Gergely Neu
INRIA Lille, SequeL
→ Universitat Pompeu Fabra, Barcelona

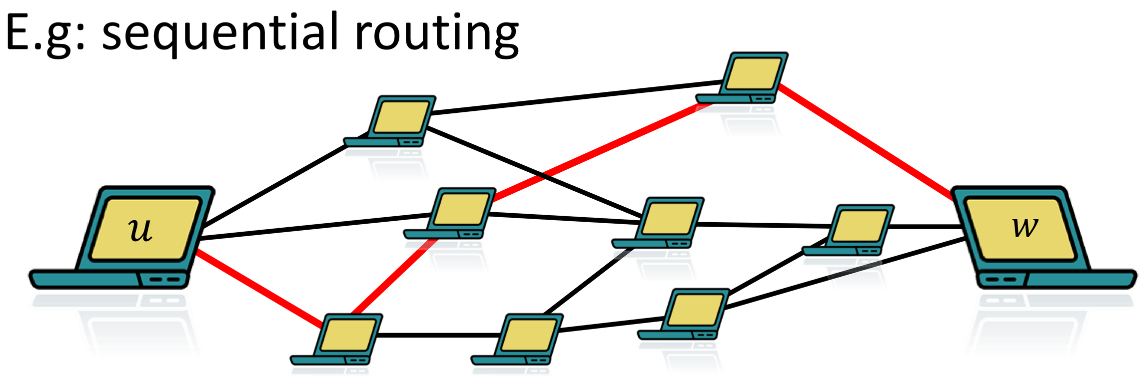
First-order regret bounds for combinatorial semi-bandits

Combinatorial semi-bandits

For each round $t = 1,2,\ldots,T$
- Environment chooses decision set $S_t \in S$
- Learner chooses 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:
$S = \{v_i\}_{i=1}^N \subseteq \{0,1\}^d$
$\|v_i\|_1 \le m$

E.g: sequential routing



- 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
$$\hat{R}_T = \Theta\left(\sqrt{mdT}\right)$$
- Best efficient algorithm (FPL) gives
$$\hat{R}_T = O\left(m\sqrt{dT\log(d)}\right)$$

Can we do better?

First-order bounds

A well-known improvement:

$$\sqrt{T} \rightarrow \sqrt{L_T^*}$$

where $L_T^* = \min_{v \in S} v^\top\left(\sum_{t=1}^T \ell_t\right)$
- Many examples for full feedback
- A handful of results for bandits:
  - Stoltz (2005): $d\sqrt{L_T^*}$
  - Allenberg et al. (2006): $\sqrt{dL_T^*}$
  - Rakhlin and Sridharan (2013): $d\sqrt{dL_T^*}$

None of these generalize efficiently to combinatorial settings!

The key idea

- A typical regret bound (Exp3, FPL,…):
$$\frac{C_1}{\eta} + \eta \cdot C_2 \sum_{t=1}^{T}\sum_{i=1}^{d} \hat{\ell}_{t,i},$$
where $\eta > 0$ is a learning rate
- If $\mathbf{E}[\hat{\ell}_{t,i}] = \ell_{t,i}$, then this becomes
$$\frac{C_1}{\eta} + \eta \cdot C_2 \cdot d \max_i L_{T,i},$$
giving $\tilde{O}\left(\sqrt{d \max_i L_{T,i}}\right) = \tilde{O}\left(\sqrt{dT}\right)$

Idea: introduce a bias in $\hat{\ell}_{t,i}$ that ensures for all i
$$\hat{L}_{T,i} \le \min_{v \in S} v^\top \hat{L}_T + \tilde{O}\left(\frac{1}{\eta}\right)$$

- This allows proving
$$\frac{C_1}{\eta} + \eta \cdot C_2 \cdot dL_T^* \rightarrow \tilde{O}\left(\sqrt{dL_T^*}\right)$$
if $\mathbf{E}\left[\min_v v^\top \hat{L}_T\right] \le L_T^*$ also holds

Algorithm: FPL-TrIX

Parameters:
non-decreasing sequences (η_t), (γ_t), (β_t)
Initialization: $\hat{L}_0 = \mathbf{0}$
For each round $t = 1,2,\ldots,T$
- Draw perturbation vector Z_t with
$$Z_{t,i} \sim f(\cdot\,|\log(1/\beta_t))$$
- Play action
$$V_t = \min_{v \in S} v^\top\left(\eta_t \hat{L}_{t-1} - Z_t\right)$$
- Compute
$$\hat{\ell}_{t,i} = \frac{\ell_{t,i} V_{t,i}}{\mathbf{E}_t[V_{t,i}] + \gamma_t}$$
- Let $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$

Trick #1

Truncated perturbations (Tr)
- $f(z|B) \propto e^{-z}\mathbf{1}_{\{z \in [0,B]\}}$
- 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

Main result

With the right tuning, FPL-TrIX guarantees
$$\hat{R}_T = O\left(m\sqrt{dL_T^*\log(d/m)}\right)$$
…and also $\hat{R}_T = O\left(m\sqrt{dT\log(d/m)}\right)$

