Online Learning Applications

Part 8: Learning in non-truthful auctions with budget constraints

Bidding in non-truthful auctions with budget

How to bid in repeated non-truthful auctions?

- Even if there is no budget it is a non-trivial online learning problem
- The problem “generalizes” MABs
- The problem has a continuous set of arms (bids)

We make the following assumption to avoid dealing with continuous sets of arms:

Assumption

There is a finite number of possible bids $\mathcal{B}$.

Bidding in non-truthful auctions with budget

How to bid in repeated non-truthful auctions?

- Even if there is no budget it is a non-trivial online learning problem
- The problem "generalizes" MABs
- The problem has a continuous set of arms (bids)

We make the following assumption to avoid dealing with continuous sets of arms:

Assumption

There is a finite number of possible bids $\mathcal{B}$.

⚠ In the previous lectures, we have seen some techniques to handle online problems with continuous action space (in a different setting).

Why does discretization work?

Discretizing the set of bids we do not loose too much utility:

- Discretize the bids into $\mathcal{B} = \{0, \epsilon, 2\epsilon, \dots, 1\}$
- Given the optimal bid b, there is a bid in $b' \in \mathcal{B}$ at most ϵ larger such that:
  - ▷ b' wins whenever b win
  - ▷ With b' we pay at most ϵ more than with b

The reward function is **one-sided Lipschitz**, i.e., the utility is Lipschitz continuous only in one direction.

Formal setting

- Sequence of T **non-truthful** auctions

Formal setting

- Sequence of T **non-truthful** auctions (for the ease of exposition, we will focus on first-price auctions)

Formal setting

- Sequence of T **non-truthful** auctions (for the ease of exposition, we will focus on first-price auctions)
- The bidder has a valuation $v \in [0, 1]$ (i.e, the utility when the ad is displayed)
- The bidder has an initial budget B

At each round $t \in [T]$:

1. The bidder chooses $b_t \in [0, 1]$
2. m_t is the maximum among the competing bids
3. The bidder utility is $f_t(b_t) = (v - b_t)\mathbf{1}[b_t \geq m_t]$
4. The bidder incurs a cost $c_t(b_t) = b_t\mathbf{1}[b_t \geq m_t]$
5. The budget is decreased by $c_t(b_t)$
6. If the budget is smaller than 1 the bidder interaction stops (this avoids spending more than the budget)

Formal setting

- We consider two possible sequences of m_t
  - ▷ Stochastic: m_t are sampled from a distribution D
  - ▷ Adversarial: no assumption on m_t

Formal setting

- We consider two possible sequences of m_t
  - Stochastic: m_t are sampled from a distribution D
  - Adversarial: no assumption on m_t
- Some notation for stochastic environments:
  - $\rho = B/T$ budget per round
  - $(f, c) \sim \mathcal{D}$ is the distribution over utility and costs induced by $m_t \sim D$
  - $\bar{f}(b) = \mathbb{E}_{(f,c)\sim\mathcal{D}} f(b)$
  - $\bar{c}(b) = \mathbb{E}_{(f,c)\sim\mathcal{D}} f(b)$
  - $\gamma \in \Delta_{\mathcal{B}}$ is a distribution over bids
  - $f(\gamma) = \mathbb{E}_{b\sim\gamma} f(b)$
  - $c(\gamma) = \mathbb{E}_{b\sim\gamma} c(b)$

Baseline (stochastic environment)

We want to have no-regret with respect to:

Baseline

The reward of the best dynamic policy when the decision maker knows the underlying distribution (but not the realizations).

- This baseline is related to the baseline in MABs in which we consider the regret with respect to the best arm **in expectation**

The baseline is upperbounded by $T \cdot OPT$ [Badanidiyuru et al., 2018], where

$$OPT = \begin{cases} \sup_{\gamma \in \Delta_{\mathcal{B}}} \bar{f}(\gamma) \\ \text{s.t. } \bar{c}(\gamma) \leq \rho \end{cases}$$

- OPT is the per-round expected utility of the best policy that satisfies the budget constraint in expectation

Generalizing multiplicative pacing

Multiplicative pacing works well in truthful auctions.

Can we generalize multiplicative pacing to non-truthful auctions?

Generalizing multiplicative pacing

Multiplicative pacing works well in truthful auctions.

Can we generalize multiplicative pacing to non-truthful auctions?

