Dual Mirror Descent for Online Allocation Problems

Santiago Balseiro [1 2] **Haihao Lu** [2] **Vahab Mirrokni** [2]

Abstract

We consider online allocation problems with concave revenue functions and resource constraints, which are central problems in revenue management and online advertising. In these settings, requests arrive sequentially during a finite horizon and, for each request, a decision maker needs to choose an action that consumes a certain amount of resources and generates revenue. The revenue function and resource consumption of each request are drawn independently and at random from a probability distribution that is unknown to the decision maker. The objective is to maximize cumulative revenues subject to a constraint on the total consumption of resources.

We design a general class of algorithms that achieve sub-linear expected regret compared to the hindsight optimal allocation. Our algorithms operate in the Lagrangian dual space: they maintain a dual multiplier for each resource that is updated using online mirror descent. By choosing the reference function accordingly, we recover dual sub-gradient descent and dual exponential weights algorithm. The resulting algorithms are simple, efficient, and shown to attain the optimal order of regret when the length of the horizon and the initial number of resources are scaled proportionally. We discuss applications to online bidding in repeated auctions with budget constraints and online proportional matching with high entropy.

1. Introduction

A central problem in revenue management and online advertising is the online allocation of requests subject to resource constraints. In revenue management, for example, firms such as hotels and airlines need to decide, when a request for a room or a flight arrives, whether to accept or decline the request (Talluri & van Ryzin, 2004). In search advertising, each time a user makes a search, the search engine has an opportunity to show an advertisement next to the organic search results (Mehta et al., 2007b). For each arriving user, the website collects bids from various advertisers who are interested in showing an ad and then needs to decide, in real time, which ad to show to the user. Such decisions are not made in isolation because resources are limited: hotels have limited number of rooms, planes have limited number of seats, and advertisers have limited budgets.

In this paper, we study allocation problems with concave revenue functions and resource constraints. Requests arrive sequentially during a finite horizon and, for each request, the decision maker needs to choose an action that consumes certain amount of resources and generates revenue. The objective of the decision maker is to maximize cumulative revenues subject to a constraint on the total consumption of resources. The revenue function and resource consumption of each request is learnt by the decision maker before making a decision. For example, airlines know the fare requested by the consumer before deciding whether to sell the ticket and search engines know advertisers' bids before deciding which ad to show. We assume that the revenue function and resource consumption of each request are drawn independently and at random from a fixed probability distribution. In practice, decision makers rarely know the probability distribution of requests in advance. Thus motivated, we consider a data-driven setting in which the underlying probability distribution is unknown to the decision maker. Performance of an online algorithm is measured using regret, which is given by the difference between the revenue attained by the optimal allocation with the benefit of hindsight (also referred as the offline optimum) and the cumulative revenues collected by the algorithm.

1.1. Our Results

We design a general class of algorithms that operate in the Lagrangian dual space. If the optimal dual variables were known in advance, the decision maker could, in principle, use these dual variables to price resources and decompose the problem across time periods. In practice, however, the optimal dual variables depend on the entire sequence of requests and are not known to the decision maker in advance. Our algorithms circumvent this issue by maintaining a dual

[1]Columbia University, New York, USA [2]Google Research, New York, USA. Correspondence to: Santiago Balseiro <srb2155@columbia.edu>, Haihao Lu <haihao.lu@chicagobooth.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

multiplier for each resource, which is updated after each request using online mirror descent. Actions are then taken using the estimated dual variables as a proxy for the opportunity cost of consuming resources. By choosing the reference function accordingly, we recover dual sub-gradient descent and dual exponential weights algorithm.

From the computational perspective, our algorithms are efficient; in many cases the dual variables can be updated after each request in linear time. This is in sharp contrast to most existing algorithms, which require periodically solving large convex optimizations problems or knowing bounds on the value benchmark (see Section 1.2). In many applications, such as online advertising, a massive number of decisions need to made in milliseconds and solving large optimizations problems is not operationally feasible.

We show that our algorithms attain regret of order $O(\sqrt{T})$ when the length of the horizon T and initial number of resources are scaled proportionally (Theorem 1). Because no algorithm can attain regret lower than $\Omega(\sqrt{T})$ under our minimal assumption (Lemma 1), these two results imply that our algorithms attain the optimal order of regret. To establish our regret bounds, we need to overcome two challenges: lower bounding the cumulative performance of our algorithm relative to the benchmark and showing that resources are not depleted too early in the horizon. We next describe these two challenges.

Recall that even though our algorithms operate in the dual space, performance is ultimately measured in the primal space. Standard results from the online mirror descent literature do not directly apply to our setting as these provide upper bounds on dual performance while the analysis requires lower bounds on primal performance. We overcome this first challenge by providing an analysis of dual online mirror descent that yields suitable lower bounds on primal performance (Proposition 3).

A requisite for obtaining good primal performance is not depleting resources too early; otherwise, the decision maker could miss good future opportunities. Our algorithms have a natural self-correcting feature that prevents them from depleting resources too early. By design, they target to consume a constant number of resources per period so as to deplete resources exactly at the end of the horizon. When a request consumes more (less) resources than the target, the corresponding dual variable is increased (decreased). Because resources are then priced higher (lower), future actions are chosen to consume resources more conservatively (aggressively). As a result, using the update rule of the dual variables, we can show that our algorithms never deplete resources too early (Proposition 2). To the best of our knowledge, this result is new to the online allocation literature and can be of interest for practitioners as, for example, advertisers have a preference for their ads to be delivered smoothly over time so as to maximize reach (Bhalgat et al., 2012; Lee et al., 2013; Xu et al., 2015). Our main result follows from combining these results together.

We then discuss applications to online bidding in repeated auctions with budget constraints and to online matching with high entropy (Section 5). As of 2019, around 85% of all display advertisements are bought programmatically – using automated algorithms (eMarketer, 2019). A common mechanism used by advertisers to buy ad slots is real-time auctions: each time a user visits a website, an auction is run to determine the ad to be shown in the user's browser. Because there is a large number of these advertising opportunities in a given day, advertisers set budgets to control their cumulative expenditure. There is thus a need to develop data-driven algorithm to optimize advertisers' bids in repeated auctions with budgets. This problem has been studied recently in Balseiro & Gur (2017; 2019), where they provide a dual sub-gradient descent algorithm that yields $O(\sqrt{T})$ regret. Our algorithms attain similar regret bounds with considerably weaker restrictions on the inputs. In particular, they assume that values and competing bids are independent, and that the dual objective is thrice differentiable and strongly convex. We require no such assumptions.

Online matching is another central problem in computer science, with applications in online advertisement allocation, job/server allocation in cloud computing, product recommendation under resource constraints, etc. It has been shown that a high-entropy proportional matching can lead to additional desirable properties, such as fairness and diversity (Lan et al., 2010; Venkatasubramanian, 2010; Qin & Zhu, 2013; Ahmed et al., 2017). We here study the online advertisement allocation problem, where at each time period, the decision maker matches an incoming impression with one advertiser (who may have a capacity constraint), aiming to maximize the total revenue over all incoming impressions while keeping a high entropy of such matchings. Recently Agrawal et al. (2018) studied a multi-round *offline* proportional matching algorithm for this problem setting. Our algorithm leads to a simple *online* counterpart to Agrawal et al. (2018) that yields similar regret/complexity bounds.

We conclude the paper by presenting numerical experiments on online proportional matching, which validate our results.

1.2. Related Work

Online allocation problems have a rich history in computer science and operations research.

Online allocation problems with linear revenue functions have been studied extensively in the so-called *random permutation model*. In the random permutation model, an adversary first selects a sequence of requests which are then

presented to the decision maker in random order. This model is more general than our setting in which requests are drawn independently and at random from an unknown distribution. Devanur & Hayes (2009) study online allocation problems in which revenues are proportional to amount of resources consumed (this is referred to as the Ad Words problem) and present a dual training algorithm with two phases: a training phase in which data is used to estimate the dual variables by solving a linear program and an exploitation phase in which actions are taken using the estimated dual variables. Their algorithm can be shown to obtain regret of order $O(T^{2/3})$. Feldman et al. (2010) present similar training-based algorithms for more general linear online allocation problems with similar regret guarantees. Pushing these ideas one step further, Agrawal et al. (2014) consider an algorithm that dynamically updates the dual variables by periodically solving a linear program using all data collected so far. This more sophisticated algorithm improves upon previous work by obtaining regret of order $O(T^{1/2})$. Compared to these papers, our algorithms work for general concave revenue functions, and for the linear case we obtain similar or better regret guarantees with simpler update rules that *do not require* solving large linear programs.

Devanur et al. (2019) study linear, online allocation problems when requests are drawn independently and at random from an unknown distribution, and provide algorithms that achieve $O(T^{1/2})$ regret. A key feature of their algorithms is that they require knowledge or estimates of the value of the benchmark (which in their case is the optimal allocation under the expected instance). When the value of the benchmark is known, they provide a simple algorithm that, similarly to ours, does not require solving a linear program in each stage. Their algorithm also maintains dual variables for the resource constraints, which are updated using an exponential update. When the value of the benchmark is unknown, they provide an algorithm that estimates the value of the benchmark by working in phases of geometrically increasing length. This algorithm, however, requires solving a linear program in each phase to estimate the value of the benchmark.

Closest to ours is Agrawal & Devanur (2015), which studies general online allocation problems that allow for concave objectives and convex feasibility constraints. They present a general class of algorithms that maintain dual variables for the constraints, which are updated using any black-box online convex optimization algorithm. When the objective is non-linear, they present fast algorithms that do not require solving a convex program when an estimate of the value of the benchmark is known. When an estimate is not available, their algorithm requires periodically solving a convex optimization program to estimate the value of the benchmark. Additionally, they allow resource constraints to be violated; they show that constraints are violated by at most $O(\sqrt{T})$. Because we require constraints to be satisfied for every realization, their algorithms are not feasible in our setting. When the objective is additively separable, they present an algorithm that updates dual variables using multiplicative weight updates and satisfies resource constraints for every realization. This algorithm, however, requires an estimate of the value of the benchmark that can be either provided as an input or obtained from solving a linear program once at the beginning of the horizon (Agrawal, 2019). Our paper extends their work by developing simple algorithms for concave, additively separable objectives that do not require estimates of the value of the benchmark and satisfy constraints for every realization under a large class of update rules. As a matter of fact, a key contribution of our work is showing that under a large class of reference functions, dual mirror descent does not deplete resources too early in every sample path (Proposition 2).

