Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Geometry of Online Packing Linear Programs

Marco Molinaro, R. Ravi

To cite this article:
Marco Molinaro, R. Ravi (2014) The Geometry of Online Packing Linear Programs. Mathematics of Operations Research 39(1):46-59. http://dx.doi.org/10.1287/moor.2013.0612

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 39, No. 1, February 2014, pp. 46–59
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.2013.0612

The Geometry of Online Packing Linear Programs

Marco Molinaro, R. Ravi
Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
{molinaro@cmu.edu, ravi@andrew.cmu.edu}

We consider packing linear programs with m rows where all constraint coefficients are normalized to be in the unit interval. The n columns arrive in random order and the goal is to set the corresponding decision variables irrevocably when they arrive to obtain a feasible solution maximizing the expected reward. Previous $(1-\epsilon)$-competitive algorithms require the right-hand side of the linear program to be $\Omega((m/\epsilon^2)\log(n/\epsilon))$, a bound that worsens with the number of columns and rows. However, the dependence on the number of columns is not required in the single-row case, and known lower bounds for the general case are also independent of n.

Our goal is to understand whether the dependence on n is required in the multirow case, making it fundamentally harder than the single-row version. We refute this by exhibiting an algorithm that is $(1-\epsilon)$-competitive as long as the right-hand sides are $\Omega((m^2/\epsilon^2)\log(m/\epsilon))$. Our techniques refine previous probably approximately correct learning based approaches that interpret the online decisions as linear classifications of the columns based on sampled dual prices. The key ingredient of our improvement comes from a nonstandard covering argument together with the realization that only when the columns of the linear program belong to few one-dimensional subspaces we can obtain such small covers; bounding the size of the cover constructed also relies on the geometry of linear classifiers. General packing linear programs are handled by perturbing the input columns, which can be seen as making the learning problem more robust.

Keywords: linear programming; online algorithms; PAC learning
MSC2000 subject classification: Primary: 90C05, 68W27; secondary: 68W20
OR/MS subject classification: Primary: programming, linear; secondary: analysis of algorithms
History: Received November 3, 2012; revised January 24, 2013, April 9, 2013. Published online in *Articles in Advance* August 1, 2013.

1. Introduction. Traditional optimization models usually assume that the input is known a priori. However, in most applications, the data is either revealed over time or only coarse information about the input is known, often modeled in terms of a probability distribution. Consequently, much effort has been directed toward understanding the quality of solutions that can be obtained without full knowledge of the input, which led to the development of online and stochastic optimization (Birge and Louveaux [6], Borodin and El-Yaniv [7]). Emerging problems such as allocating advertisement slots to advertisers and yield management on the Internet are of inherent online nature and have further accelerated this development (Agrawal et al. [1]).

Linear programming is arguably the most important and thus well-studied optimization problem. Therefore, understanding the limitations of solving linear programs when complete data is not available is a fundamental theoretical problem with a slew of applications, including the ad allocation and yield management problems above. Indeed, a simple linear program with one uniform knapsack, the secretary problem, was one of the first online problems to be considered and an optimal solution was already obtained by the early 1960s (Dynkin [13], Gilbert and Mosteller [16]). Although the single knapsack case is currently well understood under different models of how information is revealed (Babaioff et al. [3]), much less is known about problems with multiple knapsacks. Only recently, algorithms with guaranteed solution quality have been developed for these more general packing problems (Agrawal et al. [1], Devanur et al. [11], Feldman et al. [15]).

1.1. The model. We consider the following online packing linear programing problem. Consider a fixed but unknown linear program (LP) with n columns $a^t \in [0,1]^m$ (whose associated variables are constrained to be in $[0,1]$) and m packing constraints:

$$\begin{aligned} OPT = \max & \sum_{t=1}^{n} \pi_t x_t \\ & \sum_{t=1}^{n} a^t x_t \le B \qquad \text{(LP)} \\ & x_t \in [0,1]. \end{aligned}$$

Columns are presented in a random (uniform) order, and whenever a column is presented we are required to irrevocably choose the value of its corresponding variable. We assume that the number of columns n is known. (Actually knowing n up to $1 \pm \epsilon$ factor is enough. This assumption is required to allow algorithms with

nontrivial competitive ratios (Devenur and Hayes [10]).) The goal is to obtain a feasible solution to the LP while maximizing its value. Note that we use OPT to denote the optimum value of the (offline) LP.

By scaling down rows as necessary, we assume without loss of generality that all entries of B are the same, which we also denote with some overload of notation by B. Because of the packing nature of the problem, we also assume without loss of generality that all the π_t's are positive and all the a^t's are nonzero: we can simply ignore columns that do not satisfy the first property and always set to 1 the variables associated to the remaining columns that do not satisfy the second property.

The *random permutation* model, where the input is presented in a random order, has grown in popularity (Babaioff et al. [3], Devenur and Hayes [10], Goel and Mehta [17]), since it avoids strong lower bounds of the pessimistic adversarial-order model (Buchbinder and Naor [8]), while still capturing the lack of total information a priori. Moreover, the random permutation model is weaker than the *independent and identically distributed (i.i.d.)* model that assumes that the parts constituting the input are sampled independently from a fixed distribution, which is either known or unknown.