Idea: we can Lagrangify the constraint obtaining the Lagragian function

$$\bar{L}(\gamma, \lambda) = \bar{f}(\gamma) - \lambda \left[\bar{c}(\gamma) - \rho\right],$$

where

- $\gamma \in \Delta_{\mathcal{B}}$ is a randomized bidding strategy
- $\lambda \in \mathbb{R}_+$ is a Lagrange multiplier that specifies "how important is to satisfy the budget constraint"

Similarly, given two functions f_t and c_t, we let:

$$L(\gamma, \lambda, f_t, c_t) = f_t(\gamma) - \lambda \left[c_t(\gamma) - \rho\right].$$

Lagrangian game

Given the Lagrangian function $L(\cdot, \cdot, f_t, c_t)$:

- The bidder chooses γ and wants to maximize $L(\gamma, \lambda, f_t, c_t)$
- An adversary chooses λ and wants to minimize $L(\gamma, \lambda, f_t, c_t)$

Lagrangian game

Given the Lagrangian function $L(\cdot, \cdot, f_t, c_t)$:

- The bidder chooses γ and wants to maximize $L(\gamma, \lambda, f_t, c_t)$
- An adversary chooses λ and wants to minimize $L(\gamma, \lambda, f_t, c_t)$

In truthful auctions:

- λ is the pacing multiplier (updated with online gradient descent)
- It is possible to prove that $b = \frac{v}{1+\lambda} \in \arg\max_{\gamma \in \Delta_{\mathcal{B}}} L(\gamma, \lambda, f_t, c_t)$, i.e., it is an optimal bid:
  - ▷ The bidder wants to win the auction if and only if $(v - m_t) - \lambda m_t \geq 0$
  - ▷ Equivalently, $m_t \leq \frac{v}{\lambda+1}$
  - ▷ Bidding $\frac{v}{\lambda+1}$ we can guarantee to win all and only the auctions with $m_t \leq \frac{v}{\lambda+1}$
- We recover multiplicative pacing

Lagrangian game

Given the Lagrangian function $L(\cdot, \cdot, f_t, c_t)$:

- The bidder chooses γ and wants to maximize $L(\gamma, \lambda, f_t, c_t)$
- An adversary chooses λ and wants to minimize $L(\gamma, \lambda, f_t, c_t)$

In non-truthful auctions:

- λ is the pacing multiplier (we can still use online gradient descent)
- The bidder can choose $\gamma \in \Delta_b$ (and $b \sim \gamma$) using a regret minimizer for the reward function $L(\cdot, \lambda_t, f_t, c_t)$

Algorithm: Pacing strategy

```
input: Budget $B$, number of rounds $T$, learning rate $\eta$, primal regret minimizer $\mathcal{R}$;
initialization: $\rho \leftarrow B/T, \lambda_0 \leftarrow 0$;
for $t = 1, 2, \ldots, T$ do
    choose distribution over bids $\gamma_t \leftarrow \mathcal{R}(t)$;
    bid $b_t \sim \gamma_t$;
    observe $f_t(b_t)$ and $c_t(b_t)$ ;
    $\lambda_t \leftarrow \Pi_{[0,1/\rho]}(\lambda_{t-1} - \eta(\rho - c_t(b_t)))$ ;
    $B \leftarrow B - c_t(b_t)$;
    if $B < 1$ then
        terminate;
```

$\mathcal{R}$ is any regret minimizer and $\mathcal{R}(t)$ returns a distribution over bids at round t.

```
Algorithm: Pacing strategy
input: Budget $B$, number of rounds $T$, learning rate $\eta$, primal regret minimizer $\mathcal{R}$;
initialization: $\rho \leftarrow B/T, \lambda_0 \leftarrow 0$;
for $t = 1, 2, \dots, T$ do
    choose distribution over bids $\gamma_t \leftarrow \mathcal{R}(t)$;
    bid $b_t \sim \gamma_t$;
    observe $f_t(b_t)$ and $c_t(b_t)$ and $m_t$;
    $\lambda_t \leftarrow \Pi_{[0,1/\rho]}(\lambda_{t-1} - \eta(\rho - c_t(b_t)))$ ;
    $B \leftarrow B - c_t(b_t)$;
    if $B < 1$ then
        terminate;
```

$\mathcal{R}$ is any regret minimizer and $\mathcal{R}(t)$ returns a distribution over bids at round t.

Designing $\mathcal{R}$ with full feedback

Assumption

We assume to observe the highest competing bid m_t.

$\mathcal{R}$ is a regret minimizer for the **adversarial** expert problem with:

- Set of arms $\mathcal{B}$
- Reward $L(\cdot, \lambda_t, f_t, c_t)$

Designing $\mathcal{R}$ with full feedback

Assumption

We assume to observe the highest competing bid m_t.

$\mathcal{R}$ is a regret minimizer for the **adversarial** expert problem with:

- Set of arms $\mathcal{B}$
- Reward $L(\cdot, \lambda_t, f_t, c_t) \rightarrow$ **depends on λ_t that is not stochastic**

Designing $\mathcal{R}$ with full feedback

Assumption

We assume to observe the highest competing bid m_t.