Our algorithm is an online dual mirror descent algorithm. It has been known in the optimization literature that mirror descent naturally minimizes a primal-dual gap in both deterministic and stochastic setting (Bach, 2015; Lu & Freund, 2018). However, the results therein do not apply directly to our setting because (i) as we will show later, our goal is to maximize the revenue in the online setting (i.e., (1)) rather than to maximize the natural primal objective (i.e., (19) in the appendix); (ii) we do not allow violations of the budget constraints in our online setting, while in the offline setting satisfying these constraints is easy since we can always shrink variables after the fact.

We discuss additional literature in Appendix B.

1.3. Notations

We define $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n | x \geq 0\}$ and $\mathbb{R}^n_{++} := \{x \in \mathbb{R}^n | x > 0\}$. We use $[m]$ as the shorthand of $\{1, \ldots, m\}$. $\mathbb{1}$ denotes the all-one vector, and e_j is the j-th standard unit vector.

2. Problem Formulation

We consider the following generic online convex problem with resource constraints:

$$(O): \quad \max_{x: x_t \in \mathcal{X}} \sum_{t=1}^{T} f_t(x_t) \quad \text{s.t.} \sum_{t=1}^{T} b_t x_t \leq T\rho \,, \tag{1}$$

where $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ is the decision variable at time t, $f_t \in \mathbb{R}^d \to \mathbb{R}$ is the concave revenue function received at time t, $b_t \in \mathbb{R}^{m \times d}_+$ is the entry-wise non-negative cost matrix received at time t, $\rho \in \mathbb{R}^m_{++}$ is the positive resource constraint vector. In the online setting, at each time period

$1 \le t \le T$, we receive a request (f_t, b_t), and we use an algorithm A to make a real-time decision x_t based on the current request (f_t, b_t) and the previous history $\mathcal{H}_{t-1} := \{f_s, b_s, x_s\}_{s=1}^{t-1}$, i.e.,

$$x_t = A(f_t, b_t | \mathcal{H}_{t-1}) \,. \tag{2}$$

Moreover, the constraint:

$$\begin{aligned} \sum_{s=1}^{t} b_s x_s \le \rho T \\ x_t \in \mathcal{X} \end{aligned} \tag{3}$$

must be satisfied for every $t \le T$. The above process generates total revenue $\sum_{t=1}^{T} f_t(x_t)$ at the end of the T time periods, and our goal is to a design algorithm A to maximize such revenue while satisfying constraint (3).

We assume the request (f_t, b_t) is generated i.i.d. from an unknown distribution $\mathcal{P} \in \mathcal{I}$, i.e., $(f_t, b_t) \in \{(f_1, b_1), \ldots, (f_n, b_n)\}$ with probability $\mathbb{P}((f_t, b_t) = (f_i, b_i)) = p_i$, where $\mathcal{I}$ denotes a family of distributions satisfying some regularity conditions (to be further discussed in Assumption 2). In particular, we define the expected revenue of an algorithm A over distribution $\mathcal{P}$ as

$$R(A|\mathcal{P}) = \mathbb{E}_{\mathcal{P}} \left[\sum_{t=1}^{T} f_t(x_t) \right] \,,$$

where x_t is computed by (2). The baseline we compare with is the expected revenue of the optimal solution in hindsight, which is also referred as the offline problem in the computer science literature. This amounts to solving for the optimal allocation under full information of all requests and then taking expectations over all possible realizations:

$$\text{OPT}(\mathcal{P}) = \mathbb{E}_{\mathcal{P}} \left[\begin{array}{ll} \max_{x_t \in \mathcal{X}} & \sum_{t=1}^{T} f_t(x_t) \\ \text{s.t.} & \sum_{t=1}^{T} b_t x_t \le T\rho \end{array} \right] . \tag{4}$$

We further define the regret of algorithm A as:

$$\text{Regret}(A|\mathcal{P}) := \text{OPT}(\mathcal{P}) - R(A|\mathcal{P}) \,,$$

and the worst-case regret of algorithm A over a family of distributions $\mathcal{I}$ as:

$$\text{Regret}(A|\mathcal{I}) := \sup_{\mathcal{P} \in \mathcal{I}} \{\text{OPT}(\mathcal{P}) - R(A|\mathcal{P})\} \,.$$

Since the probability distribution $\mathcal{P}$ is unknown to the decision maker, our goal is to design an algorithm A that works well for any distribution $\mathcal{P} \in \mathcal{I}$, namely, it has low worst-case regret $\text{Regret}(A|\mathcal{I})$.

2.1. The Dual Problem to (1)

In this section, we provide an upper bound of $\text{OPT}(\mathcal{P})$, which we call the offline dual problem to (1), and moreover, this dual problem inspires us to develop our main algorithm (Algorithm 1 in Section 2.2) for solving (1). Such upper bound in the linear case has been considered extensively in the literature (see, e.g., Talluri & van Ryzin 1998 for an example).

Define

$$f_i^*(c) := \max_{x \in \mathcal{X}} \{f_i(x) - c^\top x\} \tag{5}$$

as the conjugate function of $f_i(x)$ (restricted in $\mathcal{X}$)[1]. And define $D(\mu) : \mathbb{R}^m \to \mathbb{R}$ as

$$D(\mu) := \sum_{i=1}^{n} p_i f_i^*(b_i^\top \mu) + \mu^\top \rho \,,$$

then $D(\mu)$ provides a valid upper bound to $\text{OPT}(\mathcal{P})$:

Proposition 1. *It holds for any $\mu \ge 0$ that*

$$\textit{OPT}(\mathcal{P}) \le T D(\mu) \,. \tag{6}$$

Furthermore, we call

$$(D): \quad \min_{\mu \ge 0} D(\mu) = \sum_{i=1}^{n} p_i f_i^*(b_i^\top \mu) + \mu^\top \rho \,. \tag{7}$$

the offline dual problem to (1).

2.2. Online Dual Mirror Descent

Online mirror descent algorithm is a standard algorithm in online convex optimization (Hazan et al., 2016). In this section, we present the online mirror descent algorithm on the dual problem (7), while our goal is to obtain a good solution to the original primal problem (1).

To discuss the mirror descent algorithm, first recall that the Bregman divergence with respect to a given convex reference function $h(\cdot)$ is defined as $V_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle$. Algorithm 1 presents the main algorithm we study in this paper. At time t, we receive a request (f_t, b_t), and we compute the optimal response $\tilde{x}_t$ that maximizes an opportunity cost-adjusted revenue of this request based on the current dual solution μ_t. We then take this action (i.e., $x_t = \tilde{x}_t$) if that does not exceed the resource constraint, otherwise we take a void action (i.e., $x_t = 0$). Notice that it follows from the definition of conjugate function (5) that $-b_t \tilde{x}_t \in \partial_\mu f_t^*(b_t^\top \mu_t)$. [2] Thus $\tilde{g}_t := -b_t \tilde{x}_t + \rho$ is an unbiased stochastic estimator of the gradient of the dual problem $D(\mu)$ at μ_t:

$$\mathbb{E}_{\mathcal{P}} [\tilde{g}_t] = \mathbb{E}_{\mathcal{P}} [-b_t \tilde{x}_t + \rho] \in \sum_{i=1}^{n} p_i \partial_\mu f_i^*(b_i^\top \mu_t) + \rho \in \partial_\mu D(\mu_t) \,.$$

[1] More precisely, $f_i^*(c)$ is the conjugate function of $f_i(x) + \mathbf{1}\{x \in \mathcal{X}\}$ under the standard definition of conjugate function, where $\mathbf{1}\{x \in \mathcal{X}\}$ is the indicator function of the constraint.

[2] ∂ here refers to the set of super-derivatives of a concave function.

Algorithm 1 Dual Mirror Descent Algorithm for (1)

Input: Initial dual solution μ_0, total time period T, remaining resources $B_0 = T\rho$, reference function $h(\cdot) : \mathbb{R}^m \to \mathbb{R}$, and step-size η.

for $t = 0, \ldots, T-1$ **do**

Receive $(f_t, b_t) \sim \mathcal{P}$, i.e., $\mathbb{P}((f_t, b_t) = (f_i, b_i)) = p_i$.

Make the primal decision and update the remaining resources:

$$\begin{aligned} \tilde{x}_t &= \arg\max_{x \in \mathcal{X}} \{f_t(x) - \mu_t^\top b_t x\} , \\ x_t &= \begin{cases} \tilde{x}_t & \text{if } b_t \tilde{x}_t \le B_t \\ 0 & \text{otherwise} \end{cases} , \\ B_{t+1} &= B_t - b_t x_t . \end{aligned}$$

Obtain a stochastic sub-gradient of $D(\mu_t)$:

$$\tilde{g}_t = -b_t \tilde{x}_t + \rho .$$

Update the dual variable by mirror descent:

$$\mu_{t+1} = \arg\min_{\mu \ge 0} \langle \tilde{g}_t, \mu \rangle + \frac{1}{\eta} V_h(\mu, \mu_t) . \qquad (8)$$

end

We then utilize $\tilde{g}_t$ to update the dual variable by performing an online mirror descent step (8) with step-size η.

Algorithm 1 only takes an initial dual variable and a step size as inputs and is simple to implement. In most cases, the mirror descent step can be computed in linear time as (8) admits a closed-form solution. For example, if the reference function is $h(\mu) = -\sum_i \mu_i \log(\mu_i)$, the dual update (8) becomes

$$\mu_{t+1} = \mu_t * \exp(-\eta \tilde{g}_t) ,$$

which recovers the online exponential weights algorithm for solving (7); if the reference function is $h(\mu) = \frac{1}{2}\|\mu\|_2^2$, the dual update (8) becomes

$$\mu_{t+1} = \text{Proj}_{\mu \ge 0}\{\mu_t - \eta \tilde{g}_t\} ,$$

which recovers the online sub-gradient descent method for solving (7).

