Fair Allocation in Online Markets

Sreenivas Gollapudi
Microsoft Research Search Labs
Mountain View, CA 94043
sreenig@microsoft.com

Debmalya Panigrahi*
Duke University
Durham, NC 27708
debmalya@cs.duke.edu

ABSTRACT

A key characteristic of a successful online market is the large participation of agents (producers and consumers) on both sides of the market. While there has been a long line of impressive work on understanding such markets in terms of revenue maximizing (also called max-sum) objectives, particularly in the context of allocating online impressions to interested advertisers, fairness considerations have surprisingly not received much attention in online allocation algorithms. Allocations that are inherently fair to participating entities, we believe, will contribute significantly to retaining current participants and attracting new ones in the long run, thereby enhancing the performance of online markets. We give two generic online allocation algorithms to address this problem: the first algorithm only focuses on the max-min fairness objective whereas the second algorithm balances between the max-min and the max-sum objectives. We compare these algorithms with existing online allocation algorithms that are based on only the max-sum objective by performing experiments on real-world data in two different online markets. Moreover, we analytically show that both our algorithms have strong theoretical guarantees for their respective objectives.

1. INTRODUCTION

As the internet achieves ubiquity and web-based services become an indispensable component of modern life, online systems have increasingly relied on efficient algorithms for allocating scarce resources to competing entities. While this problem has had a long history of algorithmic research which predates the modern-day internet by several decades, the emergence of the web has led to vigorous renewed activity, particularly in the context of online applications where the participating entities are substantially more dynamic and unpredictable than in traditional offline settings such as job scheduling. We describe such a setting as an *online market* where producers must be matched with consumers to meet specific demands that arrive over time. While this generic definition captures the essence of various real-world allocation problems, we describe some key applications in the domain of internet technologies that motivated our study.

- Our first application comes from the field of internet advertising. The central question in this domain is to assign advertising slots on webpages (impressions) that are generated by users dynamically to interested advertisers (bidders) who are constrained by their respective advertising budgets. This problem has been extensively studied recently (as evidenced by the robust literature some of which we will survey later) with the aim of maximizing revenue. However, natural logic suggests that allocation algorithms should ensure a certain degree of *fairness* to the advertisers as well — algorithms that are revenue-maximizing in the short run but are not fair to all the participating entities are likely to have negative long-term impact on revenue as participants start leaving the competition.
- For our second application, we switch our focus to the other major economic activity on the internet: e-commerce. Consider an e-commerce website that wishes to show a set of relevant products from multiple sellers in response to a user query. The set of products displayed and the corresponding sellers chosen should not only be aimed at maximizing the expected revenue that they will generate due to sales, but must also respect fairness considerations to ensure that sellers do not leave the portal in the long run.
- The emerging phenomenon of data and information markets provide yet another area where allocation problems have an inherent fairness objective. In these markets, user queries that seek some specific information or dataset are redirected to one among a set of providers who are capable of providing it. As in the previous applications, it is important to ensure that no providers is starved of requests, i.e., that the algorithm is fair to all the participating providers.
- Fairness considerations can also be driven by a need to balance requests so that individual providers are not overloaded. This is particularly relevant for applications such as cloud computing where online requests for resources and services must be distributed evenly (fairly) among available resource providers such as servers [?, ?, ?]. Similar considerations also demand

*Part of this work was done when the author was at Microsoft Research, Redmond.

a fair allocation of tasks among participants in applications such as crowdsourcing.

- Finally, the need for fairness might also be driven by other considerations such as legal ones. For example, advertisers frequently enter into agreements with ad exchanges guaranteeing them a minimum number of impressions per day. Allocation algorithms such as those in this paper that respect specified fairness goals are a tool to honoring such agreements.

In essence, any service that arbitrates among a set of available choices in order to satisfy a demand that arrives online must meet the dual objectives of revenue and fairness in its arbitration protocol. In this paper, we consider the following general setting. A set of *providers* (such as advertisers, data providers, etc.), each constrained by a *budget*, are given offline. A set of *requests* (such as advertising slots on webpages, data queries, etc.) arrive online. Each request can be allocated to one among a subset of providers, and such an allocation generates a specified amount of *revenue* subject to the constraint that the total revenue for a provider cannot exceed her budget. In the revenue-maximizing setting, the goal of the allocation algorithm is to maximize the revenue. However, to address the fairness objective, we introduce a new offline parameter called the *target* for every provider. Notionally, this represents the *minimum* revenue for a provider that gives her sufficient incentive to stay in the system. The online allocation algorithm must now ensure that each provider meets her target, even if that comes at the cost of decreased revenue. Of course, among the allocations that would meet the targets, the algorithm will continue to favor ones that generate greater revenue.

1.1 Contributions

Our main contributions in this paper can be summarized as follows.

1. We formalize the generic online allocation problem and give three objectives that represent the trade-off between the max-sum and max-min objectives. The first problem (called MAX-MIN) aims to maximize the minimum revenue (scaled by revenue targets) among all providers, the second problem (called MAX-SUM) aims to maximize the sum of revenues (capped by revenue budgets) of providers, and the third problem (called HYBRID) incorporates both max-min and max-sum objectives by aiming to maximize the sum of revenues (again, capped by revenue budgets) of providers but pays a penalty if it misses the revenue target of any provider.

2. We give efficient algorithms for the MAX-MIN and HYBRID problems, and recall an algorithm due to Devanur and Hayes [?] for the MAX-SUM problem. We show analytically that all these algorithms are nearly optimal for their respective objectives.

3. Finally, we perform large-scale experiments on real-world data to compare the MAX-MIN, MAX-SUM, and HYBRID algorithms on revenue (max-sum) and fairness (max-min) metrics. Specifically, we considered two natural real-world settings, *viz.*, online ad allocations and online information markets. For the first setting, we observe that MAX-SUM and HYBRID behave similarly, and are comprehensively outperformed by MAX-MIN in both the objectives for smaller targets/budgets. For larger targets/budgets, while MAX-MIN performs better on the fairness metric, MAX-SUM and HYBRID win the contest on the revenue metric. In the second setting of information markets, the performance of HYBRID is closer to MAX-MIN than MAX-SUM, and HYBRID and MAX-MIN consistently outperform MAX-SUM on both objectives, particularly for smaller budgets/targets.

1.2 Related Work

There has been substantial research on understanding the trade-off between fairness and other objectives, particularly maximization objectives in resource allocation such as bandwidth maximization in networks (see, e.g., [?, ?]), maximizing throughput and utilization of computing resources in scheduling (see, e.g., [?, ?]), etc. There has been some recent work on studying the trade-off between fairness and efficiency in display ad systems both for web applications as well other domains such as TV ads [?, ?, ?]. Of these, the closest to our work is [?], where the authors perform an experimental evaluation of the trade-off between efficiency and fairness. Instead, we give a concrete HYBRID formulation that simultaneously achieves fairness and efficiency, and in addition to experiments, we analytically show that our HYBRID algorithm is almost optimal for its objective function. Our HYBRID algorithm is based on a dual-based training algorithm for the MAX-SUM problem due to Devanur and Hayes [?]. There has been substantial research following up on this work applying similar technique to a more general suite of allocation problems and obtaining less restrictive versions of the so called "large budgets" condition [?, ?, ?, ?, ?]. The MAX-SUM objective has been studied extensively over the last few years, particularly in the context of online ad allocation, both in the adversarial input model [?, ?] and also in stochastic input models such as the one considered in this paper [?, ?] (see also the recent survey by Mehta [?] and references contained therein). In addition to the HYBRID and MAX-SUM objectives, we also have a formulation (the MAX-MIN problem) that solely focuses on the fairness objective. This formulation is similar to the diversification objective in [?], but whereas the goal in their work was to study the online selection problem, we study the online allocation problem.