$\mathcal{R}$ is a regret minimizer for the **adversarial** expert problem with:

- Set of arms $\mathcal{B}$
- Reward $L(\cdot, \lambda_t, f_t, c_t) \rightarrow$ **depends on λ_t that is not stochastic**
- We can use m_t to compute the reward $L(b, \lambda_t, f_t, c_t)$ for every possible bid (**full feedback**)

Designing $\mathcal{R}$ with full feedback

Assumption

We assume to observe the highest competing bid m_t.

$\mathcal{R}$ is a regret minimizer for the **adversarial** expert problem with:

- Set of arms $\mathcal{B}$
- Reward $L(\cdot, \lambda_t, f_t, c_t) \rightarrow$ **depends on λ_t that is not stochastic**
- We can use m_t to compute the reward $L(b, \lambda_t, f_t, c_t)$ for every possible bid (**full feedback**)
- ⚠ We need a regret minimizer that provides no-regret **with high probability**

Designing $\mathcal{R}$ with full feedback

Assumption

We assume to observe the highest competing bid m_t.

$\mathcal{R}$ is a regret minimizer for the **adversarial** expert problem with:

- Set of arms $\mathcal{B}$
- Reward $L(\cdot, \lambda_t, f_t, c_t)$ → **depends on λ_t that is not stochastic**
- We can use m_t to compute the reward $L(b, \lambda_t, f_t, c_t)$ for every possible bid (**full feedback**)
- ⚠ We need a regret minimizer that provides no-regret **with high probability** → **we don't want to satisfy the budget constraint in expectation**

Designing $\mathcal{R}$ with full feedback

We have seen a regret minimizer for this problem: **Hedge**.

Designing $\mathcal{R}$ with full feedback

We have seen a regret minimizer for this problem: **Hedge**.

⚠ The regret minimizer returns the distribution over arms **before** sampling. Hence, Hedge guarantees sublinear regret “deterministically”.

Designing $\mathcal{R}$ with full feedback

We have seen a regret minimizer for this problem: **Hedge**.

⚠ The regret minimizer returns the distribution over arms **before** sampling. Hence, Hedge guarantees sublinear regret “deterministically”.

Can we handle **bandit** feedback?

With bandit feedback (i.e., without observing m_t) we cannot use EXP3. We need **EXP3.P** that guarantees no-regret with high probability [Auer et al., 2002].

Stochastic environment

Theorem [Badanidiyuru et al., 2018]

Assume the sequence of m_t is stochastic. The pacing strategy with Hedge as regret minimizer $\mathcal{R}$ and $\eta = T^{-1/2}$ achieves regret

$$\widetilde{O}(\sqrt{T})$$

with high probability, where we ignore the dependency from the other parameters.

Stochastic environment

Proof sketch.

Assume that the budget is not depleted and hence the algorithm runs (almost) until round T **(we do not prove it)**. Since the reward and cost are stochastic

$$\sum_{t\in[T]} L(b, \lambda_t, f_t, c_t) \approx T\bar{L}(b, \bar{\lambda})$$

for each b with high probability, where $\bar{\lambda} = \frac{1}{T}\sum_{t\in[T]}\lambda_t$ is the average multiplier. Then, we use the no-regret property of Hedge that with high probability guarantees:

$$\sum_{t\in[T]} [f_t(b_t) - \lambda_t(c_t(b_t) - \rho)] \geq \sum_{t\in[T]} [f_t(\gamma^*) - \lambda_t(c_t(\gamma^*) - \rho)] - \widetilde{O}(\sqrt{T}),$$

where $\gamma^* \in \Delta_{\mathcal{B}}$ is the solution of the problem defining OPT (the best strategy in insight).

□

Stochastic environment

Proof sketch.

Hence,

$$\begin{aligned}\sum_{t\in[T]}[f_t(b_t)-\lambda_t(c_t(b_t)-\rho)] &\geq \sum_{t\in[T]}[f_t(\gamma^*)-\lambda_t(c_t(\gamma^*)-\rho)]-\widetilde{O}(\sqrt{T})\\ &\approx T\bar{L}(\gamma^*,\bar{\lambda})-\widetilde{O}(\sqrt{T})\\ &= T(\bar{f}(\gamma^*)-\bar{\lambda}[\underbrace{\bar{c}(\gamma^*)-\rho}_{\leq 0}])-\widetilde{O}(\sqrt{T})\\ &\geq T\ \mathrm{OPT}-\widetilde{O}(\sqrt{T}).\end{aligned}$$

Finally, $\sum_{t\in[T]}\lambda_t[c_t(b_t)-\rho]\geq -O(\sqrt{T})$ by the no-regret of gradient descent with respect to $\lambda=0$. Hence,

$$\sum_{t\in[T]}f_t(b_t)\geq T\ \mathrm{OPT}-\widetilde{O}(\sqrt{T}).$$

□

Adversarial environment: lower bound

Recall that we have shown that even in the simplest setting of truthful auctions:

Theorem

No algorithm can achieve strictly more than a $\rho := B/T$ fraction of the optimal utility.