1.2. Related work. Many different types of online problems have already been studied in the random permutation model. These include bin packing (Kenyon [20]), matchings (Goel and Mehta [17], Karp et al. [19]), the AdWords problem (Devenur and Hayes [10]), and different generalizations of the secretary problem (Babaioff et al. [3, 4], Bateni et al. [5], Im and Wang [18], Soto [25]). Closest to our work are packing problems with a single knapsack constraint. In Kleinberg [21], Kleinberg considered the B-choice secretary problem, where the goal is to select at most B items coming online in random order to maximize profit. The author presented an algorithm with competitive ratio $1 - O(1/\sqrt{B})$ and showed that $1 - \Omega(1/\sqrt{B})$ is the best possible. Generalizing the B-choice secretary problem, Babaioff et al. [2] considered the online knapsack problem and presented a $(1/10e)$-competitive algorithm. Notice that in both cases the competitive ratio does not depend on n.

Despite all these works, results for the more general online packing LPs considered here were only recently obtained by Feldman et al. [15] and Agrawal et al. [1]. The first paper presents an algorithm that obtains with high probability a solution of value at least $(1-\epsilon)OPT$ whenever $B \geq \Omega((m \log n)/\epsilon^3)$ and $OPT \geq \Omega((\pi_{\max} m \log n)/\epsilon)$, where $\pi_{\max}$ is the largest profit. In the second paper, the authors present an algorithm dubbed *DPA* (dynamic pricing algorithm), which obtains a solution of expected value at least $(1-\epsilon)OPT$ under the weaker assumptions $B \geq \Omega((m/\epsilon^2)\log(n/\epsilon))$ or $OPT \geq \Omega(((\pi_{\max} m^2)/\epsilon^2)\log(n/\epsilon))$. One other way of stating this result is that the algorithm has competitive ratio $1 - O(\sqrt{m \log(n) \log B}/\sqrt{B})$; this guarantee degrades as n increases. The current lower bound on B to allow $(1-\epsilon)$-competitive algorithms is $B \geq \log m/\epsilon^2$, also presented in Agrawal et al. [1]. We remark that these algorithms actually work for more general allocation problems, where a set of columns representing various options arrive at each step and the solution may choose at most one of the options.

Both of the above algorithms use a connection between solving the online LP and *probably approximately correct (PAC) learning* (Cucker and Zhou [9]) a linear classification of its columns, which was initiated by Devanur and Hayes [10] in the context of the AdWords problem. Here we further explore this connection, and our improved bounds can be seen as a consequence of making the learning algorithm more robust by suitably changing the input LP. Robustness is a topic well studied in learning theory (Devroye and Wagner [12], Kutin and Niyogi [22]), although existing results do not seem to apply directly to our problem. We remark that a component of robustness more closely related to the standard PAC-learning literature was also used by Devanur and Hayes [10].

In recent work, Devanur et al. [11] consider the weaker *i.i.d. model* for the general allocation problem. Whereas in the random permutation model one assumes that columns are sampled without replacement, in the i.i.d. model they are sampled with replacement. Making use of the independence between samples, Devanur et al. substantially improve requirements on B to $\Omega(\log(m/\epsilon)/\epsilon^2)$ while showing that the lower bound $\Omega(\log m/\epsilon^2)$ still holds in this model. We remark, however, that these models can present very different behaviors: as a simple example, consider an LP with n columns, $m = 1$ constraints and budget $B = 1$, where only one of the columns has $\pi_1 = a^1 = 1$ and all others have $\pi_t = a^t = 0$; in the random permutation model the expected value of the optimal solution is 1, and in the i.i.d. model this value is $1 - (1 - 1/n)^n \to 1 - 1/e$. The competitiveness of the algorithm of Devanur et al. [11] under the random permutation model is still unknown and was left as an open problem by the authors.

1.3. Our results. Our focus is to understand how large B is required to be in order to allow $(1-\epsilon)$-competitive algorithms. In particular, the best known bounds for B mentioned above degrade as the number of columns in the LP increases, whereas the minimum requirement on its magnitude does not. With the trend of handling LPs with larger number of columns (e.g., these columns correspond to the key words in the ad allocation problem, which in turn correspond to visits of a search engine's Web page), this gap is very unsatisfactory from a practical

point of view. Furthermore, given that guarantees for the single knapsack case do not depend on the number of columns, it is important to understand if the multiknapsack case is fundamentally more difficult. In this work, we give a precise indication of why the latter problem was resistant to arguments used in the single knapsack case, and overcome this difficulty to exhibit an algorithm with dimension-independent guarantee.

We show that a modification of DPA (Agrawal et al. [1]) that we call *Robust DPA* obtains a $(1-\epsilon)$-competitive solution for online packing LPs with m constraints in the random permutation model whenever $B \geq \Omega((m^2/\epsilon^2)\log(m/\epsilon))$. Another way of stating this result is that the algorithm has competitive ratio $1 - O(m\sqrt{\log B}/\sqrt{B})$. Contrasting to previous results, our guarantee does not depend on n and in the case $m = 1$ matches the bounds for the B-choice secretary problem up to lower order terms. We finally remark that we can replace the requirement $B \geq \Omega((m^2/\epsilon^2)\log(m/\epsilon))$ by $OPT \geq \Omega(((\pi_{\max} m^3)/\epsilon^2)\log(m/\epsilon))$ exactly as done in Section 5.1 of Agrawal et al. [1].

1.4. High-level outline. As mentioned before, we use the connection between solving an online LP and PAC learning a good linear classification of its columns; to obtain the improved guarantee, we focus on tightening the bounds for the generalization error of the learning problem. More precisely, solving the LP can be seen as classifying the columns into 0/1, which corresponds to setting their associated variable to 0/1. Consider a family $\mathscr{X} \subseteq \{0,1\}^n$ of linear classifications of the columns. Our algorithms essentially sample a set S of columns and learn a classification $x^S \in \mathscr{X}$ that is "good" for the columns S (i.e., obtains large proportional revenue while not filling up the proportionally scaled budget too much). The goal is to upper bound the probability that x^S is not good for the whole LP. This is typically done by union bounding over the classifications in $\mathscr{X}$ (Devenur and Hayes [10], Agrawal et al. [1]).