1.3 Problem Formulation

As described in the introduction, we will formulate three versions of the resource allocation problem that correspond to the max-min, the max-sum, and a hybrid objective. Let D be a set of n *providers*. Let B_i and T_i be the *budget* and *target* respectively for provider $i \in D$, where $T_i \leq B_i$. Intuitively, the budget and target of a provider represent her maximum and minimum revenue respectively.

An input set Q of m requests arrives over time (online), where each request $j \in Q$ has a *bid value* b_{ij} corresponding to each provider $i \in D$. The bid value represents the revenue earned if the query j is assigned to provider i, subject to the budgetary limits of provider i. (Bid values are non-negative but can be 0.) On the arrival of request j, the algorithm allocates it to a provider (denoted $d(j)$). Let $Q(i)$ be the set of requests that are allocated to provider i. The *unrestricted* *revenue* R_i corresponding to provider $i \in D$ is given by $\sum_{j \in Q(i)} b_{ij}$, while the *budgeted revenue* (or simply *revenue*) P_i corresponding to provider $i \in D$ is given by $\min(R_i, B_i)$. The *fractional coverage* c_i for provider $i \in D$ is given by the ratio $\frac{P_i}{T_i}$.

The three versions of this problem differ in the objectives that they seek to optimize.

- In the Max-Min problem, the goal is to maximize the minimum fractional coverage, i.e., $\min_{i \in D} c_i$.
- In the Max-Sum problem, the goal is to maximize the total (budgeted) revenue, i.e., $\sum_{i \in D} P_i$.
- In the Hybrid problem, we seek to maximize the total revenue as in the Max-Sum problem but we have a penalty for any provider whose revenue is less than its target. In particular, we have a *penalty coefficient* α and the goal is to maximize $\sum_{i \in D}(P_i - \alpha \max(T_i - P_i, 0))$.

We will analytically measure the quality of our algorithms by their *competitive ratio*, which is the minimum (over all input sequences) ratio of the objective of the algorithmic solution to that of the (offline) optimal solution. In the online setting, the main algorithmic challenge is to provision for the future without knowing the future input. However, in practical scenarios such as internet advertising, the impressions arriving over time have a relatively consistent pattern though they may be subject to temporary fluctuations. We formally model this by assuming that the requests are independently and identically distributed (i.i.d.) according to some probability distribution that is not known to the algorithm. Note that the variance in the distribution automatically produces a certain degree of input fluctuation; however, the mere fact that the input must be drawn from some fixed (but unknown) distribution ensures that a certain level of consistency in the long run, which our algorithms exploit. It is important to note that the support of this distribution can be extremely large (e.g., exponential in the number of providers), and therefore, we cannot hope for strong concentration bounds in the actual bid values. However, the marginals of this distribution on the advertisers must be highly concentrated if the number of requests is large compared to the number of providers, which is typically the case. Roughly speaking, this corresponds to the intuitive claim that the average (over large enough time intervals) of the bid values of a provider for the requests that arrive online remains fixed (or is subject to relatively small variations). This allows the following generic technique: initially, estimate the marginal distribution of bid values for the providers from a small constant fraction of requests (say 1%) and then use these marginals to guide allocation decisions in the future.

2. THE Max-Min PROBLEM

Recall that in the Max-Min problem, the goal is to maximize the minimum fractional coverage over all the providers, where the coverage of a provider is the ratio of her revenue to her target. For this problem, we will describe a simple and efficient algorithm (we call it the Max-Min algorithm) and theoretically verify that it is obtains a nearly optimal solution. Before describing the algorithm formally, let us give some intuition behind it.

First, note that in this problem, we can replace actual revenue (capped by budgets) by the corresponding unrestricted (i.e., uncapped) revenue for each provider. This is without loss of generality (w.l.o.g). Consider an algorithm that obtains a minimum coverage of c. If $c \geq 1$, all targets are attained and the solution is optimal. So, we focus on $c < 1$. In this case, a provider i attains a revenue of at least $c \cdot T_i$. But, note that $T_i \leq B_i$; thus,

$$c \cdot T_i \leq c \cdot B_i \leq B_i.$$

Hence, the (capped) actual revenue attained by provider i is also at least $c \cdot T_i$. Therefore, we ignore budgets in the remainder of this section (for the Max-Min problem).

Perhaps the most obvious algorithm for the Max-Min problem is one that greedily assigns each request to the provider who bids the maximum for it. However, this algorithm can be counter-productive in some scenarios, e.g., if there is a provider who bids large values but has a relatively small target and budget. In fact, consider the following scenario. Suppose at some stage of the input, the algorithm has satisfied the targets of all except one provider. Clearly, in all subsequent allocations, the algorithm must attempt to allocate the arriving request to this lone provider since such an allocation would increase the minimum coverage. More generally, this suggests that the first few units of revenue generated by a provider are more important than the latter units of revenue. To capture this intuition, we introduce a reward function that encodes the *importance* of a unit of revenue based on the current revenue of a provider. The reward function decreases as the revenue of a provider increases. The reward earned by a particular allocation is the cumulative value of the reward function earned by the provider over the interval corresponding to her previous revenue to her new revenue. The algorithm simply assigns an arriving request to the provider who earns the maximum reward for it. Note that the naïve greedy algorithm described above is simply a version of our algorithm where the reward function is uniform over the entire range of revenue.

Let us now formally define algorithm Max-Min. Let the expected optimal value of the objective function be denoted by c_{OPT}. We will assume that the algorithm knows (or has a good estimate of) the value of c_{OPT}. This assumption can be removed in exchange for a small loss in the competitive ratio of the algorithm. As described above, our algorithm uses a reward function ϕ defined as

$$\phi(k) = \left(\frac{\alpha \ln n}{c_{\text{OPT}}}\right) \exp\left(-\alpha \cdot \frac{k}{c_{\text{OPT}}} \cdot \ln n\right),$$

where α is a constant that we will fix later. We also define

$$\bar{\Phi}(k) = \int_{j=k}^{\infty} \phi(j)\ dj.$$

At any stage of the algorithm, the remaining reward for provider i is $\bar{\Phi}_i = \bar{\Phi}(c_i)$, and the overall remaining reward is $\bar{\Phi} = \sum_{i \in D} \bar{\Phi}_i$.

Let j be the current request. If the current allocation has fractional coverage c_i for some provider $i \in D_j$, then the reward r_{ij} of allocating request j to provider i is defined as the decrease in the value of $\bar{\phi}$ if request j is allocated to provider i, i.e.,

$$r_{ij} = \int_{k=c_i}^{c_i + b_{ij}/T_i} \phi(k)$$

The MAX-MIN algorithm allocates the current request j to the provider $i \in D_j$ that maximizes r_{ij}.