3. Regret Bound

In this section, we present the worst-case regret bound of Algorithm 1 for solving (1). First we state the assumptions required in our analysis.

3.1. Assumptions

Assumption 1. *(Assumptions on constraint set $\mathcal{X}$). We assume that: (i) $\mathcal{X}$ is a convex and bounded set in $\mathbb{R}_+^d$, and (ii) $0 \in \mathcal{X}$.*

The above assumption implies that we can only take non-negative actions. Moreover, we can always take the void action by choosing $x_t = 0$ in order to make sure we do not exceed the resource constraints. This guarantees the existence of a feasible solution.

Assumption 2. *(Assumptions on distribution family $\mathcal{I}$). For any $\mathcal{P} \in \mathcal{I}$, it holds that*

1. *$\mathcal{P}$ has finite support: $\mathcal{S}(\mathcal{P}) := \{(f_1, b_1), \ldots, (f_n, b_n)\}$.*
2. *For any $(f_i, b_i) \in \mathcal{S}(\mathcal{P})$, it holds that (i) $f_i(x) \ge 0, \forall x \in \mathcal{X}$; (ii) $f_i(0) = 0$; (iii) $b_i \ge 0$; and $f_i(x)$ is a concave function in $\mathcal{X}$.*
3. *There exists $\bar{f} \in \mathbb{R}_{++}$ such that $f_i(x) \le \bar{f}$ for any $x \in \mathcal{X}$ and $(f_i, b_i) \in \mathcal{S}(\mathcal{P})$.*
4. *There exists $\bar{b} \in \mathbb{R}_{++}$ such that $\|b_i x\|_\infty \le \bar{b}$ for any $x \in \mathcal{X}$ and $(f_i, b_i) \in \mathcal{S}(\mathcal{P})$.*

We herein assume $\mathcal{P}$ has finite support for simplicity of the argument. Interestingly, our regret bounds do not depend on n (the cardinality of the support of the distribution) and we conjecture our results continue to hold in the case of an infinite support. The upper bound $\bar{f}$ and $\bar{b}$ impose regularity on the probability class $\mathcal{I}$, and they will appear in the regret bound. The assumption $b_i \ge 0$ implies that we cannot replenish resources once they are consumed. The assumption $f_i(0) = 0$ is without loss of generality since we can always subtract a constant from the function $f_i(x)$.

Assumption 3. *(Assumptions on resource parameter ρ). We assume there exist $\bar{\rho}, \underline{\rho} \in \mathbb{R}_{++}$ such that for any $j \in [m]$, $\underline{\rho} \le \rho_j \le \bar{\rho}$.*

Remark 1. *Without loss of generality, we can assume $\rho_j = 1$ for any $j \in [m]$ by rescaling the j-th row in b_i. This may lead to slightly favorable regret bound, but we herein choose to keep ρ for its generality.*

Definition 1. *We define $\mu^{\max} \in \mathbb{R}^m$ such that $\mu_j^{\max} := \frac{\bar{f}}{\rho_j} + 1$.*

As we will show later in Proposition 2, as long as $0 \le \mu_0 \le \mu^{\max}$, the dual variable obtained by Algorithm 1 satisfies $\mu_t \le \mu^{\max}$ at any time t. In other words, μ_t attained by Algorithm 1 always stays in domain $\mathcal{D} := \{\mu \in \mathbb{R}^m | 0 \le \mu \le \mu^{\max}\}$.

Assumption 4. *(Assumptions on reference function $h(\cdot)$). We assume*

1. *$h(\mu)$ is coordinate-wisely separable, i.e., $h(\mu) = \sum_{j=1}^m h_j(\mu_j)$ where $h_j(\cdot)$ is a convex univariate function.*

2. *$h(\mu)$ is σ_1-strongly convex in ℓ_1-norm in $\mathcal{D}$, i.e., $h(\mu_1) \geq h(\mu_2) + \langle \nabla h(\mu_2), \mu_1 - \mu_2 \rangle + \frac{\sigma_1}{2}\|\mu_1 - \mu_2\|_1^2$ for any $\mu_1, \mu_2 \in \mathcal{D}$.*

3. *$h(\mu)$ is σ_2-strongly convex in ℓ_2-norm in $\mathcal{D}$, i.e., $h(\mu_1) \geq h(\mu_2) + \langle \nabla h(\mu_2), \mu_1 - \mu_2 \rangle + \frac{\sigma_2}{2}\|\mu_1 - \mu_2\|_2^2$ for any $\mu_1, \mu_2 \in \mathcal{D}$.*

Strong convexity of the reference function is a standard assumption for the analysis of mirror descent algorithms (Bubeck, 2015). Indeed, the strong convexity in ℓ_1-norm and ℓ_2-norm are equivalent (up to a dimension-dependent constant). We here assume strong convexity in both norms in order to obtain a tighter regret bound.

If $h(\cdot)$ is not a coordinate-wise separable function, the sub-problem (8) can be hard to solve. Furthermore, most examples in the mirror descent literature utilize coordinate-wise separable reference functions (Nemirovsky & Yudin, 1983; Beck & Teboulle, 2003; Bubeck, 2015; Lu et al., 2018; Lu, 2017).

3.2. Master Theorem

The next theorem presents the worst-case regret bound of Algorithm 1.

Theorem 1. *Consider Algorithm 1 with step-size $\eta \leq \frac{\sigma_2}{\bar{b}}$ and initial solution $\mu_0 \leq \mu^{\max}$. Suppose Assumption 1-4 are satisfied. Then it holds for any $T \geq 1$ that*

$$\begin{aligned}\mathrm{Regret}(A|\mathcal{I}) \leq & \frac{2\left(\bar{b}^2 + \bar{\rho}^2\right)}{\sigma_1}\eta T + \frac{V_h(0,\mu_0)}{\eta} \\ & + \frac{\bar{f}}{\underline{\rho}\eta}\|\nabla h(\mu^{\max}) - \nabla h(\mu_0)\|_\infty + \frac{\bar{f}\bar{b}}{\underline{\rho}}\,.\end{aligned} \tag{9}$$

When choosing $\eta = O(1/\sqrt{T})$, we obtain that $\mathrm{Regret}(A|\mathcal{I}) \leq O(\sqrt{T})$ when T is sufficiently large, and, therefore, our algorithm yields sublinear regret.

Remark 2. *In this remark, we assume $\bar{\rho} = \underline{\rho} = 1$ (this is without loss of generality as mentioned in Remark 1). Here we consider two special cases of Theorem 1 when T is sufficiently large:*

1. *Suppose $h(\mu) = \frac{1}{2}\|\mu\|^2$ and $\mu_0 = 0$, then Algorithm 1 recovers dual online sub-gradient descent, and with proper step-size η we can obtain*

$$\mathrm{Regret}(A|\mathcal{I}) \leq 2\sqrt{2m\bar{f}^2(\bar{b}^2+1)}\sqrt{T} + \bar{f}\bar{b}\,.$$

2. *Suppose $h(\mu) = -\sum_{j=1}^m \mu_j \log \mu_j$ and $\mu_0 = e^{-1}\mathbb{1}$, then Algorithm 1 recovers the dual multiplicative update algorithm, and with proper step-size η we can obtain*

$$\begin{aligned}&\mathrm{Regret}(A|\mathcal{I}) \leq \bar{f}\bar{b} + \\ &2\sqrt{2m(\bar{b}^2+1)\left(\bar{f}+1\right)\left(\bar{f}(\log\left(\bar{f}+1\right)+1) + me^{-1}\right)}\sqrt{T}\,.\end{aligned}$$

We next discuss the tightness of our regret bound. The following result, which we reproduce without proof, shows that one cannot hope to attain regret lower than $\Omega(\sqrt{T})$ under our modeling assumptions.

Lemma 1 (Lemma 1 from Arlotto & Gurvich 2019)**.** *For every $T \geq 1$, there exists a probability distribution $\mathcal{P}$ such that*

$$\inf_A \mathrm{Regret}(A|\mathcal{P}) \geq C\sqrt{T}$$

where C is a constant independent of T.

The previous result shows that, for every T, there exists a probability distribution under which all algorithms –even those that known the probability distribution– incur $\Omega(\sqrt{T})$ regret. The worst-case distribution used in the proof of the result assigns mass to three points with one point having mass of order $1/\sqrt{T}$. Because the regret bound of Algorithm 1 provided in Theorem 1 does not depend on the probability mass function of the distribution $\mathcal{P}$, it readily follows that our algorithm also attains $O(\sqrt{T})$ in such worst-case instance. This implies that our algorithm attains the optimal order of regret when the length of the horizon and initial number of resources are scaled proportionally.

We remark that the dependency of our regret bounds on the number of resources m is sub-optimal. Agrawal et al. (2014) shows that the best possible dependence on the number of resources is of order $\log(m)$ while our algorithms' dependency is of polynomial order on m (in particular, $m^{1/2}$ for dual online sub-gradient descent and $m^{3/2}$ for dual multiplicative weight updates). The algorithms in Agrawal et al. (2014), Agrawal & Devanur (2015), and Devanur et al. (2019) attain the optimal dependency on the number of resources, but, differently to ours, require either knowing an estimate on the value of benchmark or periodically solving large optimization problems.

4. Proof Sketch of Theorem 1

There are two major steps in the proof of Theorem 1. We need to show that: (i) Algorithm 1 does not deplete resources too early (Proposition 2); and (ii) before running out of the resources, the average cumulative revenue is close to a dual objective value (Proposition 3), which provides an upper bound of $\mathrm{OPT}(\mathcal{P})$. Here we present these two major steps.

Step 1 (Lower bound on the stopping time): At first, we define the stopping time of Algorithm 1:

Definition 2. *We define the stopping time τ_A of Algorithm 1 as the first time less than T that there exists resource j such that*

$$\sum_{t=1}^{\tau_A} (b_t)_j^\top x_t + \bar{b} \geq \rho_j T \, .$$

Notice that τ_A is a random variable, and moreover, we will not violate the resource constraints before the stopping time τ_A. The next proposition says the stopping time τ_A is close to the end of the horizon T.