To obtain improved guarantees, we refine this bound using an argument akin to covering: we consider *witnesses* (§2.3), which are representatives of groups of "similar" bad classifications that can be used to bound the probability that *any* classification in the group is learned; for that we need to use a nonstandard measure of similarity between classifications that is based on the budget of the LP. The problem is that, when the columns (π_t, a^t)'s do not lie in a two-dimensional subspace of $\mathbb{R}^m$, the set $\mathscr{X}$ may contain a large number of mutually dissimilar bad classifications; this is a roadblock for obtaining a small set of witnesses. In stark contrast, when these columns do lie in a two-dimensional subspace (e.g., $m = 1$), these classifications have a much nicer structure that admits a small set of witnesses. This indicates that the latter learning problem is intrinsically more robust than the former, which seems to precisely capture the increased difficulty in obtained good bounds for the multiknapsack case.

Motivated by this discussion, we first consider LPs whose columns a^t's lie in *few* one-dimensional subspaces (§2). For each of these subspaces, we are able to approximate the classifications induced in the columns lying in the subspace by considering a small subset of the induced classifications; patching together these partial classifications gives us a witness set for $\mathscr{X}$. However, this strategy as stated does not make use of the fact that the subspaces are embedded in an m-dimensional space, and hence leads to large witness sets. By establishing a connection between the "useful" patching possibilities with faces of a hyperplane arrangement in $\mathbb{R}^m$ (Lemma 8), we are able to make use of the dimension of the host space and exhibit witness sets of much smaller sizes, which leads to improved bounds.

For the general problem, the idea is to perturb the columns a^t's to make them lie in few one-dimensional subspaces, while not altering the feasibility and optimality of the LP by more than a $1 \pm \epsilon$ factor (§3). Finally, we tighten the bound by using the idea of recomputing the classification as the number of columns doubles, following (Agrawal et al. [1]) (§5).

2. OTP for almost one-dimensional columns. In this section we describe and analyze the behavior of the algorithm OLA (one-time learning algorithm) for LPs whose columns are contained in few one-dimensional subspaces of $\mathbb{R}^m$. The main idea behind the algorithm is to find an appropriate dual (perhaps infeasible) solution p for (LP) and use it to classify the columns of the LP. More precisely, given $p \in \mathbb{R}^m$, we define $x(p)_t = 1$ if $\pi_t > pa^t$ and $x(p)_t = 0$ otherwise. Thus, $x(p)$ is the result of classifying the columns (π_t, a^t)'s with the homogeneous hyperplane in $\mathbb{R}^{m+1}$ with normal $(-1, p)$. The motivation behind this classification is that it selects the columns that have positive reduced cost with respect to the dual solution p, or alternatively, it solves to optimality the Lagrangian relaxation using p as multipliers.

In this section, it will be important to have the additional assumption that the columns are in some sort of *general position*. (Given a positive integer k, we use $[k]$ as a shorthand for the set $\{1, 2, \ldots, k\}$.)

ASSUMPTION 1. *For all $p \in \mathbb{R}^m_+$ with $p \neq 0$, there are at most m different $t \in [n]$ such that $\pi_t = pa^t$.*

Typically this assumption is harmless: perturbing the input randomly by a tiny amount achieves this with probability one, and the effect of the perturbation is absorbed in the approximation guarantees (Agrawal et al. [1],

Devenur and Hayes [10]). But in order to ensure the correctness of our arguments, we keep this assumption explicitly.

2.1. Sampling LP's. To obtain a good dual solution p, we use the (random) LP consisting on the first s columns of (LP) with appropriately scaled right-hand side:

$$\begin{array}{l|l}
\max \ \sum_{t=1}^{s} \pi_{\sigma(t)} x_{\sigma(t)} \qquad (s,\delta)\text{-LP} & \min \ \left\{ \frac{s}{n}\delta B \sum_{i=1}^{m} p_i + \sum_{t=1}^{s} \alpha_{\sigma(t)} \right\} \qquad (s,\delta)\text{-Dual} \\
\sum_{t=1}^{s} a^{\sigma(t)} x_{\sigma(t)} \le \frac{s}{n}\delta B & p a^{\sigma(t)} + \alpha_{\sigma(t)} \ge \pi_{\sigma(t)} \quad t = 1,\dots,s \\
x_{\sigma(t)} \in [0,1] \quad t = 1,\dots,s. & p \ge 0 \\
 & \alpha \ge 0.
\end{array}$$

Here σ denotes the random permutation of the columns of the LP. We use $OPT(s,\delta)$ to denote the optimal value of (s,δ)-LP, and $OPT(s)$ to denote the optimal value of $(s,1)$-LP.

The static pricing algorithm OLA of Agrawal et al. [1] can then be described succinctly as follows. (To simplify the exposition, we assume that ϵn is an integer.)

1. Wait for the first ϵn columns of the LP (indexed by $\sigma(1), \sigma(2), \dots, \sigma(\epsilon n)$) and solve $(\epsilon n, 1-\epsilon)$-dual. Let (p,α) be the obtained dual optimal solution.
2. Use the classification given by p as above by setting $x_{\sigma(t)} = x(p)_{\sigma(t)}$ for $t = \epsilon n+1, \epsilon n+2, \dots$ as long as the solution obtained remains valid. From this point on set all further variables to zero.

Note that by definition this algorithm outputs a feasible solution with probability one. Our goal is then to analyze the quality of the solution produced, ultimately leading to the following theorem.