2.1 Analysis of the MAX-MIN Algorithm

Assumption. Suppose that for some $\epsilon > 0$, we have the property

$$\min_{i \in D} T_i \geq \frac{\max_{\xi, i \in D} b_i(\xi)}{2\epsilon n},$$

where ξ indexes the support of the probability distribution from which are requests are drawn. Intuitively, this assumption ensures that none of the targets are too small compared to the maximum bid value. Note that this is true in practice in typical applications; e.g., in internet advertising, typical bids are in the range of a few cents, whereas advertising targets are millions of dollars.

With this assumption, we will now prove the next theorem.

THEOREM 1. *The competitive ratio of the above algorithm for the* MAX-MIN *problem is* $1 - \epsilon$.

Before giving a formal proof, let us sketch the main ideas that we will use in our analysis. Let $\rho_{\text{OPT}} = c_{\text{OPT}} \min_{i \in D} T_i$. We will assume that

$$\rho_{\text{OPT}} \geq \frac{\beta\alpha \cdot \ln n \cdot (\max_{\xi, i \in D} b_i(\xi))}{2} \tag{1}$$

for some constant β that we will fix later. We will show later that using a carefully chosen value of β, the above assumption on ρ_{OPT} holds as a consequence of our original assumption on the value of $\min_{i \in D} T_i$. Therefore, we are not introducing any new assumptions here.

Our main technical tool is the following lemma, which lower bounds the expected decrease in the value of the potential in every step of the algorithm.

LEMMA 2. *At any stage of the algorithm, the expected (over the input) decrease in* $\bar{\Phi}$ *for the next item in the input stream is at least* $(1 - 1/\beta) \frac{\alpha \ln n}{m} \bar{\Phi}$.

We will prove this lemma shortly, but first, let us show how this lemma leads to Theorem 1. The above lemma implies that the value of $\bar{\Phi}$ decreases to at most

$$\left(1 - \left(1 - \frac{1}{\beta}\right) \frac{\alpha \ln n}{m}\right) \cdot \bar{\Phi} \leq n^{-(1-1/\beta)\frac{\alpha}{m}} \cdot \bar{\Phi},$$

since $1 - x \leq e^{-x}$. Using this repeatedly over the m requests, and observing that the initial value of the potential is n, we can upper bound the expected value of the potential at the end of the algorithm.

COROLLARY 3. *The expected value of* $\bar{\Phi}$ *when the algorithm terminates is at most* $n^{1-(1-1/\beta)\alpha}$.

We set $\alpha = \frac{c_{\text{OPT}}}{n \ln n}$ and $\beta = 1/\epsilon$, which satisfies Eqn. (1) subject to our original assumption. Now, we are ready to complete the proof of Theorem 1.

PROOF OF THEOREM 1. Suppose not, and let $i_{\min}$ be the provider with the minimum fractional coverage at the end of the algorithm. Then,

$$\bar{\Phi} \geq \bar{\Phi}_{i_{\min}} > \frac{c_{\text{OPT}}}{\alpha \ln n} \cdot n^{-\alpha(1-\epsilon)} = n^{1-(1-1/\beta)\alpha},$$

where the last equation follows from the choice of α and β. This violates Corollary 3. □

$$\begin{aligned}
\sum_{\xi} p(\xi) \cdot w_i(\xi) \cdot b_i(\xi) &\geq \frac{c_{\text{OPT}} T_i}{m} && \forall\, i \in D \\
\sum_{i \in S} w_i(\xi) &= 1 && \forall\, i \in D, \xi \\
0 \leq w_i(\xi) &\leq 1 && \forall\, i \in D, \xi
\end{aligned}$$

Figure 1: LP relaxation of the MAX-MIN problem.

We are left with proving Lemma 2. To prove this lemma, we use a linear programming (LP) relaxation of the MAX-MIN problem (Figure 1). Here, ξ indexes the support of the distribution from which requests are drawn, $p(\xi)$ denotes the probability of any request in the input stream having type ξ, and $w_i(\xi)$ is the fraction to which such a request is allocated to provider $i \in D$. The first constraint asserts that the expected revenue for any provider is at least $c_{\text{OPT}} \cdot T_i$ and the second constraint requires that each request be assigned to exactly one provider (in the fractional version, the sum of the fractional assignments for a provider is exactly 1). Since the expected optimal value of the objective is c_{OPT}, the optimal solution is feasible for this LP. It is important to note that we are using this LP only for analysis; the algorithm cannot use this LP since is does not know the distribution from which the input is drawn. Moreover, this LP can be exponential in size! We will compare the expected decrease of potential in our algorithm to that by a hypothetical (fractional) algorithm (we call it the LP-based algorithm) that exactly follows the assignment given by the optimal fractional solution for the above LP. Since our algorithm maximizes the decrease in potential in any step, any lower bound on the expected decrease of potential for the LP-based algorithm is also a lower bound for our algorithm. We can show such a lower bound of $(1 - 1/\beta) \frac{\alpha \ln n}{m} \bar{\Phi}$, thereby proving Lemma 2. The details of the proof are in the appendix.

3. THE MAX-SUM PROBLEM

Recall that in the MAX-SUM problem, the goal is to maximize the revenue subject to budgets, and there are no targets. The algorithm for the MAX-SUM problem is due to Devanur and Hayes [?]; we give the algorithm for completeness. Consider the LP relaxation of the MAX-SUM problem given in Figure 3(a). Here, x_{ij} is the fraction of request j that is assigned to provider i. The dual of this LP is given in Figure 3(b).

The algorithm has the following steps:

1. For the first ϵ fraction of the input, the algorithm solves the dual LP (Figure 2(b)) with B_i replaced by ϵB_i. Suppose the resulting dual variable values are z_i^* and w_i^*.

2. For every subsequent request j, the algorithm assigns it greedily to $\arg\max_{i \in D} b_{ij}(1 - z_i^*)$.

The next theorem is due to Devanur and Hayes [?].

THEOREM 4. *The competitive ratio of the above algorithm for the* MAX-SUM *problem is* $1 - O(\epsilon)$.

4. THE HYBRID PROBLEM

Maximize $\sum_{j\in Q}\sum_{i\in D} b_{ij}x_{ij}$ subject to

$$\begin{aligned}\sum_{i\in D} x_{ij} &\leq 1 \qquad \text{for all } j\in Q\\ \sum_{j\in Q} b_{ij}x_{ij} &\leq B_i \qquad \text{for all } i\in D\\ x_{ij} &\geq 0 \qquad \text{for all } i\in D, j\in Q\end{aligned}$$

(a) The primal LP for the MAX-SUM problem.

Minimize $\sum_{j\in Q} y_j + \sum_{i\in D} z_iB_i$ subject to

$$\begin{aligned} y_j &\geq b_{ij}(1-z_i) \qquad \text{for all } i\in D, j\in Q\\ y_j &\geq 0 \qquad \text{for all } j\in Q\\ z_i &\geq 0 \qquad \text{for all } i\in D\end{aligned}$$

(b) The dual LP for the MAX-SUM problem.