Proposition 2. *Consider Algorithm 1 with step-size $\eta \leq \frac{\sigma_2}{\bar{b}}$. Then it holds that $\mu_t \leq \mu^{\max}$ for any $t \leq T$. Furthermore, it holds with probability 1 that*

$$T - \tau_A \leq \frac{1}{\eta \underline{\rho}} \|\nabla h(\mu^{\max}) - \nabla h(\mu_0)\|_\infty + \frac{\bar{b}}{\underline{\rho}} \, . \qquad (10)$$

Step 2 (Primal-dual bound on the cumulative revenue): We here study the primal-dual gap until the stopping-time τ_A. Notice that before the stopping time τ_A, Algorithm 1 performs the standard mirror descent steps on the dual function.

Let us denote the random variable γ_t to be the type of request in time period t, i.e., γ_t is the random variable that determines the (stochastic) sample i in the t-th iteration of Algorithm 1. Then μ_{t+1} is a random variable which depends on all previous values $\gamma_0, \ldots, \gamma_t$ and we denote this string of random variables $\xi_t = \{\gamma_0, \ldots, \gamma_t\}$.

The next Proposition presents a primal-dual bound on the cumulative revenue of Algorithm 1 before the stopping time τ_A.

Proposition 3. *Consider the Algorithm 1 with given step-size η under Assumptions* 1-4. *Let τ_A be the stopping time defined in Definition 2. Denote $\bar{\mu}_{\tau_A} = \frac{\sum_{t=1}^{\tau_A} \mu_t}{\tau_A}$. Then the following inequality holds:*

$$\mathbb{E}_{\mathcal{P}} \left[\tau_A D(\bar{\mu}_{\tau_A}) - \sum_{t=1}^{\tau_A} f_t(x_t) \right] \leq \frac{2\left(\bar{b}^2 + \bar{\rho}^2\right)}{\sigma_1} \eta \mathbb{E}_{\mathcal{P}} \left[\tau_A\right] + \frac{V_h(0, \mu_0)}{\eta} \, .$$

Together with the above two steps, we can show that the cumulative revenue till the stopping time is not far away from the optimal revenue to the offline problem (1) by using Proposition 1. We present the proof of Proposition 2, Proposition 3 and Theorem 1 in Appendix D, Appendix E and Appendix F, respectively.

5. Applications

In this section, we discuss applications of Algorithm 1 to online matching with high entropy and bidding in repeated auctions with budgets.

5.1. Bidding in Repeated Auctions with Budgets

Most online advertisements are sold using auctions in which advertisers bid based on viewer-specific information. Typically advertisers participate in a large number of auctions on a given day, and they set budgets to control their cumulative expenditure throughout the day. We discuss how to apply our methods to the problem of bidding in repeated auctions with budgets.

We consider an advertiser with a budget ρT that limits the cumulative expenditure over T auctions. Each request corresponds to an auction in which an *impression* becomes available for sale. When the t-th impression arrives, the advertiser first learns a value v_t for winning the impression based viewer-specific information and then determines a bid w_t to submit to the auction. We assume that impressions are sold using a second-price auction. Denoting by b_t the highest bid submitted by competitors, the advertiser wins whenever his bid is the highest (i.e., $w_t \geq b_t$) and pays the second-highest bid in case of winning (i.e., $b_t \mathbf{1}\{w_t \geq b_t\}$). To simplify the exposition, we assume that ties are broken in favor of the advertiser. At the point of bidding, the advertiser does not know the highest competing bid. Consistent with practice, we assume that the advertiser only observes his payment in case of winning.

Values and competing bids are drawn i.i.d. from an unknown, discrete distribution. The assumption that competing bids are i.i.d. can be motivated by mean-field models in which each agent is assumed to compete with a stationary bidding landscape (see, e.g., Iyer et al. 2014; Balseiro et al. 2015). These are predicated on the fact that, typically, the number of bidders in online advertising markets is large (in the orders of hundreds or thousands) and that in each auction an advertiser competes with a small random set of different bidders. As a result, the competing bids an advertiser faces tends to be independently distributed and exogenous (i.e., not affected by the past bids of the decision maker).

The problem of bidding in repeated auctions with budgets has been studied recently in Balseiro & Gur (2017; 2019). In their paper, they present an adaptive pacing strategy that attempts to learn an optimal Lagrange multiplier using subgradient descent. Their adaptive pacing strategy is shown to attain $O(\sqrt{T})$ regret under restrictive assumptions on the distribution of inputs. Specifically, they assume that values and competing bids are independent, and that $D(\mu)$ is thrice differentiable and strongly convex. In practice, however, values and competing bids are positively correlated. Our algorithms attain similar regret bounds without such restrictive assumptions on the inputs.

With the benefit of hindsight, a decision maker can win an auction by bidding an amount equal to the highest competing bid (i.e., $w_t = b_t$). Therefore, the optimal solution in hindsight reduces to solving a knapsack problem in which

the impressions to be won are chosen to maximize the net utility subject to the budget constraint. The problem is given by:

$$\max_{x_t \in \{0,1\}} \sum_{t=1}^{T} (v_t - b_t) x_t$$
$$\text{s.t.} \sum_{t=1}^{T} b_t x_t \le T\rho,$$

where $x_t \in \{0,1\}$ is a decision variable indicating whether the advertiser wins the t-th impression.

Note that the informational assumptions are different from the ones of our baseline model because the competing bid b_t is not assumed to be known at the point of bidding. Interestingly, because ads are sold using an ex-post incentive compatible auction, such information is not necessary for our algorithm: the algorithm only needs to know the payment incurred. As a matter of fact, our analysis applies to any other ex-post incentive compatible auction.

This problem can be mapped to our framework by setting $f_t(x) = (v_t - b_t)x$. Denoting by $\mu_t \ge 0$ the dual multiplier of the budget constraint, the primal decision in Algorithm 1 is given by

$$\tilde{x}_t = \arg\max_{x \in \{0,1\}} \{f_t(x) - \mu_t b_t x\} = \mathbf{1}\{v_t \ge (1+\mu_t) b_t\}$$

This decision can be implemented by bidding $w_t = v_t/(1+\mu_t)$ without knowing the maximum competing bid. We present the formal algorithm (Algorithm 2) in Appendix A. Theorem 1 readily implies that choosing $\eta \sim 1/\sqrt{T}$ yields a regret of $O(\sqrt{T})$.

5.2. Proportional Matching with High Entropy

We consider an online matching problem using the terminology from online advertising. Suppose there are n different impressions and m advertisers. At time period t, an impression with revenue vector $r_t \in \mathbb{R}^m$ arrives, i.e., if we allocate it to advertiser $j \in [m]$, then it generates revenue $(r_t)_j$.

In the online setting, the impressions arrive sequentially. For each time period t, we decide an assignment probability variable $x_t \in \mathcal{X} := \{x \in \mathbb{R}^m_+ | \sum_{i=1}^m x_i \le 1\}$, and assign the arriving impression to advertiser j with probability $(x_t)_j$. Notice that with probability $1 - \sum_{j=1}^m (x_t)_j$ the impression is not assigned to any advertiser, and in practice, such impressions will go to other traffic. Suppose there are in total T time periods, and we assume the capacity of the j-th advertiser is $\rho_j T$. Define

$$H(x) := -\sum_{j=1}^{m} x_j \log(x_j) - \left(1 - \sum_{j=1}^{m} x_j\right) \log\left(1 - \sum_{j=1}^{m} x_j\right)$$

to be the entropy function of assignment probability x.

We herein study the high entropy fractional matching, where the goal is to find a fractional matching $\{x_t\}_t$ to maximize the revenue with an entropy regularizer. The hindsight problem is:

$$\max_{x_t \in \mathcal{X}} \sum_{t=1}^{T} r_t^\top x_t + \lambda H(x_t)$$
$$\text{s.t.} \sum_{t=1}^{T} v_t \le T\rho\,, \tag{11}$$

where λ is the parameter of the entropy regularizer and v_t is a random variable defined by (12).

A matching with high entropy has been shown to possess many additional desirable properties, for example, higher fairness and higher diversity (Lan et al., 2010; Venkatasubramanian, 2010; Qin & Zhu, 2013; Ahmed et al., 2017). Recently (Agrawal et al., 2018) designed a multi-round offline proportional allocation algorithm for solving (11). Our algorithm, in contrast, is a simpler online algorithm and does not need to be run multi-rounds. A major difference is that their capacity constraints can be violated during the runs because they allow rescaling the variables at the end of the run to satisfy the capacity constraints, while we do not allow such violation in our online setting. Refer to Agrawal et al. (2018) for a more detailed literature review on the background of this problem.

Algorithm 3 in Appendix A is a variant of Algorithm 1 for the above proportional matching problem (11), with $f_t(x) = r_t^\top x + \lambda H(x)$ and $b_t = I$. The only difference is that in the constraints we need to take into account the actual realization of the probabilistic matching. Define the random variable

$$v_t = \begin{cases} e_j & \text{w.p. } x_j \\ 0 & \text{w.p. } 1 - \sum_{j=1}^m x_j \end{cases}, \tag{12}$$

where $e_j \in \mathbb{R}^m$ is the j-th standard unit vector in $\mathbb{R}^m$. Then v_t characterizes the realized assignment of the impression at time t. In the online problem, the constraint of (11) is stated in terms of v_t, i.e., the random realization of the decision variable x_t. Let ζ denote the random variable determines the realization in the above process, then the results in Theorem 1 still holds after taking the expectation over ζ on the left-hand of (9):

Proposition 4. *Consider Algorithm 3 with step-size $\eta \sim O(\frac{1}{\sqrt{T}})$ and initial solution $\mu_0 = 0$ for solving (11). Then it holds that*

$$\text{Regret}(A|\mathcal{I}) := \sup_{\mathcal{P} \in \mathcal{I}} \{\textit{OPT}(\mathcal{P}) - \mathbb{E}_\zeta R(A|\mathcal{P})\} \le O\left(\sqrt{T}\right).$$

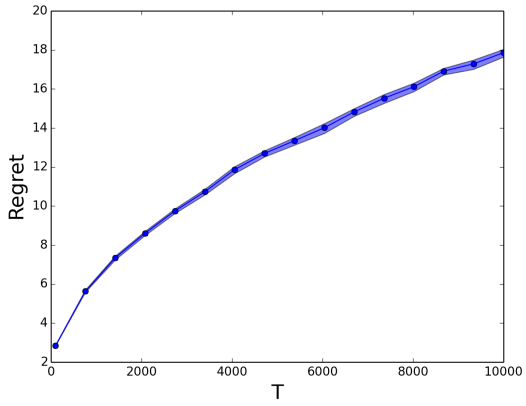

Figure 1. Regret versus horizon T for the numerical experiments on proportional matching with high entropy.