THEOREM 1. *Fix $\epsilon \in (0,1]$. Consider an instance of* (LP) *such that* (i) *Assumption 1 holds*; (ii) *there are $K \ge m$ one-dimensional subspaces of $\mathbb{R}^m$ containing the columns a^t's*; (iii) $B \ge \Omega((m/\epsilon^3)\log(K/\epsilon))$. *Then algorithm* OLA *returns a feasible solution with expected value at least* $(1-5\epsilon)OPT$.

Let $S = \{\sigma(1), \dots, \sigma(\epsilon n)\}$ be the (random) index set of the columns sampled by OLA. We use p^S to denote the optimal dual solution obtained by OLA; notice that p^S is completely determined by S. To simplify the notation, we also use x^S to denote $x(p^S)$.

Notice that, for all the scenarios where x^S is feasible, the solution returned by OLA is identical to x^S with its components $x^S_{\sigma(1)}, \dots, x^S_{\sigma(\epsilon n)}$ set to zero. Given this observation, we can actually focus on proving that x^S is a good solution.

LEMMA 1. *Fix $\epsilon \in (0,1]$. Suppose that* (i) *Assumption 1 holds*; (ii) *there are $K \ge m$ one-dimensional subspaces of $\mathbb{R}^m$ containing the columns a^t's*; (iii) $B \ge \Omega((m/\epsilon^3)\log(K/\epsilon))$. *Then with probability at least* $1-\epsilon$, *x^S is a feasible solution for* (LP) *with value at least* $(1-3\epsilon)OPT$.

To see how Theorem 1 follows from this, let $\mathscr{E}$ denote the event that x^S is feasible for (LP) with value at least $(1-3\epsilon)OPT$, which occurs with probability at least $(1-\epsilon)$. Notice that in any scenario in $\mathscr{E}$ we have $x_{\sigma(t)} = x^S_{\sigma(t)}$ for all $t > \epsilon n$. By the nonnegativity of the profits, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{n} \pi_{\sigma(t)} x_{\sigma(t)}\right] \ge \mathbb{E}\left[\sum_{t=1}^{n} \pi_{\sigma(t)} x_{\sigma(t)} \,\middle|\, \mathscr{E}\right] \Pr(\mathscr{E}) = \mathbb{E}\left[\sum_{t>\epsilon n} \pi_{\sigma(t)} x^S_{\sigma(t)} \,\middle|\, \mathscr{E}\right] \Pr(\mathscr{E}). \tag{1}$$

Again using the definition of $\mathscr{E}$, we have $\mathbb{E}[\sum_{t=1}^{n} \pi_{\sigma(t)} x^S_{\sigma(t)} \mid \mathscr{E}] \Pr(\mathscr{E}) \ge (1-4\epsilon)OPT$. Moreover, we have the inequality

$$\mathbb{E}\left[\sum_{t\le\epsilon n} \pi_{\sigma(t)} x^S_{\sigma(t)} \,\middle|\, \mathscr{E}\right] \Pr(\mathscr{E}) \le \mathbb{E}\left[\sum_{t\le\epsilon n} \pi_{\sigma(t)} x^S_{\sigma(t)}\right] \le \epsilon OPT$$

(see, e.g., Lemma 2.4 of Agrawal et al. [1]). Using linearity of expectation and combining the two previous bounds, we get

$$\mathbb{E}\left[\sum_{t>\epsilon n} \pi_{\sigma(t)} x^S_{\sigma(t)} \,\middle|\, \mathscr{E}\right] \ge (1-5\epsilon)OPT,$$

and the result follows from (1).

2.2. Connection to PAC learning. We assume from now on that $B \geq \Omega((m/\epsilon^3)\log(K/\epsilon))$. Let $\mathcal{X} = \{x(p)\colon p \in \mathbb{R}^m_+\} \subseteq \{0,1\}^n$ denote the set of all possible linear classifications of the LP columns that can be generated by OLA. With slight overload in the notation, we identify a vector $x \in \{0,1\}^n$ with the set $\{t \in [n]\colon x_t = 1\}$ indicated by it. The following definition is motivated by Lemma 1.

Definition 1 (Bad Solution). Given a scenario, we say that x^S is *bad* if it is either an infeasible solution for (LP) or has value less than $(1-3\epsilon)OPT$. We say that x^S is *good* otherwise.

As noted in previous work, since our decisions are made based on reduced costs it suffices to analyze the *budget occupation* (or complementary slackness) of the solution in order to understand its *value*. To make this precise, given $x \in \{0,1\}^n$ let $a_i(x) = \sum_{t\in x} a_i^t$ be its occupation of the ith budget and let $a_i^S(x) = (1/\epsilon)\sum_{t\in x\cap S} a_i^t$ be its appropriately scaled occupation of ith budget in the sampled LP (recall $|S| = \epsilon n$). For completeness, we present the proof of the following lemma in Appendix B, which can be seen as an approximate version of an observation on Lagrangian relaxation made by Everett in the early 1960s (Everett [14]) and is also related to the approximate complementary slackness conditions in Vazirani [27].

Lemma 2. *Consider a scenario where x^S satisfies* (i) *for all $i \in [m]$, $a_i(x^S) \leq B$ and* (ii) *for all $i \in [m]$ with $p_i^S > 0$, $a_i(x^S) \geq (1-3\epsilon)B$. Then x^S is good.*

Recall that the solution x^S is obtained by selecting the columns with positive reduced cost with respect to the optimal dual solution p^S. Therefore, it is intuitively clear that x^S resembles an optimal solution for $(\epsilon n, 1-\epsilon)$-LP and thus should (approximately) be feasible and satisfy complementary slackness conditions. Using the assumption that the input is in general position, this is made formal in the following lemma, whose proof is also deferred to Appendix B.