Figure 2: The MAX-SUM problem

Maximize $\sum_{j\in Q}\sum_{i\in D} b_{ij}x_{ij} - \alpha\sum_{i\in D}\pi_i$ subject to

$$\begin{aligned}\sum_{i\in D} x_{ij} &\leq 1 \qquad \text{for all } j\in Q\\ \sum_{j\in Q} b_{ij}x_{ij} &\leq B_i \qquad \text{for all } i\in D\\ \pi_i &\geq T_i - \sum_{j\in Q} b_{ij}x_{ij} \qquad \text{for all } i\in D\\ x_{ij} &\geq 0 \qquad \text{for all } i\in D, j\in Q\\ \pi_i &\geq 0 \qquad \text{for all } i\in D\end{aligned}$$

(a) The primal LP for the HYBRID problem.

Minimize $\sum_{j\in Q} y_j + \sum_{i\in D} z_iB_i$ subject to

$$\begin{aligned} y_j &\geq b_{ij}(1-z_i) \qquad \text{for all } i\in D, j\in Q\\ y_j &\geq 0 \qquad \text{for all } j\in Q\\ z_i &\geq 0 \qquad \text{for all } i\in D\end{aligned}$$

(b) The dual LP for the HYBRID problem.

Figure 3: The HYBRID problem

Recall that the HYBRID problem is a generalization of the MAX-SUM problem where a penalty of $\alpha\cdot\pi_i$ is charged for missing target T_i by a margin of π_i. Our algorithm for the HYBRID problem is a generalization of the MAX-SUM algorithm in the previous section. Consider the LP relaxation of the HYBRID problem given in Figure 3(a). Here, x_{ij} is the fraction of request j that is allocated to provider i, and $\alpha\cdot\pi_i$ is the penalty paid for failing to meet the minimum threshold T_i for provider i. The dual of this LP is given in Figure 3(b).

Before describing the algorithm formally, let us give an intuitive sketch. This algorithm explicitly uses the first ϵ fraction of the requests (call this the *initial portion* of the input) to learn the input distribution. However, the input distribution might have exponential support, and therefore a constant fraction is not sufficient for learning the distribution *per se*. Instead, the algorithm uses the initial portion of the input to learn the optimal dual variables. Once the initial portion of the input has arrived, the algorithm feeds it into the dual program and solves it. The key observation is that even though the initial portion is not a good representative of the overall input distribution, the dual variables obtained are approximately equal to the expected value of the optimal dual. Once the dual variables have been determined, the algorithm uses these dual variables to make assignment choices using a principle known as *complementary slackness*. This translates to the following rule: assign each request to the provider who has the highest bid for it, where the bids are discounted by a parameter dependent on the optimal dual variables.

Formally, the HYBRIDalgorithm has the following steps:

1. For the first ϵ fraction of the input, the algorithm solves the dual LP (Figure 3(b)) with B_i and T_i replaced by ϵB_i and ϵT_i respectively. Suppose that the resulting dual variable values are z_i^* and w_i^*.

2. For every subsequent request j, the algorithm assigns it greedily to $\arg\max_{i\in D} b_{ij}(1-z_i^*+w_i^*)$.

4.1 Analysis of the HYBRIDAlgorithm

Assumption. We will assume that OPT is *large*; in particular, that

$$\text{OPT} \geq (1+\alpha)\sqrt{\frac{n\ln(n/\epsilon^3)\cdot\sum_{j\in Q}(\max_{i\in D} b_{ij})}{3\epsilon}}.$$

Note that in practice, OPT and $\sum_{j\in Q}(\max_{i\in D} b_{ij})$ are orders of magnitude larger than all the other terms in the above expression. Therefore, the assumption essentially states that the square of the maximum revenue should dominate the sum of maximum bids. This clearly holds in practice because one would expect the optimal revenue to be comparable to the sum of maximum bids even without squaring.

It will be convenient to introduce a new notation β defined as

$$\beta = \frac{\text{OPT}}{\sum_{j\in Q}(\max_{i\in D} b_{ij})}. \tag{2}$$

Then, the assumption on the value of OPT becomes

$$\text{OPT} \geq \frac{(1+\alpha)^2 n\ln(n/\epsilon^3)}{3\epsilon\beta}. \tag{3}$$

With this assumption, we will now prove the next theorem.

THEOREM 5. *The competitive ratio of the above algorithm for the* HYBRID *problem is* $1-O(\epsilon)$.

Let us first establish the notation that we will use in the analysis. Let

$$x_{ij}(z,w) = \begin{cases} 1 & \text{if } i = \arg\max_{i\in D} b_{ij}(1-z_i+w_i)\\ 0 & \text{otherwise.}\end{cases}$$

So, $x_{ij}(z,w)$ is the indicator for whether request j will be assigned to provider i by the algorithm, if the dual variables are z and w. Let S denote the first ϵ fraction of the requests, and let $S^c = \{1,\ldots,m\}\setminus S$.

For the other notations, we use the following generic rule. Let R denote unrestricted revenue (i.e. not capped at the budgets), P and X denote the values of the primal and dual objectives, and π denote the penalty in the primal objective (before scaling by the penalty coefficient α). A subscript of i for any of these notations denotes the corresponding parameter for provider $i \in D$, whereas if there is no subscript, then we mean the sum of the parameter values over all $i \in D$. Further, these parameters are functions of (z, w, S) where z, w are the values of the dual variables and S is the set of requests over which the parameter is computed. If $S = Q$, we will drop it from the list of arguments. As examples of this notation scheme,

$$X_i(z, w, S) = (1 - z_i + w_i) \sum_{j \in S} x_{ij} b_{ij} + (z_i \cdot \epsilon B_i - w_i \cdot \epsilon T_i)$$

$$\pi(z, w) = \sum_{i \in D} \max\left(T_i - \sum_{j \in Q} x_{ij} b_{ij}, 0\right).$$

The first lemma states that the unrestricted revenue for any particular provider estimated from the initial portion of the input is an accurate estimate of the overall unrestricted revenue scaled by ϵ. This formalizes the intuition that we described earlier about the initial portion of the input providing good estimates of cumulative parameters even though it cannot help estimate the input probability distribution itself. The lemma is a direct consequence of Chernoff bounds (see, e.g., [**?**]).

LEMMA 6. *Suppose $b_{ij} \leq 1$ for all $i \in D, j \in Q$. For each $i \in D$ and any z, w, t_i,*

$$\mathbb{P}[|R_i(z, w, S) - \epsilon R_i(z, w)| > t_i] < 2e^{-\frac{t_i^2}{3\epsilon R_i(z,w)}}.$$

Next, we aggregate the concentration bounds obtained from the above lemma and claim that the sum of deviation of the revenue estimates obtained from the initial portion of the input for all the providers is a small fraction of the overall optimal revenue. (Due to space constraints, we omit the proof of this lemma.)

LEMMA 7. *There exists a set of values t_i, $i \in D$, such that*

$$\sum_{i \in D} t_i \leq \frac{\epsilon^2}{1 + \alpha} \text{OPT},$$

and with probability at least $1 - \epsilon$,

$$|R_i(z, w, S) - \epsilon R_i(z, w)| \leq t_i \text{ for all } i \in D.$$

Note that Lemma 7 essentially says that if all the budgets were ∞ and if there were no targets, then our algorithm is nearly optimal. We will now show that this claim about the restricted revenue can be extended to the budgeted revenue in the presence of penalties imposed by targets. Our main technical claim is to show that complementary slackness conditions are approximately satisfied by z^*, w^*. Let $a_i = t_i/\epsilon$. Note that by weak duality, if we can show that

$$X_i - P_i \leq (1 + \alpha) a_i \quad \text{for all } i \in D, \tag{4}$$

then it implies that

$$P \geq (1 - \epsilon)\text{OPT}. \tag{5}$$

Therefore, it is sufficient to show Eqn. (4).