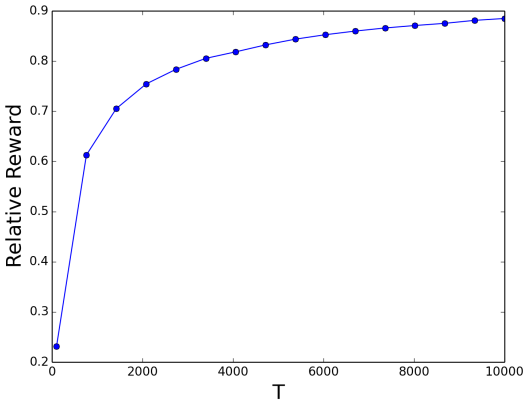

Figure 2. Relative reward versus horizon T for the numerical experiments on proportional matching with high entropy.

6. Numerical Experiment

Here, we present a numerical experiment on proportional matching with high entropy (Section 5.2) to verify our results.

The dataset is generated following the procedures stated in Balseiro et al. (2014). For each value of T, we plot the average regret and its 95% confidence interval over 400 random trials. For all experiments, we start from $\mu_0 = 0$, utilize $h(x) = \frac{1}{2}\|x\|_2^2$ as our reference function (thus the algorithm is dual sub-gradient descent), and choose $\eta = \frac{1}{\sqrt{T}}$ as the step-size. The details of the numerical experiment and data generation are presented in Appendix H, and the code to reproduce the results is in supplementary materials.

Figure 1 plots the regret versus horizon T, from which we can clearly see that the regret grows at the rate of $\sqrt{T}$, which verifies the results in Theorem 1.

Figure 2 plots the relative reward (ratio between the reward collected by the online algorithm and the offline optimal) versus horizon T. As we can see in Figure 2, the relative reward gets to around 90% with $10,000$ online iterations, which showcases the effectiveness of our proposed algorithm.

We present additional results and discussions in Appendix H.

7. Conclusion

In this paper, we present a class of simple and efficient algorithms for online allocation problems with concave revenue functions. We show that our algorithms attain $O(\sqrt{T})$ regret, which matches the lower bound. Numerical experiments validate our results. Interesting future research directions are to explore whether better regret bounds can be obtained under more restrictive assumptions on the inputs and to study the performance of online dual mirror descent on more general inputs (e.g., non-stationary or adversarial inputs).

Acknowledgement

The authors would like to thank Balasubramanian Sivan, Rad Niazadeh, Shipra Agrawal and three anonymous reviewers for providing a number of valuable comments that greatly improved the paper.

References

Agrawal, S. Private communications. 2019.

Agrawal, S. and Devanur, N. R. Fast algorithms for online stochastic convex programming. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pp. 1405–1424, USA, 2015. Society for Industrial and Applied Mathematics.

Agrawal, S., Wang, Z., and Ye, Y. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62 (4):876–890, 2014.

Agrawal, S., Devanur, N. R., and Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18, 2016.

Agrawal, S., Zadimoghaddam, M., and Mirrokni, V. Proportional allocation: Simple, distributed, and diverse matching with high entropy. In *International Conference on Machine Learning*, pp. 99–108, 2018.

Ahmed, F., Dickerson, J. P., and Fuge, M. Diverse weighted bipartite b-matching. *arXiv preprint arXiv:1702.07134*, 2017.

Arlotto, A. and Gurvich, I. Uniformly bounded regret in the multisecretary problem. *Stochastic Systems*, 9(3):231–260, 2019.

Bach, F. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

Badanidiyuru, A., Langford, J., and Slivkins, A. Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134, 2014.

Balseiro, S. R. and Gur, Y. Learning in repeated auctions with budgets: Regret minimization and equilibrium. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, pp. 609, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345279.

Balseiro, S. R. and Gur, Y. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968, 2019.

Balseiro, S. R., Feldman, J., Mirrokni, V., and Muthukrishnan, S. Yield optimization of display advertising with ad exchange. *Management Science*, 60(12):2886–2907, 2014.

Balseiro, S. R., Besbes, O., and Weintraub, G. Y. Repeated auctions with budgets in ad exchanges: Approximations and design. *Management Science*, 61(4):864–884, 2015.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Bhalgat, A., Feldman, J., and Mirrokni, V. Online allocation of display ads with smooth delivery. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pp. 1213–1221, New York, NY, USA, 2012. Association for Computing Machinery.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Chen, G. and Teboulle, M. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Devanur, N. R. and Hayes, T. P. The adwords problem: online keyword matching with budgeted bidders under random permutations. In *Proceedings of the 10th ACM conference on Electronic commerce*, EC '09, pp. 71–78. ACM, 2009.

Devanur, N. R., Jain, K., Sivan, B., and Wilkens, C. A. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *J. ACM*, 66(1), jan 2019.

eMarketer. Us programmatic ad spending forecast 2019. April 2019. Retrieved from http://www.emarketer.com.

Feldman, J., Korula, N., Mirrokni, V., Muthukrishnan, S., and Pál, M. Online ad assignment with free disposal. In *Proceedings of the 5th International Workshop on Internet and Network Economics*, WINE '09, pp. 374–385. Springer-Verlag, 2009.

Feldman, J., Henzinger, M., Korula, N., Mirrokni, V. S., and Stein, C. Online stochastic packing applied to display ad allocation. In *Proceedings of the 18th annual European conference on Algorithms: Part I*, ESA'10, pp. 182–194. Springer-Verlag, 2010.

Hazan, E. et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Iyer, K., Johari, R., and Sundararajan, M. Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12): 2949–2970, 2014.

Jasin, S. Performance of an lp-based control for revenue management with unknown demand parameters. *Operations Research*, 63(4):909–915, 2015.

Lan, T., Kao, D., Chiang, M., and Sabharwal, A. *An axiomatic theory of fairness in network resource allocation*. IEEE, 2010.

Lee, K.-C., Jalali, A., and Dasdan, A. Real time bid optimization with smooth budget delivery in online advertising. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD '13, New York, NY, USA, 2013. Association for Computing Machinery.

Li, X. and Ye, Y. Online linear programming: Dual convergence, new algorithms, and regret bounds. 2019.

Lu, H. "Relative-continuity" for non-lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *arXiv preprint arXiv:1710.04718*, 2017.

Lu, H. and Freund, R. M. Generalized stochastic frank-wolfe algorithm with stochastic" substitute" gradient for structured convex optimization. *arXiv preprint arXiv:1807.07680*, 2018.

Lu, H., Freund, R., and Nesterov, Y. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Mehta, A., Saberi, A., Vazirani, U., and Vazirani, V. Adwords and generalized online matching. *J. ACM*, 54:22:1–22:19, October 2007a.

Mehta, A., Saberi, A., Vazirani, U., and Vazirani, V. Adwords and generalized online matching. *J. ACM*, 54(5):22–es, October 2007b. ISSN 0004-5411.

Nemirovsky, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York, 1983.

Qin, L. and Zhu, X. Promoting diversity in recommendation by entropy regularizer. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Talluri, K. and van Ryzin, G. An analysis of bid-price controls for network revenue management. *Management Science*, 44(11): 1577–1593, 1998.

Talluri, K. T. and van Ryzin, G. J. *The Theory and Practice of Revenue Management*. International Series in Operations Research & Management Science, Vol. 68. Springer, 2004.

Venkatasubramanian, V. Fairness is an emergent self-organized property of the free market for labor. *Entropy*, 12(6):1514–1531, 2010.

Xu, J., Lee, K.-c., Li, W., Qi, H., and Lu, Q. Smart pacing for effective online ad campaign optimization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 2217–2226, New York, NY, USA, 2015. Association for Computing Machinery.

A. Formal Algorithms for Solving the Two Applications in Section 5

Algorithm 2 Online Dual Mirror Descent Algorithm for Bidding in Repeated Auctions

Input: Initial dual solution μ_0, reference function $h(\cdot) : \mathbb{R} \to \mathbb{R}$, step-size η.

for $t = 0, \ldots, T-1$ **do**

Receive an impression with value v_t.

Bid $w_t = \min\{\tilde{w}_t, B_t\}$ where $\tilde{w}_t = v_t/(1+\mu_t)$ and B_t is the remaining budget.

Observe the payment $q_t = b_t \mathbf{1}\{w_t \geq b_t\}$.

Obtain a stochastic dual sub-gradient

$$\tilde{g}_t = -q_t + \rho \, .$$

Update the dual variable using mirror descent:

$$\mu_{t+1} = \arg\min_{\mu \geq 0} \langle \tilde{g}_t, \mu \rangle + \frac{1}{\eta} V_h(\mu, \mu_t) \, .$$

end

Algorithm 3 Online Dual Mirror Descent Algorithm for Proportional Matching Problems with High Entropy

Input: Initial dual solution μ_0, and step-size η.

for $t = 0, \ldots, T-1$ **do**

Receive an impression with revenue vector r_t, and regularized revenue function $f_t(x) = r_t^\top x + \lambda H(x)$.