Lemma 3. *Suppose that Assumption 1 holds. Then in every scenario, x^S satisfies the following*: (i) *for all $i \in [m]$, $a_i^S(x^S) \leq (1-\epsilon)B$ and* (ii) *for every $i \in [m]$ with $p_i^S > 0$, $a_i^S(x^S) \geq (1-2\epsilon)B$.*

Given the properties of x^S guaranteed by Lemma 3, together with the observation that $a_i(x) = \mathbb{E}[a_i^S(x)]$ for all x, the idea is to use concentration inequalities to argue that the conditions in Lemma 2 hold with good probability. Although concentration of $a_i^S(x)$ for *fixed* x can be achieved via Chernoff-type bounds, the quantity $a_i^S(x^S)$ has undesired correlations; obtaining an effective bound is the main technical contribution of this paper.

Definition 2 (Badly Learned). For a given scenario, we say that $x \in \mathcal{X}$ can be *badly learned for budget i* if either (i) $a_i^S(x) \leq (1-\epsilon)B$ and $a_i(x) > B$ or (ii) $a_i^S(x) \geq (1-2\epsilon)B$ and $a_i(x) < (1-3\epsilon)B$.

Essentially these are the classifications that look good for the sampled $(\epsilon n, 1-\epsilon)$-LP but are actually bad for (LP). More precisely, Lemmas 2 and 3 give the following.

Observation 2. Consider a scenario for which x^S is bad. Then $x^S = x$ for some x that can be badly learned in this scenario for some budget $i \in [m]$.

This observation directly implies that

$$\Pr(x^S \text{ is bad}) \leq \Pr\left(\bigvee_{i\in[m],\, x\in\mathcal{X}} x \text{ can be badly learned for budget } i\right). \tag{2}$$

Notice that indeed the right-hand side of this inequality does not depend on x^S, it is only a function of how skewed $a_i^S(x)$ is as compared to its expectation $a_i(x)$ (over all $x \in \mathcal{X}$).

From this point on, usually the right-hand side in the previous equation is upper bounded by taking a union bound over all its terms (Agrawal et al. [1]). However, this strategy can be too wasteful, because if x and x' are "similar" there is a large overlap between the scenarios where $a_i^S(x)$ is skewed and those where $a_i^S(x')$ is skewed. To obtain improved guarantees, we introduce in the next section a new way of bounding the right-hand side of the above expression; we use something akin to a covering argument, although we need to use a suitable (and nonstandard) measure to capture the similarity between classifications.

2.3. Similarity via witnesses. First, we partition the classifications that can be badly learned for budget i (for some scenario) into two sets, depending on why they are bad: for $i \in [m]$, let $\mathcal{X}_i^+ = \{x \in \mathcal{X}\colon a_i(x) > B\}$ and $\mathcal{X}_i^- = \{x \in \mathcal{X}\colon a_i(x) < (1-3\epsilon)B\}$. To simplify the notation, given a set x we define $\text{skewm}_i(\epsilon, x)$ to be the event that $a_i^S(x) \leq (1-\epsilon)B$ and $\text{skewp}_i(\epsilon, x)$ to be the event that $a_i^S(x) \geq (1-2\epsilon)B$. Notice that if $x \in \mathcal{X}_i^+$, then $\text{skewm}_i(\epsilon, x)$ is the event that $a_i^S(x)$ is significantly smaller than its expectation (skewed in the minus direction),

whereas for $x \in \mathscr{X}_i^-$ $\text{skewp}_i(\epsilon, x)$ is the event that $a_i^S(x)$ is significantly larger than its expectation (skewed in the plus direction). These definitions directly give the equivalence

$$\Pr\left(\bigvee_{i,\, x\in\mathscr{X}} x \text{ can be badly learned for budget } i\right) = \Pr\left(\bigvee_{i,\, x\in\mathscr{X}_i^+} \text{skewm}_i(\epsilon, x) \vee \bigvee_{i,\, x\in\mathscr{X}_i^-} \text{skewp}_i(\epsilon, x)\right). \tag{3}$$

To introduce the concept of witnesses, consider two sets x, x', say, in $\mathscr{X}_i^+$. Take a subset $w \subseteq x \cap x'$; the main observation is that, since $a^t \geq 0$ for all t, for all scenarios we have $a_i^S(w) \leq a_i^S(x)$ and $a_i^S(w) \leq a_i^S(x')$. In particular, the event $\text{skewm}_i(\epsilon, x) \vee \text{skewm}_i(\epsilon, x')$ is contained in $\text{skewm}(\epsilon, w)$. The set w serves as a witness for scenarios that are skewed for either x or x'; if additionally $a_i(w)$ are reasonably larger than $(1-\epsilon)B$, we can then use concentration inequalities over $\text{skewm}_i(\epsilon, w)$ in order to bound probability of $\text{skewm}(\epsilon, x) \vee \text{skewm}(\epsilon, x')$. This ability of bounding multiple terms of the right-hand side of (3) simultaneously is what gives an improvement over the naive union bound.

Definition 3 (Witness). We say that $\mathscr{W}_i^+$ is a *witness set* for $\mathscr{X}_i^+$ if (i) for all $w \in \mathscr{W}_i^+$, $a_i(w) \geq (1-\epsilon/2)B$ and (ii) for all $x \in \mathscr{X}_i^+$ there is $w \in \mathscr{W}_i^+$ contained in x. Similarly, we say that $\mathscr{W}_i^-$ is a *witness set* for $\mathscr{X}_i^-$ if (i) for all $w \in \mathscr{W}_i^-$, $a_i(w) \leq (1-3\epsilon/2)B$ and (ii) for all $x \in \mathscr{X}_i^-$ there is $w \in \mathscr{W}_i^-$ containing x.