We consider the two cases $w_i^* = 0$ and $w_i^* > 0$ separately. Let us outline the case of $w_i^* = 0$. Note that the complementary slackness conditions hold for the LP when restricted to the impressions in S since we solved this LP exactly when setting the dual variables. Therefore, $\pi_i^* = 0$. This observation now leads to a bound on π_i since π_i only depends on the unrestricted revenue and we have already shown that the overall unrestricted revenue R_i has small deviation from $R_i(S)/\epsilon$. Next, we use this observation about penalties and the complementary slackness conditions to obtain an upper bound on the gap between the solution obtained by the LP and a feasible dual solution, thereby proving Eqn. (4). We present the details of this proof in the appendix.

5. EXPERIMENTS

In this section, we set out to validate the effectiveness of our allocation algorithms. Specifically, we employ the algorithms in two settings, *viz.*, online ad allocations and online information markets. We will describe each setting in more detail next.

5.1 Online Ad Allocations

One of the most common applications of online allocation algorithms is in sponsored search. In this application, we map the producers to the advertisers who are interested in advertising slots associated with user queries which are the *requests*. The queries arrive in an online manner and the advertisers are typically budget constrained and specify their valuations (bids) for a set of keywords via a campaign which is run for a specified duration. We call a selected advertiser for a given ad slot as an *impressed* ad. An advertiser continues to participate in the allocation process for different keywords specified by her campaign so long as her budget is not exhausted. An advertiser's target is realized if her ad is shown in response to as many user queries as possible subject to her budget constraint. We report the following measures of effectiveness of all the three algorithms — total *revenue* generated by all the advertisers; the *fractional target realization* for an advertiser, i.e., the ratio of budget spent to the overall budget for an advertiser; *winning rate* of an advertiser, i.e., the impression-to-participation ratio of an advertiser; and, the *coverage* of advertisers, i.e., number of unique advertisers that won at least one impression. The first measure focuses on the central performance objective of the algorithms, i.e., revenue maximization, while the other two measures focus on the fairness aspect of the algorithms.

In this experiment, we sampled $5,000,000$ queries and the corresponding bids from the logs for a single day of a commercial ad delivery engine, each of which had at least 10 participating advertisers. The *relative budgets* of advertiser is computed as their expected cost over all their bids. This value is obtained by summing the product of the bid and the clickthrough rate (probability of the ad being clicked) over all the queries. The real budget of an advertiser is now obtained by scaling the relative budget using a scaling factor β that we vary over our experiments. For every impressed ad, we reduce an advertiser's remaining budget by her expected cost, i.e, the product of her bid and the clickthrough rate of her impressed ad. We ran the MAX-MIN, MAX-SUM, and HYBRID algorithms considering all participating advertisers for a given query and report the above three measures of effectiveness. In the case of the HYBRID algorithm, we chose a penalty coefficient $\alpha = 30$.

In the first experiment, we measured the effect of the budget size on the revenue and fairness of the resulting allocation from each algorithm. Clearly, as the scaling factor β increases, the amount of budget left for any advertiser to

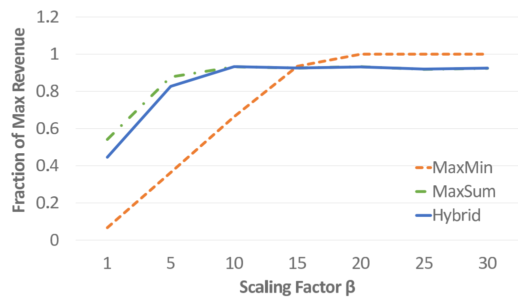


Figure 4: The effect of the scaling factor β on revenue in online ad allocation

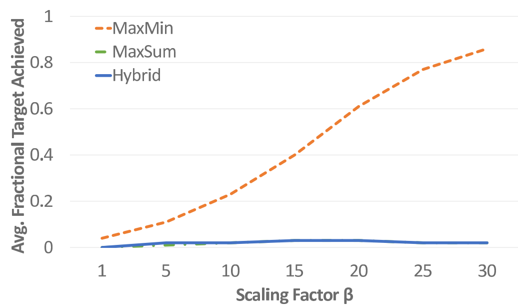


Figure 5: The effect of the scaling factor β on fairness (fractional target coverage) in online ad allocation

participate in an auction decreases and therefore becomes a critical factor for the allocation algorithm to consider. In this experiment, we varied β from 1 to 30 and measured the total expected revenue resulting from the allocation. We set the target at 20% of the budget for each advertiser. Figure **??** illustrates the revenue derived from each allocation for different values of β.

For smaller values of β when the corresponding budgets are large, the Max-Sum and Hybrid algorithms earn more revenue than Max-Min since they optimize revenue. However, somewhat surprisingly, as the budget decreases, i.e., at larger values of β, these algorithms start to underperform w.r.t. Max-Min. We suspect this is so because the Max-Min algorithm is more egalitarian in its initial allocations and therefore, uses up the budgets on the advertisers in a uniform manner. On the other hand, the allocations in the Max-Sum and Hybrid algorithms result in significant skew in budget utilization of the advertisers and some of the advertisers' budgets are never used up.

Next, we measured the fairness of the resulting allocations for the same values of β. Our metric for fairness is the average, over the lowest $k\%$ for some k, of the fractional coverage of the advertisers. Figure **??** shows the results of this experiment, where $k = 1$.

Even when the budgets are large, Max-Min does a better job of allocating the ad slots to more advertisers than Hybrid and Max-Sum. Now, as the budgets decrease, Max-Min does a significantly better allocation in terms of fairness compared to the other two algorithms. In fact, in this range, it does better than the other two algorithms for *both* fairness and revenue objectives. When we increase k, this effect becomes more pronounced. This is illustrated in

Bucket	Max-Min	Max-Sum	Hybrid
1	0.004	0.05	0.05
2	0.004	0.06	0.06
3	0.005	0.06	0.07
4	0.007	0.07	0.08
5	0.02	0.10	0.10
6	0.02	0.11	0.11
7	0.03	0.11	0.13
8	0.04	0.16	0.15
9	0.79	0.25	0.22
10	0.08	0.03	0.03

Table 1: Distribution of advertisers in terms of fractional target coverage in online ad allocation ($\beta = 15$)

Figures **??** and **??** where we range k from 1 to 10. We used two values of the scaling parameter of $\beta = 5$ and $\beta = 15$ in this experiment.

