX guarantees  

$$\hat{R}_T = O\left(m \sqrt{d L_T^*} \log(d/m)\right)$$
...and also  $\hat{R}_T = O(m \sqrt{d T} \log(d/m))$

**Proof steps**

Let  $D = \log(d/m)$ ,  $\beta_t = \log(1/\beta_t)$

- A key result about the bias of  $\hat{\ell}_{t,i}$ :  
**Lemma 2:** For any  $i$  and  $v$ ,  

$$L_{T,i} \leq v^\top L_T + \frac{m(D + \beta_T)}{\eta_T} + \frac{1}{\gamma_T}$$
- The regret of FPL-TRIX:  
**Theorem 3:** If  $\beta_t d \leq \gamma_t$ , then  

$$\sum_{t=1}^T \sum_{i=1}^d \hat{\ell}_{t,i} \leq v^\top L_T + \frac{mD}{\eta_T} + \sum_{t=1}^T (\eta_t m + \beta_t d + \gamma_t) \sum_{i=1}^d \hat{\ell}_{t,i}$$
- This suggests  $\gamma_t = \eta_t m = \beta_t d$
- Static learning rates:  
**Corollary 4:**  
Setting  $\eta = \sqrt{3(D+1)/d L_T^*}$  gives  

$$\hat{R}_T \leq 5.2 m \sqrt{d L_T^* (D+1)} + O(\log T)$$
- Self-confident learning rates:  
**Theorem 5:**  
Setting  $S_t = \frac{1}{D} + \sum_{k=1}^t \sum_{i=1}^d \hat{\ell}_{k,i}$   
and  $\eta_t = \sqrt{D/S_{t-1}}$  gives  

$$\hat{R}_T \leq 1.3 m \sqrt{d L_T^* (D+1)} + O(\log T)$$
- Proof: quite tricky as  $S_t \neq \Theta(\dots)$   
...but it's much more practical than using a doubling trick

**Why does it work?**

- Truncation actually not necessary
- Implicit exploration is necessary

**The IX effect**

\*without estimator with and without implicit exploration