As indicated by the previous discussion, given witness sets $\mathscr{W}_i^+$ and $\mathscr{W}_i^-$ for $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$, we directly get the bound

$$\Pr\left(\bigvee_{i,\, x\in\mathscr{X}_i^+} \text{skewm}_i(\epsilon, x) \vee \bigvee_{i,\, x\in\mathscr{X}_i^-} \text{skewp}_i(\epsilon, x)\right) \leq \Pr\left(\bigvee_{i,\, w\in\mathscr{W}_i^+} \text{skewm}_i(\epsilon, w) \vee \bigvee_{i,\, w\in\mathscr{W}_i^-} \text{skewp}_i(\epsilon, w)\right). \tag{4}$$

Using concentration inequalities, we can now bound the probability that x^S is bad in terms of the size of witnesses sets.

Lemma 4. *Suppose that, for all $i \in [m]$, there are witness sets for $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$ of size at most M. Then* $\Pr(x^S \text{ is bad}) \leq 8mM\exp(-\epsilon^3 B/33)$.

Proof. Combining Equations (2), (3), and (4) and union bounding over all terms in the disjunction, we have that

$$\Pr(x^S \text{ is bad}) \leq \sum_{i,\, w\in\mathscr{W}_i^+} \Pr(\text{skewm}_i(\epsilon,\, w)) + \sum_{i,\, w\in\mathscr{W}_i^-} \Pr(\text{skewp}_i(\epsilon, w)).$$

Thus, it suffices to show that for all $w \in \mathscr{W}_i^+$ (respectively, $w \in \mathscr{W}_i^-$), the event $\text{skewm}_i(\epsilon, w)$ (respectively, $\text{skewp}_i(\epsilon, w)$) occurs with probability at most $2\exp(-\epsilon^3 B/33)$. The following simple inequalities will be helpful:

$$\text{For } \epsilon, \alpha, \beta \geq 0, \quad \frac{1-\alpha\epsilon}{1+\beta\epsilon} \geq 1-(\alpha+\beta)\epsilon \qquad \text{and} \qquad \frac{1-\alpha\epsilon}{1-\beta\epsilon} \leq 1-(\alpha-\beta)\epsilon. \tag{5}$$

Take $w \in \mathscr{W}_i^+$. By definition of this set, $a_i(w) \geq (1-\epsilon/2)B$, so the event $\text{skewm}(\epsilon, w)$ is contained in the event that $a_i^S(w) \leq (1-\epsilon)a_i(w)/(1-\epsilon/2)$, which is contained in the event $a_i^S(w) \leq (1-\epsilon/2)a_i(w)$. Using a Chernoff-type bound (more explicitly, Corollary 1 with $\tau = \epsilon^2 a_i(w)/2$), we obtain that $\Pr(\text{skewm}(\epsilon, w)) \leq 2\exp(-\epsilon^3 B/33)$.

Similarly, take $w \in \mathscr{W}_i^-$, such that $a_i(w) \leq (1-3\epsilon/2)B$. It is easy to check that the event $\text{skewp}(\epsilon, w)$ is contained in $a_i^S(w) \geq (1+\epsilon/2)a_i(w)$, so using Corollary 1 with $\tau = \epsilon^2 B/2$ we get that $\Pr(\text{skewm}(\epsilon, w)) \leq 2\exp(-\epsilon^3 B/33)$. This concludes the proof of the lemma.

The usefulness of defining witnesses as such is of course contingent upon the ability of finding witness sets that are much smaller than $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$. One reasonable choice of a witness set for, say, $\mathscr{X}_i^+$ is the collection of all of its minimal sets; unfortunately, this may not give a witness set of small enough size. However, notice that a witness set need not be a subset of $\mathscr{X}_i^+$ (or even $\mathscr{X}$). Allowing elements outside $\mathscr{X}_i^+$ gives the flexibility of obtaining witnesses that are associated to multiple "similar" minimal elements of $\mathscr{X}_i^+$, which is effective in reducing the size of witness sets.

2.4. Good witnesses for almost one-dimensional columns. Given the previous lemma, our task is to find small witness sets. Unfortunately, when the (π_t, a^t)'s lie in a space of dimension at least 3, $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$ may contain many (dependent on n) disjoint sets (see Figure 1), which shows that in general we cannot find small