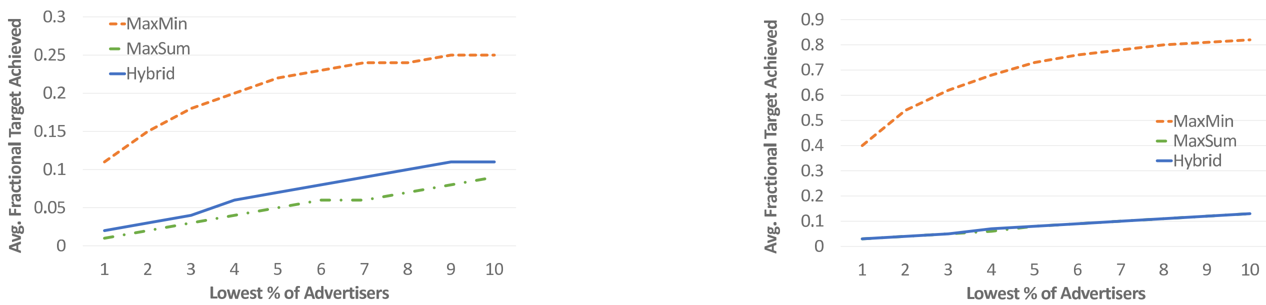
Recall that the number reported in this figure is the average of fractional target achieved over all the advertisers in the lowest $k\%$. As we increase k, the average levels off toward the average for the entire set of advertisers. Initially, the increase in the fairness of the allocation from Max-Min is more pronounced. On the other hand, the increase in the allocations from Max-Sum and Hybrid are more gradual and only realize around 15% of the possible target for $k = 10$ for both values of β. This behavior suggests that Max-Min uniformly increases the allocations to all the advertisers, while Max-Sum and Hybrid have a skewed allocation at the end. To verify this observation, we fixed the scaling factor to 15 (i.e., use small budgets) to verify whether Max-Min would push all the advertisers toward realizing their targets. We bucketed the advertisers based on fraction of their targets realized with bucket 1 corresponding to 10% of the realized target and bucket 10 corresponding to the target being realized. Table **??** shows the resulting histogram validating our hypothesis.

Finally, we measured other indicators of performance of each algorithm. Specifically, we measured the winning rate of the advertisers as the ratio of the number of impressions to the number of non-zero bids. Figure **??** highlights the performance of the algorithms with respect to the advertisers in the lowest $k\%$ (ordered by the winning rate). Even though the difference is very small, we did find a consistent trend of the Max-Min algorithm improving the winning rates of the advertisers and is consistent with the other performance indicators.

5.2 Online Information Markets

Another important and emerging paradigm of online markets are information markets wherein data providers can sell raw data or services utilizing the data. Two large scale examples are the Google Custom Search[1] and the Bing Search Service[2] announced by Microsoft on its Azure marketplace[3]. Other examples include `http://www.globalfinancialdata.com` and `http://www.elasticpath.com/products`. To validate our algorithms in this setting, we implemented a scaled down version of such a service using data from three prin-

[1] `https://www.google.com/cse/`
[2] `http://datamarket.azure.com/dataset/bing/search`
[3] `https://datamarket.azure.com`



(a) $\beta = 5$ (b) $\beta = 15$

Figure 6: Fractional target coverage for the lowest k% of advertisers in online ad allocation

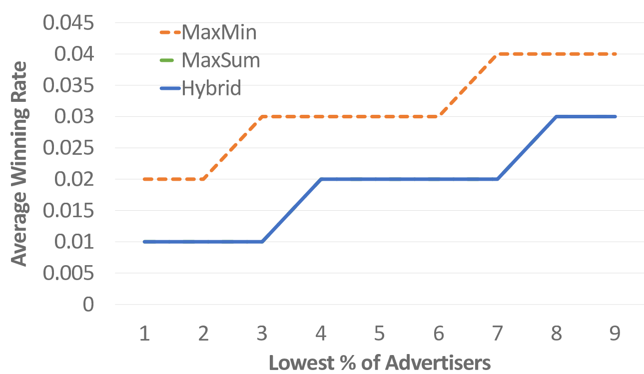


Figure 7: The winning rates of advertisers in online ad allocation ($\beta = 15$)

Index	Category	Score
0	computing|computers|laptop computers	0.337
0	clothing & shoes|costumes	0.489
1	cameras & optics|cameras|digital cameras	0.316
1	computing|computers|laptop computers	0.314
1	electronics|audio electronics|mp3 players	0.327
7	electronics|cell phones & plans|cell phones	0.253
7	electronics|audio electronics|speakers	0.316
7	home furnishings|furniture|beds	0.346
12	electronics|tv & video|televisions	0.430
12	electronics|audio electronics|mp3 players	0.319
12	electronics|audio electronics|speakers	0.322
12	electronics|electronics accessories|headphones	0.315
19	electronics|tv & video|televisions	0.288
19	cameras & optics|camcorders	0.442

Table 2: A view into the indexes supporting different subsets of categories with different scores for online information markets

cipal sources used in a commercial Shopping service, *viz.*, the commerce queries to the service, the taxonomy of the products in the commerce index, and the product catalog. Specifically, we considered the top 20 categories in the service's Shopping taxonomy ranked by popularity in the user queries. Let us denote them by $\mathcal{C}$. We then considered all products restricted to these categories from the catalog. We considered a sample of commerce queries received by the shopping service in a single day of October 2012. Restricting the queries in $\mathcal{Q}$ to those that are associated with the important categories in $\mathcal{C}$ yielded us a set of around 200,000 queries $\mathcal{Q}_{\sqsubseteq}$. These queries served as *requests* in our experiment. Next, we describe how we instantiate the *providers*.

We created 1000 indexes in the backend to service the product queries. We proceeded with the index creation as follows. We randomly allocated the products to different indices. Thus, each index served as a proxy for a *provider* and was associated with a subset of commerce categories along with a subset of products in each of the associated categories. As a measure of the quality of each provider, we used some well-known measures of economic relevance [**?**, **?**] of the products for all the associated categories. Specifically, we used the *review count*, *average review score*, *number of merchant offers*, *price index*, and the *product differentiating index* i.e., how differentiating are the set of products it contains. Note that these are query independent features and can readily be computed by any provider. Further the associated data is very succinct and can readily be stored and used at query time by our algorithms. Table **??** summarizes a subset of the resulting set of providers.

Unlike the previous experiment, there are no explicit targets associated with each provider. Instead we use the category specific quality score of each provider as a proxy for the fraction of queries it expects to serve. Since there is no specific auction in this setting either, we measured only the *fraction of achieved target* for each provider. For every query q, we compute the score of an index I as

$$r(q, I) = \sum_{C \in \mathcal{C}} \mathbb{P}(q \in C) \cdot \mathbb{P}(q \in I | q \in C),$$

where the $\mathbb{P}(.|.)$ denotes the conditional probability of the respective events. We associated a target for each index I to be the $\sum_{q \in \mathcal{Q}_v} r(q, I)$ scaled by a factor β which we vary in the experiments to control the targets appropriately. We varied β from 1 to 30 in steps of 5.

In the first experiment, we measured the effect of β on both the revenue and fairness resulting from the allocations. Figure **??** illustrates the change in revenue as the targets are reduced. While the MAX-MIN algorithm very slightly underperforms the MAX-SUM and HYBRID algorithms for larger targets (i.e., smaller values of β), it begins to outperform MAX-SUM for smaller targets. One interesting observation here is that HYBRID exhibits similar performance as MAX-MIN unlike in the previous experiments involving online ads. Also, unlike the previous case, none of the allocations result in indexes reaching their targets. This is because

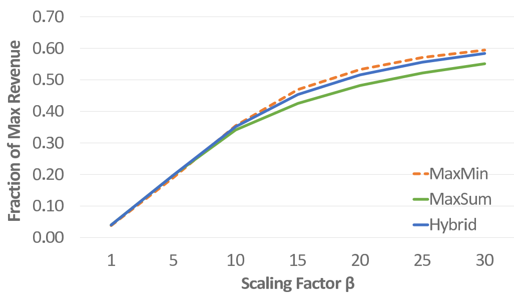


Figure 8: The effect of the scaling factor β on revenue in online information markets

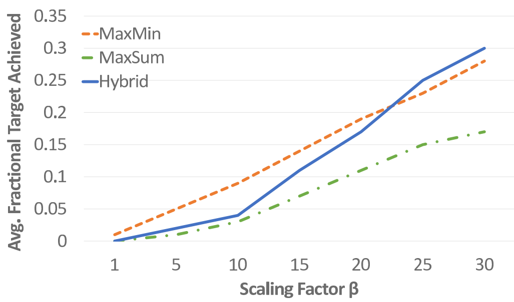


Figure 9: The effect of the scaling factor β on fairness (fractional target coverage) in online information markets

the bid-to-budget ratio is much smaller in this setting.