Decide the assignment probability and update the remaining capacity:

$$x_t = \arg\max_{x \in \mathcal{X}} \{f_t(x) - \mu_t^T x\}$$

or equivalently

$$(x_t)_j = \frac{\exp((r_t(j) - \mu_t(j))/\lambda)}{\sum_{l=1}^m \exp((r_t(l) - \mu_t(l))/\lambda) + 1} \, .$$

Make the allocation decision: v_t is set base on x_t by (12) if it does not excess the capacity constraint, otherwise set $v_t = 0$.

Obtain a stochastic dual sub-gradient:

$$\tilde{g}_t = -x_t + \rho \, .$$

Update the dual variable using mirror descent:

$$\mu_{t+1} = \arg\min_{\mu \geq 0} \langle \tilde{g}_t, \mu \rangle + \frac{1}{\eta} V_h(\mu, \mu_t) \, .$$

end

B. Additional Literature Review

There is a stream of literature that studies online allocation problems with linear utility functions when the input is adversarial (Mehta et al., 2007a; Feldman et al., 2009). In this case, it is generally impossible to attain sublinear regret and, instead, the focus is on designing algorithms that obtain constant factor approximations to the offline optimum solution.

Our algorithms attain regret of order $O(\sqrt{T})$, which is tight under our minimal assumptions on the input (Lemma 1). Jasin (2015) studies linear allocation problems and shows that it is possible to attain $O(\log T)$ regret when the expected instance is non-degenerate. His algorithm periodically re-estimates the distribution of requests and computes a primal control by periodically solving a linear program with the re-estimated parameters. Li & Ye (2019) study linear allocation problems under the assumption that the distribution of requests is absolutely continuous with uniformly bounded densities. They present a dual algorithm that attains $O(\log T)$ regret. Their algorithm updates dual variables by solving a dual, linear program in each stage using all data collected so far. The assumptions of these two papers are essentially imposing that the *dual objective* is strongly convex at the optimal dual variables. In comparison, under our weaker assumptions, the dual objective cannot be guaranteed to be strongly convex, which leads to a $\Omega(\sqrt{T})$ lower bound on regret. Similar distinctions arise in online convex optimization where convexity vs. strong convexity of the *primal objective* functions determine whether $\Theta(\sqrt{T})$ vs. $\Theta(\log T)$ regret is attainable (see, e.g., Hazan et al. 2016).

Our work is also related to the literature on multi-arm bandits with knapsacks. Our feedback structure is stronger because we get to observe the reward function and consumption matrix before making a decision, while, in the bandit literature, these are revealed after making a decision. While algorithms for bandits with knapsacks are not directly applicable, our problem can be thought of as a *contextual* multi-arm bandit problem with knapsacks, where the context would correspond to the information of the request. The algorithms of Badanidiyuru et al. (2014) and Agrawal et al. (2016) can be applied to our setting after discretizing the context and action space. Discretization, however, leads to sub-optimal performance guarantees. In particular, the cardinality of the support of the action space does not appear in our regret bounds, while it must appear in the bandit setting since bandit algorithms need to explore the rewards for different actions. For example, in the problem of bidding in second-price auctions, contextual bandits algorithms with discretization lead to $O(T^{3/4})$ regret while our algorithms lead to $O(T^{1/2})$ regret.

C. Proofs of Proposition 1

Notice that for any $\mu \geq 0$, it holds that

$$\begin{aligned}
&\mathrm{OPT}(\mathcal{P}) \\
=&\mathbb{E}_{\mathcal{P}}\left[\begin{array}{ll} \max_{x_t \in \mathcal{X}} & \sum_{t=1}^T f_t(x_t) \\ \text{s.t.} & \sum_{t=1}^T b_t x_t \leq T\rho \end{array}\right] \\
\leq&\mathbb{E}_{\mathcal{P}}\left[\max_{x_t \in \mathcal{X}} \sum_{t=1}^T f_t(x_t) + T\mu^\top \rho - \mu^\top \sum_{t=1}^T b_t x_t\right] \\
=&T\mathbb{E}_{\mathcal{P}}\left[\max_{x \in \mathcal{X}} f(x) - \mu^\top b x + \mu^\top \rho\right] \\
=&T\left(\sum_{i=1}^n p_i \max_{x \in \mathcal{X}}\{f_i(x) - \mu^\top b_i x\} + \mu^\top \rho\right) \\
=&T\left(\sum_{i=1}^n p_i f_i^*(b_i^\top \mu) + \mu^\top \rho\right) ,
\end{aligned} \tag{13}$$

where the first inequality is because of the feasibility of x and $\mu \geq 0$ and the last equality is due to the definition of f_i^*. This finishes the proof. □

D. Proofs of Proposition 2

The key step in the proof of Proposition 2 is the following lemma, which shows that the dual update (8) never exceeds the upper bound $\mu^{\max}$ when the step-size η is small enough.

Lemma 2. *Let $\tilde{g} = b\nabla f^*(b^\top \mu) + \rho$ with $(b, f) \in \{(b_1, f_1), \ldots, (b_n, f_n)\}$, and $\mu^+ = \arg\min_{\hat{\mu} \geq 0}\langle \tilde{g}, \hat{\mu}\rangle + \frac{1}{\eta} V_h(\hat{\mu}, \mu)$. Suppose $\mu \leq \mu^{\max}$ and $\eta \leq \frac{\sigma_2}{\bar{b}}$, then it holds that $\mu^+ \leq \mu^{\max}$.*

Proof. Denote $J := \{j | \mu_j^+ > 0\}$, then we just need to show $\mu_j^+ \leq \mu_j^{\max}$ for any $j \in J$. Following the update rule (8), it holds for any $j \in J$ that

$$\dot{h}_j(\mu_j^+) = \dot{h}_j(\mu_j) - \eta \tilde{g}_j = \dot{h}_j(\mu_j) - \eta (b)_j^\top \nabla f^*(b^\top \mu) - \eta \rho_j . \tag{14}$$

Define $h_j^*(c) = \max_{\mu_j}\{c\mu_j - h_j(\mu_j)\}$ as the conjugate function of $h_j(\mu_j)$, then by the property of conjugate function it holds that $h_j^*(\cdot)$ is a $\frac{1}{\sigma_2}$-smooth univariate convex function. Furthermore, $\dot{h}_j^*(\cdot)$ is increasing, and $\dot{h}_j^*(\dot{h}_j(\mu_j)) = \mu_j$.

Now define $\tilde{x} := \arg\max_{x \in \mathcal{X}}\{f(x) - \mu^\top b x\} = -\nabla f^*(b^\top \mu)$. Then it holds that $0 = f(0) \leq f(\tilde{x}) - \mu^\top b\tilde{x} \leq \bar{f} - \mu^\top b\tilde{x}$, whereby $\mu^\top b\tilde{x} \leq \bar{f}$. Since $\mu \geq 0, b \geq 0, \tilde{x} \in \mathcal{X} \subseteq \mathbb{R}_+^d$, it holds for any $j \in J$ that $(b)_j^\top \tilde{x} \leq \frac{\bar{f}}{\mu_j}$. Meanwhile, it follows by the definition of $\bar{b}$ that $(b)_j^\top \tilde{x} \leq \bar{b}$. Together with (14), it holds that

$$\dot{h}_j(\mu_j^+) \leq \dot{h}_j(\mu_j) + \eta \min\left(\frac{\bar{f}}{\mu_j}, \bar{b}\right) - \eta\rho_j . \tag{15}$$

If $\frac{\bar{f}}{\rho_j} \leq \mu_j \leq \mu_j^{\max}$, we have $\min\left(\frac{\bar{f}}{\mu_j}, \bar{b}\right) - \rho_j \leq 0$, thus it holds that $\mu_j^+ \leq \mu_j \leq \mu_j^{\max}$ by utilizing (15) and convexity of $\dot{h}_j$. Otherwise, $\mu_j \leq \frac{\bar{f}}{\rho_j}$, and furthermore,

$$\begin{aligned}
\mu_j^+ &= \dot{h}_j^*(\dot{h}_j(\mu_j^+)) \leq \dot{h}_j^*(\dot{h}_j(\mu_j) + \eta\bar{b}) \\
&\leq \dot{h}_j^*(\dot{h}_j(\mu_j)) + \frac{\eta\bar{b}}{\sigma_2} \leq \frac{\bar{f}}{\rho_j} + 1 = \mu_j^{\max} ,
\end{aligned}$$

where the first inequality is from (15) and the monotonicity of $\dot{h}_j^*(\cdot)$, the second inequality is from $\dot{h}_j^*(\dot{h}_j(\mu_j)) = \mu_j$ and the $\frac{1}{\sigma_2}$-smoothness of $h_j^*(\cdot)$, the last inequality utilizes $\eta \leq \frac{\sigma_2}{\bar{b}}$, and the last equality follows from Definition 1. This finishes the proof of Lemma 2. □

Proof of Proposition 2: First, a direct application of Lemma 8 shows that for any t, $\mu_t \leq \mu^{\max}$. Next, it follows by the definition of τ_A (Definition 2) that there exist j such that $\sum_{t=1}^{\tau_A}(b_t)_j^\top x_t + \bar{b} \geq \rho_j T$. By the definition of $\tilde{g}_t$, we have

$$\sum_{t=1}^{\tau_A}(\tilde{g}_t)_j = \rho_j\tau_A - \sum_{t=1}^{\tau_A}(b_t)_j^\top x_t \leq \rho_j\tau_A - \rho_j T + \bar{b} ,$$

thus

$$T - \tau_A \leq \frac{\bar{b} - \sum_{t=1}^{\tau_A}(\tilde{g}_t)_j}{\rho_j} . \tag{16}$$

On the other hand, it follows the update rule (8) that for any $t \leq \tau_A$,

$$\dot{h}_j((\mu_{t+1})_j) \geq \dot{h}_j((\mu_t)_j) - \eta(\tilde{g}_t)_j .$$

Thus,

$$\begin{aligned}
\sum_{t=1}^{\tau_A} -(\tilde{g}_t)_j &\leq \frac{1}{\eta}\left(\dot{h}_j((\mu_{\tau_A+1})_j) - \dot{h}_j((\mu_0)_j)\right) \\
&\leq \frac{1}{\eta}\left(\dot{h}_j(\mu_j^{\max}) - \dot{h}_j((\mu_0)_j)\right) ,
\end{aligned} \tag{17}$$

where the last inequality is due to the monotocity of $\dot{h}_j(\cdot)$. Combining (16) and (17), we reach

$$T - \tau_A \leq \max_j\left\{\frac{\dot{h}_j(\mu_j^{\max}) - \dot{h}_j((\mu_0)_j)}{\eta\rho_j} + \frac{\bar{b}}{\rho_j}\right\} .$$

