Joint Online Learning and Decision-making via Dual Mirror Descent

Alfonso Lobos, Paul Grigas, Zheng Wen

International Conference on Machine Learning (ICML)

July 21, 2021

Motivating Toy Example

Decision Problem

Decide energy production for the incoming week. Production decisions are taken every 15 minutes.

Global constraints: Preferable no less than m megawatts (MWs) should be produced and no more than M MWs.

Motivating Toy Example: Notation

- T: Total 15 minutes periods.
- $z^t \in \mathcal{Z}$: Energy decision at period $t \in T$. (In general $\mathcal{Z} \subset \mathbb{R}^d$).
- $w^t \in W$: Context vector observed at period $t \in T$.
- $f(\cdot;\cdot) : \mathcal{X} \times \mathcal{W} \to \mathbb{R}, c(\cdot;\cdot) : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$: Revenue and cost functions when all problem parameters are known.
- Global cost constraints:

$$m \leq \sum_{t=1}^{T} c(z^t; w^t) \leq M$$

Motivating Toy Example: Notation (Cont.)

- $\theta^* \in \Theta$: Possibly unknown problem parameters that we may aim to learn.
- $f(\cdot;\cdot,\cdot) : \mathcal{Z} \times \Theta \times \mathcal{W} \to \mathbb{R}, c(\cdot;\cdot,\cdot) : \mathcal{Z} \times \Theta \times \mathcal{W} \to \mathbb{R}$: Revenue and cost functions when learning θ^* is required.

Motivating Toy Example: Overall Goal

Derive an energy production policy for every 15 minutes that aims to maximize the total revenue while taking into account lower and upper bound cost global constraints. The setup may need to learn the vector θ^* of unknown parameters if necessary.

Contribution 1: Algorithmic Setup

- We propose a novel family of algorithms to tackle a joint online learning and decision making problem.
- Algorithm can be seen as a combination between a dual mirror descent scheme and generic learning steps.
- Allows arbitrary variable space and general revenue and cost functions that can even take negative values.
- Allows both lower and upper bound global cost constraints.

Contribution 2: Benchmark

- Novel benchmark used to measure the regret of our algorithm.
- Benchmark generalizes the 'best dynamic solution in hindsight' benchmark.
- Well-suited for settings with possible "infeasible sequence of context vector arrivals".

Contribution 3: Regret Bounds

- When θ^* is known, our algorithm achieves a $O(\sqrt{T})$ regret.
- Result relies on bounding the dual variables using a Slater type of condition.
- In the general case, regret is bounded by the sum between $O(\sqrt{T})$ 'terms' coming from the 'known θ^*' case plus learning terms.

Contribution 4: Worst-constraint Violation

- Algorithm may violate a lower bound constraint by at most by $O(\sqrt{T})$ when θ^* is know.
- By construction, algorithm satisfies the upper bound constraints.

Contribution 5: Experiments

- Two experiments performed. First is related to bidding strategies in online advertising. Second, is a linear contextual bandits problem with bounded number of actions.
- Experiments show the robustness and flexibility of our approach.

Thank you for Listening.