Adversarial environment: regret guarantees

Theorem [Castiglioni et al., 2022]

The pacing strategy with Hedge as regret minimizer $\mathcal{R}$ and $\eta = T^{-1/2}$ guarantees utility at least:

$$\rho\ T\ OPT - \widetilde{O}(\sqrt{T}),$$

where

- OPT is the per-round reward of the best fixed **distribution** over bids
- $\rho := B/T$ is the per-round budget
- We ignore the dependency on the other parameters

Adversarial environment: regret guarantees

Theorem [Castiglioni et al., 2022]

The pacing strategy with Hedge as regret minimizer $\mathcal{R}$ and $\eta = T^{-1/2}$ guarantees utility at least:

$$\rho \ T \ OPT - \widetilde{O}(\sqrt{T}),$$

where

- OPT is the per-round reward of the best fixed **distribution** over bids
- $\rho \coloneqq B/T$ is the per-round budget
- We ignore the dependency on the other parameters

- If the environment is well-behaved then we can expect much better performance.
- If the environment changes "slightly" the guarantees approaches a $\widetilde{O}(\sqrt{T})$ regret.

A UCB-like approach

A UCB-like approach

Consider a **stochastic** environment and **bandit** feedback.

Natural approach: Estimate the parameters of the problem $\bar{f}$ and $\bar{c}$

As in the case of stochastic MABs, we want to be **optimistic** to incentivize **exploration**.

Idea: At each round t

- Estimate $\bar{f}$ with an **upper** confidence bound $\bar{f}_t^{UCB}$
- Estimate $\bar{c}$ with a **lower** confidence bound $\bar{c}_t^{LCB}$

Then, we play the optimal distribution γ_t over $\mathcal{B}$ using estimates:

$$OPT_t = \begin{cases} \sup\limits_{\gamma \in \Delta_{\mathcal{B}}} \ \bar{f}_t^{UCB}(\gamma) \\ \text{s.t. } \ \bar{c}_t^{LCB}(\gamma) \leq \rho \end{cases}$$

A UCB-like approach

Algorithm: UCB-Bidding Algorithm

```
input: Budget $B$, number of rounds $T$, learning rate $\eta$;
for $t = 1, \dots, T$ do
    for $b \in \mathcal{B}$ do
        $\bar{f}_t(b) \leftarrow \frac{1}{N_{t-1}(b)} \sum_{t'=1}^{t-1} f_{t'}(b)\mathbb{I}(b_{t'} = b)$;
        $\bar{f}_t^{UCB}(b) \leftarrow \bar{f}_t(b) + \sqrt{\frac{2\log(T)}{N_{t-1}(b)}}$;
        $\bar{c}_t(b) \leftarrow \frac{1}{N_{t-1}(b)} \sum_{t'=1}^{t-1} c_{t'}(b)\mathbb{I}(b_{t'} = b)$;
        $\bar{c}_t^{LCB}(b) \leftarrow \bar{c}_t(b) - \sqrt{\frac{2\log(T)}{N_{t-1}(b)}}$;
    compute $\gamma_t$ solution of the LP defining $\text{OPT}_t$;
    bid $b_t \sim \gamma_t$;
    observe $f_t(b_t)$ and $c_t(b_t)$ ;
    $B \leftarrow B - c_t(b_t)$;
    if $B < 1$ then
        terminate;
```

A UCB-like approach

Theorem [Agrawal and Devanur, 2014]

Assume the sequence of m_t is stochastic. The UCB-Bidding Algorithm provides regret $\widetilde{O}(\sqrt{T})$, where we ignore the dependence from the other parameters.

A UCB-like approach

Theorem [Agrawal and Devanur, 2014]

Assume the sequence of m_t is stochastic. The UCB-Bidding Algorithm provides regret $\widetilde{O}(\sqrt{T})$, where we ignore the dependence from the other parameters.

No guarantees for the adversarial setting since confidence bounds are designed for stochastic environments.

References

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. ACM*, 65 (3), mar 2018. ISSN 0004-5411. doi: 10.1145/3164539. URL `https://doi.org/10.1145/3164539`.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Matteo Castiglioni, Andrea Celli, and Christian Kroer. Online learning with knapsacks: the best of both worlds. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2767–2783, 2022.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.