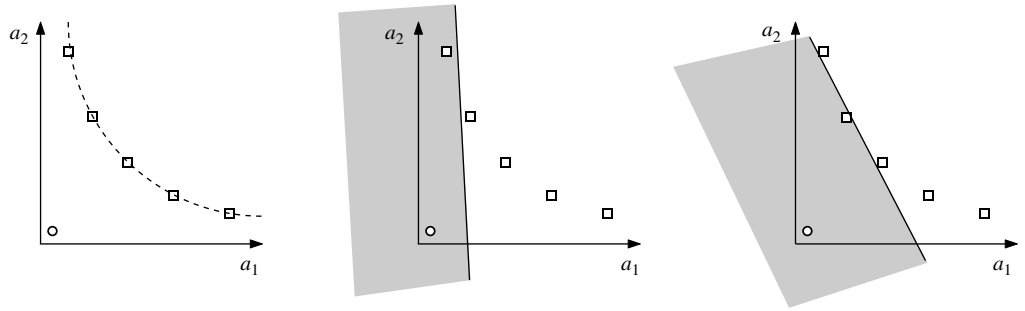

FIGURE 1. Example of LP with $m = 2$ constraints whose witness sets depend on the number of columns. Given a fixed B and ϵ, consider the LP where all the π_t's equal 1, and the columns a^t's are depicted in the first figure: the circle represents a cluster of columns arbitrarily close to each other and that has total occupation $(1-\epsilon)B$ of both budgets 1 and 2; each square represents a cluster of columns arbitrarily close to each other that has total occupation at least $2\epsilon B$ for some budget. The second and the third picture show two linear classifications which are bad (they belong to $\mathscr{X}_2^+$ for some 2), and whose intersection is the circle-cluster. Notice that there is no witness for both of these classifications simultaneously. Generalizing this LP by considering M clusters evenly spaced in the dashed semicircle of the first picture, we get that any witness set for $\mathscr{X}_i^+$ has $\Omega(M)$ elements, which can be made arbitrarily large.

witness sets directly. This sharply contrasts with the case where the (π_t, a^t)'s lie in a two-dimensional subspace of $\mathbb{R}^{m+1}$. In this case, it is not difficult to show that $\mathscr{X}$ is a union of two chains with respect to inclusion. In the special case where the a^t's lie in a one-dimensional subspace of $\mathbb{R}^m$, we show that $\mathscr{X}$ is actually a single chain (Lemma 6), and therefore we can take $\mathscr{W}_i^+$ as *the* minimal set of $\mathscr{X}_i^+$ and $\mathscr{W}_i^-$ as *the* maximal set of $\mathscr{X}_i^-$.

Because of the above observations, we focus on LPs whose a^t's lie in few one-dimensional subspaces. In this case, $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$ are sufficiently well behaved so that we can find small (independent of n) witness sets.

LEMMA 5. *Suppose that there are $K \geq m$ one-dimensional subspaces of $\mathbb{R}^m$ that contain the a^t's. Then there are witness sets for $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$ of size at most $(O((K/\epsilon)\log(K/\epsilon)))^m$.*

We use the rest of the section to prove this lemma. So assume its hypothesis and partition the index set $[n]$ into $C_1, C_2, \ldots, C_K$ such that for all $j \in [K]$ the columns $\{a^t\}_{t \in C_j}$ belong to the same one-dimensional subspace. Equivalently, for each $j \in [K]$ there is a vector c^j of l_∞-norm 1 such that for all $t \in C_j$ we have $a^t = \|a^t\|_\infty c^j$. An important observation is that now we can order the columns (locally) by the ratio of profit over budget occupation: without loss of generality assume that for all $j \in [K]$ and $t, t' \in C_j$ with $t < t'$, we have $\pi_t/\|a^t\|_\infty \geq \pi_{t'}/\|a^{t'}\|_\infty$ (notice that this ratio is well defined since by assumption $a^t \neq 0$ for all $t \in [n]$).

Given a classification x, we use $x|_{C_j}$ to denote its projection onto the coordinates in C_j; so $x|_{C_j}$ is the induced classification on columns with indices in C_j. Similarly, we define $\mathscr{X}|_{C_j} = \{x|_{C_j} : x \in \mathscr{X}\}$ as the set of all classifications induced by $\mathscr{X}$ in the columns in C_j.

Strengthening a previous observation, the main property that we get from working with one-dimensional subspaces is the following.

LEMMA 6. *For each $j \in [K]$, the sets in $\mathscr{X}|_{C_j}$ are prefixes of C_j.*

PROOF. Fix $j \in [K]$. Consider a set $x \in \mathscr{X}$ and let p be a dual vector such that $x(p) = x$. Let t' be the last index of C_j that belongs to $x|_{C_j}$; this implies that $\pi_{t'} > pa^{t'} = pc^j\|a^{t'}\|_\infty$, or alternatively $\pi_{t'}/\|a^{t'}\|_\infty > pc^j$. By the ordering of the columns, for all $t \in C_j$ smaller than t' we have $\pi_t/\|a^t\|_\infty \geq \pi_{t'}/\|a^{t'}\|_\infty > pc^j$ and hence $t \in x|_{C_j}$. By definition of t' it follows that $x|_{C_j} = \{t \in C_j : t \leq t'\}$, a prefix of C_j; this concludes the proof.

To simplify the notation fix $i \in [m]$ for the rest of this section, so we aim at providing witness sets for $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$. The idea is to group the classifications according to their budget occupation caused by the different column classes C_j's. To make this formal, start by covering the interval $[0, B+m]$ with intervals $\{I_l\}_{l \in L}$, where $I_0 = [0, \epsilon B/(4K))$ and $I_l = [(\epsilon B/(4K))(1+\epsilon/4)^{l-1}, (\epsilon B/(4K))(1+\epsilon/4)^l)$ for $l > 0$ and $L = \{0, \ldots, \lceil \log_{1+\epsilon/4}(8K/\epsilon)\rceil\}$ (note that since $B \geq m$, we have $B + m \leq 2B$). Define $\mathscr{B}_{i,j}^l$ as the set of partial classifications $y \in \mathscr{X}|_{C_j}$ whose budget occupation $a_i(y)$ lies in the interval I_l. For $v \in L^K$ define the family of classifications $\mathscr{B}_i^v = \{(y^1, y^2, \ldots, y^K) : y^j \in \mathscr{B}_{i,j}^{v_j}\}$. The $\mathscr{B}_i^v$'s then provide the desired grouping of the classifications. Note that the $\mathscr{B}_i^v$'s may include classifications not in $\mathscr{X}$ and may not include classifications in $\mathscr{X}$ that have occupation $a_i(\cdot)$ greater than $B + m$.