Proof steps

Let $D = \log(d/m)$, $B_t = \log(1/\beta_t)$
- A key result about the bias of $\hat{\ell}_{t,i}$:

Lemma 2: For any i and v,
$$\hat{L}_{T,i} \le v^\top \hat{L}_T + \frac{m(D + B_T)}{\eta_T} + \frac{1}{\gamma_T}$$

- The regret of FPL-TrIX:

Theorem 3: If $\beta_t d \le \gamma_t$, then
$$\sum_{t=1}^{T} V_t^\top \ell_t \le v^\top \hat{L}_T + \frac{mD}{\eta_T} + \sum_{t=1}^{T}\left(\eta_t m + \beta_t d + \gamma_t\right)\sum_{i=1}^{d}\hat{\ell}_{t,i}$$

- This suggests $\gamma_t = \eta_t m = \beta_t d$
- Static learning rates:

Corollary 4:
Setting $\eta = \sqrt{3(D+1)/dL_T^*}$ gives
$$\hat{R}_T \le 5.2\, m\sqrt{dL_T^*(D+1)} + O(\log T)$$

- Self-confident learning rates:

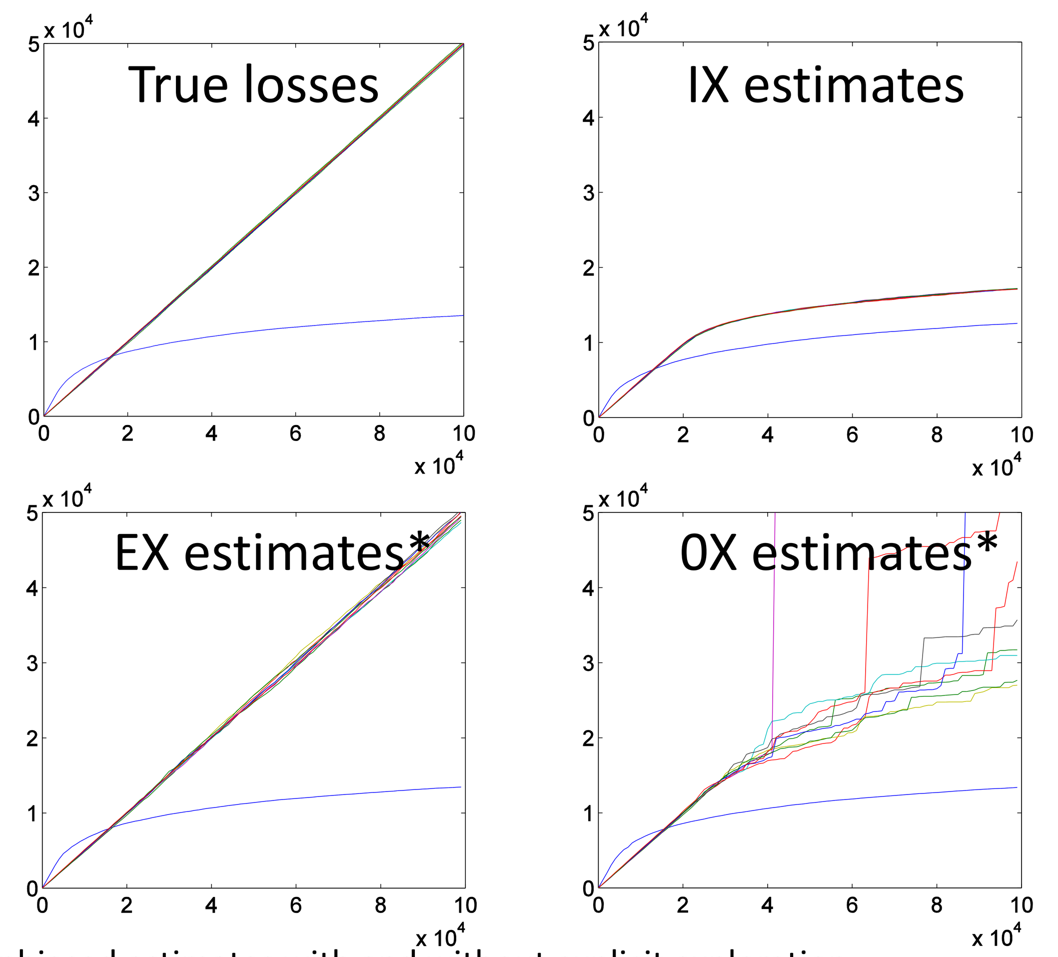
Theorem 5:
Setting $S_t = \frac{1}{D} + \sum_{k=1}^{t}\sum_{i=1}^{d}\hat{\ell}_{k,i}$ and $\eta_t = \sqrt{D/S_{t-1}}$ gives
$$\hat{R}_T \le 13\, m\sqrt{dL_T^*(D+1)} + O(\log T)$$

- Proof: quite tricky as $S_t \ne O(t)$…
…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



*unbiased estimates with and without explicit exploration