Nested Bandits

International Conference on Machine Learning

Matthieu Martin, Panayotis Mertikopoulos, Thibaud Rahier, **Houssam Zenati**

July, 2022

Criteo AI Lab
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

CRITEO AI Lab

ICML International Conference On Machine Learning

The Blue Bus / Red Bus Paradox (1/2)

Choosing a mean of transportation

Alternatives:

- a **car**, which takes on average 15 mins ($v_{car} = -15$)
- a **bus**, which takes on average 20 mins ($v_{bus} = -20$)

Logit choice [1, 2]

- $\mathbb{P}(\text{car}) = \frac{\exp(v_{car})}{\exp(v_{car})+\exp(v_{bus})} \approx 0.62$ most probable choice
- $\mathbb{P}(\text{bus}) = \frac{\exp(v_{bus})}{\exp(v_{car})+\exp(v_{bus})} \approx 0.38$

Choosing a mean of transportation

Alternatives:

- a **car**, which takes on average 15 mins ($v_{\text{car}} = -15$)
- a **blue bus**, which takes on average 20 mins ($v_{\text{bus}} = -20$)
- a **red bus**, identical to the blue bus (except its color)

Logit choice [1, 2]

- $\mathbb{P}(\text{car}) = \frac{\exp(v_{\text{car}})}{\exp(v_{\text{car}})+2\exp(v_{\text{bus}})} = 0.45$ no longer most probable!
- $\mathbb{P}(\text{blue bus}) = \mathbb{P}(\text{red bus}) = \frac{\exp(v_{\text{bus}})}{\exp(v_{\text{car}})+2\exp(v_{\text{bus}})} = 0.27$

Problem

Logit choice no longer reasonable: an irrelevant alternative switches choice odds!

Adversarial Bandits (1/2)

Notations and incurred regret of EXP3

- $(v_{a,t})_{a\in\mathcal{A}}$ *payoff vector* of stage $t = 1, 2, \ldots T$
- $P_t(a)$ probability of choosing arm a at stage t (n arms)
- $r_t = v_{a_t,t}$ *reward* received at stage t from arm $a_t \sim P_t$

$$\mathrm{Reg}(T) \leq \sqrt{2n\log(n)T}$$

Blue Bus / Red Bus situation

Two alternatives $a_1, a_2 \in \mathcal{A}$ generate consistently same reward:

can we avoid considering both alternatives in a bandit algorithm?

More general: $\mathcal{A}$ has an inherent structure?

If n very big but some alternatives have very similar rewards:

can we exploit this side information to design a more efficient algorithm?

Notations and incurred regret of EXP3

- $(v_{a,t})_{a\in\mathcal{A}}$ *payoff vector* of stage $t = 1, 2, \dots T$
- $P_t(a)$ probability of choosing arm a at stage t (n arms)
- $r_t = v_{a_t,t}$ *reward* received at stage t from arm $a_t \sim P_t$

$$\mathrm{Reg}(T) \leq \sqrt{2n\log(n)T}$$

Nested Exponential Weights algorithm

If we exploit side-information on the structure of $\mathcal{A}$ and regularity of $(v_a)_{a\in\mathcal{A}}$, we propose to use the **Nested Exponential Weights** (NEW) algorithm to obtain

$$\mathrm{Reg}(T) \leq \sqrt{2n_{\mathrm{eff}}\log(n)T}$$

where n_{eff} is typically much smaller than n and we always have $n_{\mathrm{eff}} \leq n$.

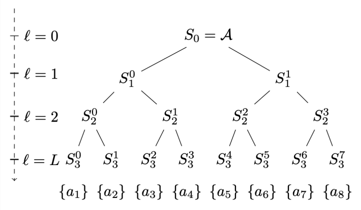

Figure 1: Nested structure: ($L = 3$)

- $\mathcal{A} := \{a_i : i = 1, \ldots, n\}$ set of *alternatives*
- $\{\mathcal{A}\} =: \mathcal{S}_0 \succcurlyeq \cdots \succcurlyeq \mathcal{S}_L := \{\{a\} : a \in \mathcal{A}\}$ tower of *partitions*