This finishes the proof by noticing that $\rho_j \geq \rho$ and $\dot{h}_j(\mu_j^{\max}) - \dot{h}_j((\mu_0)_j) \leq \|\nabla h(\mu^{\max}) - \nabla h(\mu_0)\|_\infty$. □

E. Proof of Proposition 3

Before proving Proposition 3, we first introduce some new notations which are used in the proof. By the definition of

conjugate function, we can rewrite the dual problem (7) as the following saddle-point problem:

$$(S): \quad \min_{0\le\mu}\max_{y\in p\mathcal{X}} L(y,\mu) := \sum_{i=1}^{n} p_i f_i(y_i/p_i) - \mu^\top By + \mu^\top \rho\,, \tag{18}$$

where $y := [y_1,\dots,y_n] \in \mathbb{R}^{nd}$, $B := [b_1;\dots;b_n] \in \mathbb{R}^{m\times nd}$ and $p\mathcal{X} := \{y | y_i \in p_i\mathcal{X}\} \subseteq \mathbb{R}_+^{nd}$. By minimizing over μ in (18), we obtain the following primal problem:

$$(P): \quad \max_{y} P(y) := \sum_{i=1}^{n} p_i f_i(y_i/p_i) \tag{19}$$

$$\text{s.t.} By \le \rho \tag{20}$$

$$y \in p\mathcal{X}\,. \tag{21}$$

The decision variable $y_i/p_i \in \mathcal{X}$ can be interpreted as the expected action to be taken when a request of type i arrives. Therefore, (P) can be interpreted as a deterministic optimization problem in which resource constraints can be satisfied in expectation. In the linear case, this problem is sometimes referred as the deterministic linear program (Talluri & van Ryzin, 1998) or the expected instance (Devanur et al., 2019). Moreover, we define an auxiliary primal variable sequence $\{z_t\}_{t=1,\dots,T}$:

$$z_t = \arg\max_{z\in p\mathcal{X}} L(z,\mu_t)\,. \tag{22}$$

As a direct consequence of (18) and (22), we obtain:

$$g_t := -Bz_t + \rho = \nabla_\mu L(z_t,\mu_t) \in \partial_\mu D(\mu_t)\,. \tag{23}$$

Proof of Proposition 3. It follows by the definition of $\tilde{g}_t$, $\bar{b}$ and $\bar{\rho}$ that

$$\mathbb{E}_{\gamma_t}\|\tilde{g}_t\|_\infty^2 \le 2\left(\mathbb{E}_{\gamma_t}\|b_t x_t\|_\infty^2 + \|\rho\|_\infty^2\right) \le 2\left(\bar{b}^2+\bar{\rho}^2\right)\,. \tag{24}$$

Note that $\mu_t \in \sigma(\xi_{t-1})$, $g_t \in \sigma(\xi_{t-1})$, and $\tilde{g}_t \in \sigma(\xi_t)$, where $\sigma(X)$ denotes the sigma algebra generated by a stochastic process X. Notice $\mathbb{E}_{\gamma_t}\tilde{g}_t = g_t$, thus it holds for any $\mu \in \mathcal{D}$ that

$$\begin{aligned}
&\langle g_t, \mu_t - \mu\rangle \\
=&\langle \mathbb{E}_{\gamma_t}[\tilde{g}_t|\mu_t], \mu_t - \mu\rangle \\
\le&\mathbb{E}_{\gamma_t}\Big[\langle \tilde{g}_t, \mu_t - \mu_{t+1}\rangle + \frac{1}{\eta}V_h(\mu,\mu_t) \\
&\qquad - \frac{1}{\eta}V_h(\mu,\mu_{t+1}) - \frac{1}{\eta}V_h(\mu_{t+1},\mu_t)|\mu_t\Big] \\
\le&\mathbb{E}_{\gamma_t}\Big[\langle \tilde{g}_t, \mu_t - \mu_{t+1}\rangle + \frac{1}{\eta}V_h(\mu,\mu_t) \\
&\qquad - \frac{1}{\eta}V_h(\mu,\mu_{t+1}) - \frac{\sigma_1}{2\eta}\|\mu_{t+1}-\mu_t\|_1^2|\mu_t\Big] \\
\le&\mathbb{E}_{\gamma_t}\left[\frac{\eta}{\sigma_1}\|\tilde{g}_t\|_\infty^2 + \frac{1}{\eta}V_h(\mu,\mu_t) - \frac{1}{\eta}V_h(\mu,\mu_{t+1})|\mu_t\right] \\
\le&\frac{2\eta}{\sigma_1}\left(\bar{b}^2+\bar{\rho}^2\right) + \frac{1}{\eta}V_h(\mu,\mu_t) - \mathbb{E}_{\gamma_t}\left[\frac{1}{\eta}V_h(\mu,\mu_{t+1})|\mu_t\right]\,,
\end{aligned} \tag{25}$$

where the first inequality follows from Three-Point Property stated in Lemma 3.2 of Chen & Teboulle (1993), the second inequality is by strongly convexity of h, the third inequality uses that $a^2+b^2 \ge 2ab$ for $a,b\in\mathbb{R}$ and Cauchy-Schwarz to obtain

$$\begin{aligned}
\frac{\sigma_1}{2\eta}\|\mu_{t+1}-\mu_t\|_1^2 + \frac{\eta}{\sigma_1}\|\tilde{g}_t\|_\infty^2 &\ge \|\mu_{t+1}-\mu_t\|_1\|\tilde{g}_t\|_\infty \\
&\ge |\langle \tilde{g}_t, \mu_t - \mu_{t+1}\rangle|\,,
\end{aligned}$$

and the last inequality follows from (24). Taking expectation with respect to ξ_{t-1} and multiplying by η on both sides of (25) yields:

$$\begin{aligned}
&\mathbb{E}_{\xi_{t-1}}[\eta\langle g_t, \mu_t - \mu\rangle] \\
&\le \frac{2\left(\bar{b}^2+\bar{\rho}^2\right)}{\sigma_1}\eta^2 + \mathbb{E}_{\xi_{t-1}}[V_h(\mu,\mu_t)] - \mathbb{E}_{\xi_t}[V_h(\mu,\mu_{t+1})]\,.
\end{aligned} \tag{26}$$

Consider the process $M_t = \sum_{s=1}^t \eta\langle g_s, \mu_s - \mu\rangle - \mathbb{E}_{\xi_{s-1}}[\eta\langle g_s, \mu_s - \mu\rangle]$, which is martingale with respect to ξ_t (i.e., $M_t \in \sigma(\xi_t)$ and $\mathbb{E}[M_{t+1}|\xi_t] = M_t$) with increments bounded by

$$\begin{aligned}
|M_t - M_{t-1}| &\le \eta(\|g_t\|_\infty + \mathbb{E}_{\xi_{t-1}}\|g_t\|_\infty)\|\mu_t - \mu\|_1 \\
&\le 2(\bar{b}+\bar{\rho})m\|\mu_t - \mu\|_\infty \\
&\le 4m(\bar{b}+\bar{\rho})\|\mu^{\max}\|_\infty \\
&= 4m(\bar{b}+\bar{\rho})\left(\frac{\bar{f}}{\underline{\rho}}+1\right) < \infty\,,
\end{aligned}$$

where the first inequality is Cauchy-Schwarz, the second inequality is from $\|g_t\|_\infty \le \bar{b}+\bar{\rho}$ almost surely, and the last inequality uses $\mu_t \in \mathcal{D}$ by Lemma 2. Since τ_A is a stopping time with respect to ξ_t and τ_A is bounded, the Optional Stop-

ping Theorem implies that $\mathbb{E}[M_{\tau_A}] = 0$. Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{\tau_A} \eta\langle g_t, \mu_t - \mu\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{\tau_A} \mathbb{E}_{\xi_{t-1}}[\eta\langle g_t, \mu_t - \mu\rangle]\right] \leq \frac{2\left(\bar{b}^2 + \bar{\rho}^2\right)}{\sigma_1}\eta^2\mathbb{E}[\tau_A] + V_h(\mu, \mu_0)\,. \tag{27}$$

where the inequality follows from summing up (26) from $t = 1$ to $t = \tau_A$, telescoping, and using that the Bregman divergence is non-negative.