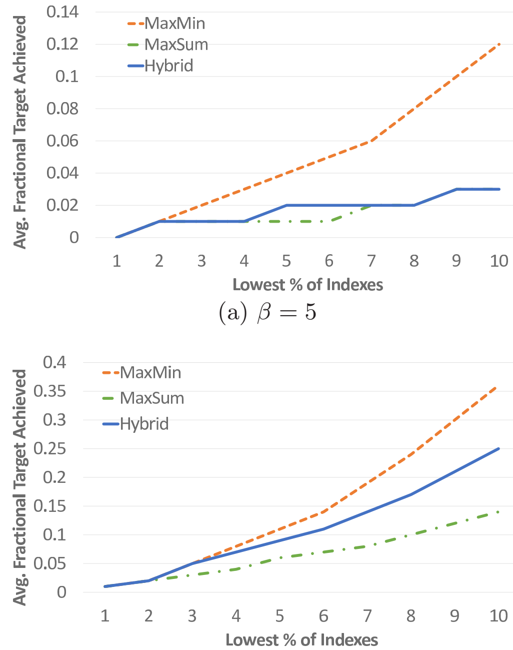
Continuing with the experiment, we measured the fairness, *viz.*, the average value of the fractional targets achieved in the lowest $k\%$ of indexes. Figure **??** shows that both the Max-Min and Hybrid algorithms substantially outperform Max-Sum, particularly for smaller targets.

Next, we measured the fairness of the allocations as k is increased from the bottom 1% to 10%. This was to observe how the allocations changes as one goes up in the order. We used two values of the scaling parameter β to cover both small and large targets. As Figures **??** and **??** show, the Max-Min algorithm does a better job of trying to equalize the fractional realized targets across all indexes compared to the Max-Sum and Hybrid algorithms. For larger targets, both Max-Sum and Hybrid do not make any progress toward balancing the allocations even for higher values of k.

Finally, we measure the winning rate of selection of indexes to answer user queries. Figure **??** shows that the Max-Min algorithm produces allocations in which the lowest 1% of indexes get selected, on average, almost 8% of the time they qualify for answering the user query, while the corresponding numbers for both Max-Sum and Hybrid are next to zero even as the percentage of indexes included increases from the lowest 1% to lowest 10%. In the same range, the performance of Max-Min improves from 8% to nearly 13%.

6. CONCLUSIONS

In this paper, we formulated three generic online allocation problems Max-Min, Max-Sum, and Hybrid that aim at a combination of fairness and revenue optimization objectives. We gave algorithms for these problem that we proved analytically are nearly optimal in their respective objectives.



(a) $\beta = 5$

(b) $\beta = 15$

Figure 10: Fractional target coverage for the lowest $k\%$ of indexes in online information markets

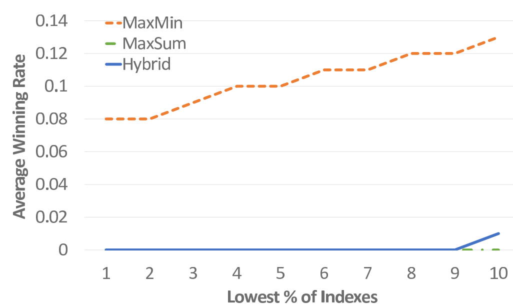


Figure 11: The average winning rate of the indexes in online information markets ($\beta = 15$)

We compared the algorithms on fairness and revenue metrics on two large real-world data sets corresponding online ad allocation and online information markets, and concluded that revenue-maximizing algorithms are consistently outperformed by fairness-aware algorithms on multiple fairness objectives. Moreover, the fairness-aware algorithms generate revenue that is comparable to (in some input ranges, even greater than) the revenue-maximizing algorithms. These observations offer compelling proof of the importance of incorporating fairness objectives in online allocation algorithms.

7. REFERENCES

[1] Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *CoRR*, abs/0911.2974, 2009.

[2] Sanjoy K. Baruah, N. K. Cohen, C. Greg Plaxton, and Donald A. Varvel. Proportionate progress: A notion of fairness in resource allocation. *Algorithmica*, 15(6):600–625, 1996.

[3] Niv Buchbinder, Kamal Jain, and Joseph Naor. Online primal-dual algorithms for maximizing ad-auctions revenue. In *ESA*, pages 253–264, 2007.

[4] Nikhil R. Devanur and Thomas P. Hayes. The adwords problem: online keyword matching with budgeted bidders under random permutations. In *ACM Conference on Electronic Commerce*, pages 71–78, 2009.
[5] Nikhil R. Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A. Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. In *ACM Conference on Electronic Commerce*, pages 29–38, 2011.
[6] Jon Feldman, Monika Henzinger, Nitish Korula, Vahab S. Mirrokni, and Clifford Stein. Online stochastic packing applied to display ad allocation. In *ESA (1)*, pages 182–194, 2010.
[7] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica. Dominant resource fairness: Fair allocation of multiple resource types. In *NSDI*, 2011.
[8] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica. Dominant resource fairness: fair allocation of multiple resource types. In *USENIX NSDI*, page 24, 2011.
[9] Ashish Goel, Adam Meyerson, and Serge A. Plotkin. Combining fairness with throughput: Online routing with multiple objectives. *J. Comput. Syst. Sci.*, 63(1):62–79, 2001.
[10] Gagan Goel and Aranyak Mehta. Online budgeted matching in random input models with applications to adwords. In *SODA*, pages 982–991, 2008.
[11] Carlee Joe-Wong, Soumya Sen, Tian Lan, and Mung Chiang. Multiresource allocation: Fairness-efficiency tradeoffs in a unifying framework. *IEEE/ACM Trans. Netw.*, 21(6):1785–1798, 2013.
[12] Thomas Kesselheim, Klaus Radke, Andreas Tönnis, and Berthold Vöcking. Primal beats dual on online packing lps in the random-order model. In *STOC*, 2014.
[13] Jon Kleinberg, Yuval Rabani, and Eva Tardos. Fairness in routing and load balancing. In *FOCS*, pages 568–578, 1999.
[14] Aranyak Mehta. Online matching and ad allocation. *Foundations and Trends in Theoretical Computer Science*, 8(4):265–368, 2013.
[15] Aranyak Mehta and Vahab Mirrokni. Online ad serving: Theory and practice. `http://www.sigmetrics.org/sigmetrics2011/tutorials/tutorial3.pdf`, 2011.
[16] Aranyak Mehta, Amin Saberi, Umesh V. Vazirani, and Vijay V. Vazirani. Adwords and generalized online matching. *J. ACM*, 54(5), 2007.
[17] Marco Molinaro and R. Ravi. Geometry of online packing linear programs. In *ICALP (1)*, pages 701–713, 2012.
[18] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1997.
[19] Noam Nisan, Jason Bayer, Deepak Chandra, Tal Franji, Robert Gardner, Yossi Matias, Neil Rhodes, Misha Seltzer, Danny Tom, Hal R. Varian, and Dan Zigmond. Google's auction for tv ads. In *ICALP (2)*, pages 309–327, 2009.
[20] Debmalya Panigrahi, Atish Das Sarma, Gagan Aggarwal, and Andrew Tomkins. Online selection of diverse results. In *WSDM*, pages 263–272, 2012.
[21] J. Rowley. Product search in e-shopping: a review and research propositions. *Journal of Consumer Marketing*, 17(1):20–35, 2000.
[22] Saswati Sarkar and Kumar N. Sivarajan. Fairness in cellular mobile networks. *IEEE Transactions on Information Theory*, 48(8):2418–2426, 2002.
[23] W. R. Smith. Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, 21(1):3 – 8, 1956.

