First-order regret bounds for combinatorial semi-bandits

Gergely Neu
INRIA Lille, SequeL team
→ Universitat Pompeu Fabra, Barcelona
Combinatorial semi-bandits

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

- 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 $V_t^T \ell_t$
- Learner observes losses $V_{t,i} \ell_{t,i}$
Combinatorial semi-bandits

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

- 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 $V_t^T \ell_t$
- Learner observes losses $V_{t,i} \ell_{t,i}$
Combinatorial semi-bandits

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

- 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 $V_t^T \ell_t$
- 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

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

- Minimax regret is

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

- Best efficient algorithm (FPL) gives

\[ \hat{R}_T = O\left(m \sqrt{dT \log(d)}\right) \]
Regret

- Goal: minimize regret

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

- Minimax regret is

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

- Best efficient algorithm (FPL) gives

Can we do better?
First-order bounds

• A well-known improvement:

\[ \sqrt{T} \rightarrow \sqrt{L_T^*} \]

where \( L_T^* = \min_{\nu \in S} \nu^\top (\sum_{t=1}^{T} \ell_t) \)
First-order bounds

• A well-known improvement:

\[
\sqrt{T} \rightarrow \sqrt{L_T^*}
\]

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

• 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:
  \[
  \sqrt{T} \rightarrow \sqrt{L^*_T}
  \]
  where \( L^*_T = \min_{\nu \in S} \nu^T (\sum_{t=1}^T \ell_t) \)

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

None of these generalize efficiently to combinatorial settings!
This paper

• Algorithm: FPL-TRIX

  - Follow the perturbed leader
  - Truncated exponential perturbations
  - Implicit exploration
This paper

- Algorithm: FPL-TRIX

- Follow the perturbed leader

- Truncated exponential perturbations

- Implicit exploration

These allow proving

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

instead of the usual

\[ \mathbb{E} \left[ \sum_{v \in S} p_t(v) (v^\top \hat{\ell}_t)^2 \right] \leq m d T \]
This paper

• Algorithm: FPL-TRIX

Follow the perturbed leader

Implicit exploration

Truncated exponential perturbations

These allow proving

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

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!

True losses

IX estimates

EX estimates*

OX estimates*