On the other hand, it holds that by choosing $\mu = 0$

$$\begin{aligned}
&\sum_{t=1}^{\tau_A} \eta\langle g_t, \mu_t - \mu\rangle \\
=&\sum_{t=1}^{\tau_A} \eta\langle \nabla_\mu L(z_t, \mu_t), \mu_t - \mu\rangle \\
=&\sum_{t=1}^{\tau_A} \eta\left(L(z_t, \mu_t) - L(z_t, \mu)\right) \\
=&\sum_{t=1}^{\tau_A} \eta\left(L(z_t, \mu_t) - P(z_t) - \mu(\rho - Bz_t)\right) \\
=&\sum_{t=1}^{\tau_A} \eta\left(D(\mu_t) - P(z_t) - \mu(\rho - Bz_t)\right) \\
\geq&\tau_A\eta\left(D(\bar{\mu}_{\tau_A}) - \frac{\sum_{t=1}^{\tau_A} P(z_t)}{\tau_A}\right) - \sum_{t=1}^{\tau_A}\mu(\rho - Bz_t) \\
=&\tau_A\eta\left(D(\bar{\mu}_{\tau_A}) - \frac{\sum_{t=1}^{\tau_A} P(z_t)}{\tau_A}\right)\,,
\end{aligned} \tag{28}$$

where the first equality uses (23), the second equality is because $L(z, \mu)$ is linear in μ, the third equality is from $z_t = \arg\min_z L(z, \mu_t)$, the first inequality uses convexity of $D(\cdot)$ over μ, and the last equality is because $\mu = 0$. Combining (27) and (28) and choosing $\mu = 0$, we obtain:

$$\begin{aligned}
&\mathbb{E}\left[\tau_A D(\bar{\mu}_{\tau_A}) - \sum_{t=1}^{\tau_A} P(z_t)\right] \\
&\leq \frac{2\left(\bar{b}^2 + \bar{\rho}^2\right)}{\sigma_1}\eta\mathbb{E}[\tau_A] + \frac{V_h(\mu, \mu_0)}{\eta}\,.
\end{aligned} \tag{29}$$

Notice that μ_t and z_t are measurable given the sigma algebra $\sigma(\xi_{t-1})$. From the update of x_t and z_t, we know that if a request of type i-th is realized in the t-th iteration, then $x_t = (z_t)_i/p_i$. Thus it holds for any $t \leq \tau_A$ that

$$\mathbb{E}_{\gamma_t}[f_t(x_t)|\xi_{t-1}] = \sum_{i=1}^{n} p_i f_i((z_t)_i/p_i) = P(z_t)\,.$$

Therefore, another martingale argument yields that

$$\mathbb{E}\left[\sum_{t=1}^{\tau_A} f_t(x_t)\right] = \mathbb{E}\left[\sum_{t=1}^{\tau_A} P(z_t)\right]\,. \tag{30}$$

Combining (29) and (30) finishes the proof. □

F. Proof of Theorem 1

Proof of Theorem 1. For any $\mathcal{P} \in \mathcal{I}$, we have for any τ_A that

$$\begin{aligned}
\mathrm{OPT}(\mathcal{P}) &= \frac{\tau_A}{T}\mathrm{OPT}(\mathcal{P}) + \frac{T - \tau_A}{T}\mathrm{OPT}(\mathcal{P}) \\
&\leq \tau_A D(\bar{\mu}_{\tau_A}) + (T - \tau_A)\,\bar{f}\,,
\end{aligned}$$

where the inequality uses (6) and the fact that $\mathrm{OPT}(\mathcal{P}) \leq \bar{f}$. Therefore,

$$\begin{aligned}
&\mathrm{Regret}(A|\mathcal{P}) \\
=&\mathrm{OPT}(\mathcal{P}) - R(A|\mathcal{P}) \\
\leq&\mathbb{E}_{\mathcal{P}}\left[\tau_A D(\bar{\mu}_{\tau_A}) + (T - \tau_A)\bar{f} - \sum_{t=1}^{T} f_t(x_t)\right] \\
\leq&\mathbb{E}_{\mathcal{P}}\left[\left(\tau_A D(\bar{\mu}_{\tau_A}) - \sum_{t=1}^{\tau_A} f_t(x_t)\right)\right] \\
&+ \mathbb{E}_{\mathcal{P}}\left[(T - \tau_A)\bar{f}\right] \\
\leq&\frac{2\left(\bar{b}^2 + \bar{\rho}^2\right)}{\sigma_1}\eta\mathbb{E}_{\mathcal{P}}[\tau_A] + \frac{V_h(0, \mu_0)}{\eta} \\
&+ \frac{\bar{f}}{\underline{\rho}\eta}\|\nabla h(\mu^{\max}) - \nabla h(\mu_0)\|_\infty + \frac{\bar{f}\bar{b}}{\underline{\rho}} \\
\leq&\frac{2\left(\bar{b}^2 + \bar{\rho}^2\right)}{\sigma_1}\eta T + \frac{V_h(0, \mu_0)}{\eta} \\
&+ \frac{\bar{f}}{\underline{\rho}\eta}\|\nabla h(\mu^{\max}) - \nabla h(\mu_0)\|_\infty + \frac{\bar{f}\bar{b}}{\underline{\rho}}\,,
\end{aligned} \tag{31}$$

where the second inequality is because $\tau_A \leq T$ and $f_t(x_t) \geq 0$, the third inequality uses Proposition 2 and Proposition 3, and the last inequality is from $\tau_A \leq T$ almost surely. Moreover, (31) holds for any $\mathcal{P} \in \mathcal{I}$, which finishes the proof of Theorem 1. □

G. Proof of Proposition 4

Proof of Proposition 4. The proof essentially follows exactly from the proof of Theorem 1 after taking the expectation on ζ. Notice that $\tilde{g}_t$ does not depend on the realization of ζ, thus Proposition 3 still holds. The only part requiring major modification in the proof is the bound on stopping time (i.e., Proposition 2). Actually (10) no longer holds almost surely, but it still holds in expectation, which is enough to show Proposition 4. The difficulty in establishing this result is that the stopping time is now defined in term of the realized allocation v_t instead of the action x_t. For the consistency of the proof, we still use the notations for the general problem (1) herein. Our goal is now to show that

$$\mathbb{E}_\zeta\left[T-\tau_A\right] \leq \max_j \left\{ \frac{\dot{h}_j(\mu_j^{\max}) - \dot{h}_j((\mu_0)_j)}{\eta\rho_j} + \frac{\bar{b}}{\rho_j} \right\} . \tag{32}$$

Consider the sigma algebra $\mathcal{F}_t = \sigma(\mathcal{H}_t, f_{t+1}, b_{t+1})$ where $\mathcal{H}_t = \{f_s, b_s, v_s\}_{s=1}^t$ is the previous history. Note that $x_{t+1} \in \mathcal{F}_t$ but $v_{t+1} \in \mathcal{F}_{t+1}$. At first, it holds that $M_t := \sum_{s=1}^t b_s(v_s - x_s)$ is a martingale with respect to $\mathcal{F}_t$, because $M_t \in \mathcal{F}_t$, $\mathbb{E}_\zeta(\|M_t\|) \leq 2\bar{b}t$ is bounded for any t, and

$$\mathbb{E}_\zeta\left[M_{t+1} - M_t | \mathcal{F}_t\right] = b_{t+1}\mathbb{E}\left[v_{t+1} - x_{t+1} | x_{t+1}\right] = 0 .$$

Recall that τ_A is the first time that

$$\sum_{t=0}^{\tau_A} (b_t)_j^\top v_t + \bar{b} \geq \rho_j T . \tag{33}$$

Notice the $(b_t)_j^\top v_t$ in the left-hand-side of (33) is measurable with respect to $\mathcal{F}_t$, thus τ_A is a stopping time with respect to $\mathcal{F}_t$. It then follows by Martingale Optional Stopping Theorem that $\mathbb{E}_\zeta[M_{\tau_A}] = \mathbb{E}_\zeta[M_1] = 0$. Therefore, because $\tilde{g}_t = -x_t + \rho$ we obtain

$$\begin{aligned}
\mathbb{E}_\zeta\left[\sum_{t=1}^{\tau_A} (\tilde{g}_t)_j\right] &= \mathbb{E}_\zeta\left[\rho_j\tau_A - \sum_{t=1}^{\tau_A} (b_t)_j^\top x_t\right] \\
&= \mathbb{E}_\zeta\left[\rho_j\tau_A - \sum_{t=1}^{\tau_A} (b_t)_j^\top v_t\right] \\
&\leq \mathbb{E}_\zeta\left[\rho_j\tau_A - \rho_j T + \bar{b}\right] ,
\end{aligned}$$

where the second equality is from $\mathbb{E}_\zeta[M_{\tau_A}] = 0$ and the inequality from (33). Thus

$$\mathbb{E}_\zeta\left[T - \tau_A\right] \leq \mathbb{E}_\zeta\left[\frac{\bar{b} - \sum_{t=1}^{\tau_A}(\tilde{g}_t)_j}{\rho_j}\right] .$$

Notice that $\tilde{g}_t$ does not depend on the realized allocation ξ_t, thus Lemma 2 and (17) still holds, which finishes the proof of (32). Proposition 4 can be then proved by following the exact steps in the proof of Theorem 1 after taking an additional expectation over ζ. □

H. Additional details on the numerical experiments

Here we present additional details in the numerical experiments.

Data generation: We use the dataset introduced by Balseiro et al. (2014). They consider the problem faced by a publisher who has to deliver impressions to advertisers so as to maximize click-through rates. (They consider the secondary objective of maximizing revenue from a spot market, which we do not take into account for this experiments). We incorporate the entropy regularizer $H(x)$ to the objective with parameter $\lambda = 0.0002$, which was tuned to balance diversity and efficiency of the allocation. In each problem instance there are m advertisers; advertiser j can be assigned at most $\rho_j T$ impressions. The revenue vector r_t gives the expected click-through rate of assigning the impression to each advertiser. In their paper, they parametrically estimate click-through rates using mixtures of log-normal distributions. Because they do not report the actual data used to estimate their model, we instead take their estimate model as a generative model and sample impressions from the distributions provided in their paper. We generated $100,000$ samples for each publisher, and we present results for publisher 2 from their dataset, which has 12 advertisers.

Random trials: There are two layers of randomness in Algorithm 3: randomness coming from the data (i.e., $\mathcal{P}$), and randomness coming from the proportional matching (i.e., ζ). In the numerical experiments, we first obtain 20 random datasets with size T uniformly randomly chosen from $100,000$ samples (for the first layer of randomness), and for each dataset, we run Algorithm 3 20 times (for the second layer of randomness). In total, we run 400 random trials, and report the average regret with 95% confidence interval in Figure 1.

Regret and relative reward computation: For each random trial with given round T, we compute the cumulative revenue obtained by Algorithm 3. Once an advertiser does not have any remaining budget, we rule it out from the future allocation. We then compute the average cumulative revenue over the 400 trials as our expected revenue of Algorithm 3, i.e., $R(A|\mathcal{P})$. We compute OPT by solving the offline problem (1) with $100,000$ samples. We report the following regret: $\text{Regret}(A|\mathcal{P}) = T/100000 \times \text{OPT} - R(A|\mathcal{P})$ and its 95% confidence interval any different value of T in Figure 1. We report the following relative reward: $R(A|\mathcal{P})/(T/100000 \times \text{OPT})$ in Figure 2.