Reward & Feedback

For all $a \in \mathcal{A}$ and $a \equiv S_L \lhd S_{L-1} \lhd \cdots \lhd S_0 \equiv \mathcal{A}$ its *lineage*,

$$v_a = \sum_{\ell=1}^{L} r_{S_\ell}$$

Semi-bandit feedback: at each round, the learner *observes each* r_{S_ℓ}

$$r_{S_\ell} \in [0, R_\ell] \quad \text{for all } S_\ell \in \mathcal{S}_\ell,\ \ell = 1, \ldots, L,$$

where $R_\ell \geq 0$ represents the *reward variability* for $\mathcal{S}_\ell$

Algorithm

For each stage $t = 1, 2, \ldots,$ given $y_t \in \mathbb{R}^{\mathcal{A}}$ (current score), η_t (learning rate) and μ_ℓ (uncertainty level parameter), the learner:

1. computes choice probability P_t from **Nested Logit Choice** (NLC) $P_{S_\ell|S_{\ell-1}}(y)$ and y_t using **upward pass** on level scores y_{S_ℓ}

$$P_{S_\ell|S_{\ell-1}}(y) = \frac{\exp(y_{S_\ell}/\mu_\ell)}{\exp(y_{S_{\ell-1}}/\mu_\ell)} \tag{NLC}$$

2. selects action $a_t \in \mathcal{A}$ following **downward pass** in (NLC)

$$a_t \sim P_t(\eta_t y_t)$$

3. uses level rewards $r_{S,t}$ for each class $S \ni a_t$ and constructs a **Nested Importance Weighted Estimator** (NIWE) $\hat{v}_t$ of the payoff vector of stage t
4. updates their score: $y_{t+1} \leftarrow y_t + \hat{v}_t$ and the process repeats

Regret guarantees for NEW

Theorem

Defining $\sqrt{n_{\mathrm{eff}}} = \sum_{\ell=1}^{L} \sqrt{n_\ell} R_\ell$*, if NEW is run with* $\eta_t = \sqrt{\log n/(2t)}$*, we have*

$$\mathbb{E}[\mathrm{Reg}_p(T)] \leq 2\sqrt{2 n_{\mathrm{eff}} \log n \cdot T}.$$

Comparison to EXP3

Regret guarantees of NEW and EXP3 differ by a factor of

$$\alpha = \sqrt{n/n_{\mathrm{eff}}},$$

Suppose red bus / blue bus problem with

- $n_1 = 2$ classes and $n_2 = 100$ alternatives per class
- negligible intra-class reward differential ($R_2 \approx 0$)

regret guarantees improves by a factor of $\alpha \approx 10$

Benefits in the red bus / blue bus problem

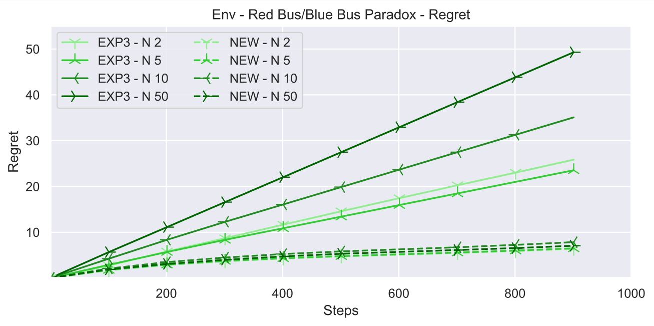

Figure 2: Regret of EXP3 and NEW in the red bus / blue bus problem with different numbers of buses *N*.

Interpretation

NEW systematically achieves better regret than EXP3 and is far less sensible to *N*

Conclusion

The **Nested Exponential Weights** (NEW) algorithm combines:

- the **Nested Logit Choice** (NLC) rule
- the **Nested Importance Weighted Estimator** (NIWE)

resulting in an improved adversarial bandit algorithm exploiting side-information on the **structure** of $\mathcal{A}$ and **regularity of** $(v_a)_{a \in \mathcal{A}}$

Thank you!

References

[1] R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

[2] D. L. McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, NY, 1974.