APPENDIX

Proof of Lemma 2. Consider a hypothetical algorithm that allocates request j of type ξ with probability $w_i(\xi)$ to provider $i \in D$. The expected decrease of $\bar{\Phi}$ for this algorithm at any stage is

$$\begin{aligned}
&\sum_{\xi} p(\xi) \sum_{i\in D} w_i(\xi) \int_{k=c_i}^{c_i+b_i(\xi)/T_i} \phi(k) \\
= \quad & \sum_{\xi} p(\xi) \sum_{i\in D} w_i(\xi) \left(\frac{\phi(c_i)\cdot b_i(\xi)}{T_i}\right)\left(\frac{c_{\text{OPT}}\cdot T_i}{\alpha\cdot b_i(\xi)\ln n}\right) \times \\
& \times \left(1 - e^{-\frac{\alpha\cdot b_i(\xi)\ln n}{c_{\text{OPT}}\cdot T_i}}\right),
\end{aligned}$$

by using the definition of ϕ and algebraic manipulations. Since $\frac{1-e^{-x}}{x} \geq 1 - x/2$, the expected decrease of $\bar{\Phi}$ is at least

$$\begin{aligned}
&\sum_{\xi} p(\xi) \sum_{i\in D} w_i(\xi) \left(\frac{\phi(c_i)\cdot b_i(\xi)}{T_i}\right)\left(1 - \frac{\alpha\cdot b_i(\xi)\ln n}{2c_{\text{OPT}}\cdot T_i}\right) \\
\geq \quad & \sum_{\xi} p(\xi) \sum_{i\in D} w_i(\xi) \left(\frac{\phi(c_i)\cdot b_i(\xi)}{T_i}\right)\left(1 - \frac{1}{\beta}\right),
\end{aligned}$$

by Eqn. (1)). Rearranging terms, we get

$$\begin{aligned}
&\left(1-\frac{1}{\beta}\right) \sum_{i\in D} \frac{\phi(c_i)}{T_i} \sum_{\xi} p(\xi)\cdot w_i(\xi)\cdot b_i(\xi) \\
\geq \quad & \left(1-\frac{1}{\beta}\right) \sum_{i\in D} \frac{\phi(c_i)\cdot c_{\text{OPT}}}{m} = \left(1-\frac{1}{\beta}\right)\frac{\alpha\ln n}{m}\bar{\Phi}
\end{aligned}$$

by the definition of ϕ. Since the Max-Min algorithm assigns the new request j to the provider that *maximizes* the decrease in the value of $\bar{\Phi}$, the lemma follows.

Proof of Theorem 5. For simplicity of notation, we drop z^*, w^* from the arguments since those are the only values of z, w that we consider.

First, let us consider $w_i^* = 0$. By complementary slackness conditions on the LP restricted to the requests in S, we have $\pi_i^* = 0$, and therefore, $R_i(S) = \sum_{j\in S} x_{ij}b_{ij} \geq \epsilon T_i$. Multiplying by $1/\epsilon$ and rearranging terms, we get $\pi_i = T_i - R_i \leq t_i/\epsilon = a_i$, which implies that $P_i = \min(B_i, R_i) - \alpha\pi_i \geq \min(B_i, R_i) - \alpha a_i$. Since $w_i^* = 0$, $X_i = (1-z_i^*)R_i + z_i^* B_i$. We have two subcases.

1. $z_i^* > 0$: In this case, we have
$$\begin{aligned}
& X_i - P_i \leq (1-z_i^*)R_i + z_i^*B_i - \min(B_i, R_i) + \alpha a_i \\
\leq \quad & \max(B_i, R_i) - \min(B_i, R_i) + \alpha a_i = |B_i - R_i| + \alpha a_i.
\end{aligned}$$
Since $z_i^* > 0$, from complementary slackness conditions on the LP restricted to requests in S, we have that $R_i(S) = \epsilon B_i$. Now, from Lemma 7, we have that $|B_i - R_i| \leq a_i$.
2. $z_i^* = 0$: In this case, $X_i = R_i$. First, consider the case $R_i \leq B_i$. Then, $P_i \geq \min(R_i, B_i) - \alpha a_i = R_i - \alpha a_i$, which implies that $X_i - P_i \leq \alpha a_i \leq (1+\alpha)a_i$. Now, consider the case $R_i \geq B_i$. We have $X_i = R_i \leq R_i(S)/\epsilon + a_i$. Since $R_i(S) \leq \epsilon B_i$, we have $X_i \leq B_i + a_i$. Now, $X_i = R_i \leq B_i + a_i$ and $P_i \geq \min(R_i, B_i) - \alpha a_i$ imply $X_i - P_i \leq (1+\alpha)a_i$.

Similar arguments (omitted due to space constraints) are used for the case $w_i^* > 0$.

Finally, we will prove that $P(S^c) \geq (1-\epsilon)P - \sum_{i\in D} t_i$. We have already noted that if $\pi_i(S) = 0$, then $\pi_i \leq a_i$. If $\pi_S > 0$, then

$$\pi_i(S) = \epsilon T_i - R_i(S) \geq \epsilon T_i - \epsilon R_i - t_i.$$

Multiplying by $1/\epsilon$, we get $\pi_i \leq \frac{\pi_i(S)}{\epsilon} + a_i$, and therefore,

$$\pi_i(S^c) = \pi_i - \pi_i(S) \leq (1-\epsilon)\pi_i + a_i.$$

Lemma 7 says that $R_i(S^c) > (1-\epsilon)R_i - t_i$ for all $i \in D$.

$$\begin{aligned} \text{Thus,} \qquad P_i(S^c) &= \min((1-\epsilon)B_i, R_i(S^c)) - \alpha\pi_i(S^c) \\ &\geq \min((1-\epsilon)B_i, (1-\epsilon)R_i - t_i) - (1-\epsilon)\pi_i - \alpha a_i \\ &\geq (1-\epsilon)P_i - (1+\alpha)a_i \quad \text{(since } t_i \leq a_i\text{)}. \end{aligned}$$

Since $\sum_{i \in D}(1+\alpha)a_i \leq \epsilon$ OPT, Theorem 5 follows.