Now consider a nonempty $\mathscr{B}_i^v$. Let $\underline{w}_i^v$ be the inclusion-wise smallest element in $\mathscr{B}_i^v$. Notice that such unique smallest element exists: since $\mathscr{X}|_{C_j}$ is a chain, so is $\mathscr{B}_{i,j}^{v_j}$, and hence $\underline{w}_i^v$ is union (over j) of the smallest elements in the sets $\{\mathscr{B}_{i,j}^{v_j}\}_j$. Similarly, let $\bar{w}_i^v$ denote the largest element in $\mathscr{B}_i^v$. Intuitively, $\underline{w}_i^v$ and $\bar{w}_i^v$ will serve as witnesses for all the sets in $\mathscr{B}_i^v$.

Finally, define the witness sets by adding the $\underline{w}_i^v$ and $\bar{w}_i^v$'s of appropriate size corresponding to meaningful $\mathscr{B}_i^v$'s: set $\mathscr{W}_i^+ = \{\underline{w}_i^v\colon \mathscr{B}_i^v \cap \mathscr{X} \neq \varnothing, a_i(\underline{w}_i^v) \geq (1-\epsilon/2)B\}$ and $\mathscr{W}_i^- = \{\bar{w}_i^v\colon \mathscr{B}_i^v \cap \mathscr{X} \neq \varnothing, a_i(\bar{w}_i^v) \leq (1-3\epsilon/2)B\}$.

It is not too difficult to see that indeed, say, $\mathscr{W}_i^+$ is a witness set for $\mathscr{X}_i^+$: If $x \in \mathscr{X}_i^+$ belongs to some $\mathscr{B}_i^v$, then $\underline{w}_i^v$ belongs to $\mathscr{W}_i^+$ and is easily shown to be a witness for x. However, if x does not belong to any $\mathscr{B}_i^v$, by having too large $a_i(x)$, the idea is to find a smaller set $x' \subseteq x$ that belongs to some $\mathscr{B}_i^v$ *and* to $\mathscr{X}$, and then use $\underline{w}_i^v$ as a witness for x. We note that ignoring induced classifications with occupation larger than $B+m$ and ignoring B_i^v's that do not intersect $\mathscr{X}$ is very important for guaranteeing that $\mathscr{W}_i^+$ and $\mathscr{W}_i^-$ are small. The following lemma is proved formally in Appendix C.

Lemma 7. *The sets $\mathscr{W}_i^+$ and $\mathscr{W}_i^-$ are witness sets for $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$.*

2.5. Bounding the size of witness sets. Clearly these witness sets $\mathscr{W}_i^+$ and $\mathscr{W}_i^-$ have size at most $|L|^K$. Although this size is independent of n, it is still unnecessarily large since it only uses locally (for each C_j) the fact that $\mathscr{X}$ consists of linear classifications; in particular, it does not use the dimension of the ambient space $\mathbb{R}^m$. Suppose that $J \subseteq K$, of cardinality m, is such that the directions $\{c^j\}_{j\in J}$ form a basis of $\mathbb{R}^m$. Knowing the partial classification $x(p)|_{C_j}$, or more precisely the value of pc^j, for all $j \in J$ completely determines the whole classification $x(p)$. Similarly, knowing that $x(p)|_{C_j} \in \mathscr{B}_i^{v_j}$ for all $j \in J$ should give some information about which $\mathscr{B}_i^{v_j}$'s $x(p)|_{C_j}$ can belong to for $j \notin J$; this indicates that there are not enough degrees of freedom to allow a linear classification in $\mathscr{B}_i^v$ for each $v \in L^K$. The difficulty in making this argument formal is that the latter information does not completely determine which $\mathscr{B}_i^v$ the classification $x(p)$ belongs to. The idea is not to use a fixed set J of indices, but look at the whole K simultaneously.

Lemma 8. *At most $(O((K/\epsilon)\log(K/\epsilon)))^m$ of the $\mathscr{B}_i^v$'s contain an element from $\mathscr{X}$.*

Proof. To capture the fact that our classification is obtained via dual vectors in $\mathbb{R}^m$, we move from analyzing classifications to analyzing dual vectors. For $v \in L^K$ define P^v as the set of nonnegative dual vectors p such that $x(p)$ belongs to $\mathscr{B}_i^v$. It suffices to prove that at most $(O((K/\epsilon)\log(K/\epsilon)))^m$ of the families P^v's are nonempty. The main idea is to use that fact that the P^v's come from a hyperplane arrangement (Matoušek [23]) in $\mathbb{R}^m$.

To start, for $j \in [K]$ and $l \in L$ define $P_j^l = \{p \in \mathbb{R}_+^m\colon x(p)|_{C_j} \in \mathscr{B}_{i,j}^l\}$. Since $x(p) \in \mathscr{B}_i^v$ if and only if for all $j \in [K]$ we have $x(p)|_{C_j} \in \mathscr{B}_{i,j}^{v_j}$, it follows that $P^v = \bigcap_j P_j^{v_j}$. Let τ_j^l denote the first index in C_j such that the prefix $\{t \in C_j\colon t \leq \tau_j^l\}$ occupies the budget i to an extent in I_l. Using Lemma 6 and the fact that the a^t's are nonnegative, we get that $\mathscr{B}_{i,j}^l$ is the set of all prefixes of C_j that contain τ_j^l but do not contain τ_j^{l+1}. Moreover, notice that the set $x(p)|_{C_j}$ contains τ_j^l if and only if $\pi_{\tau_j^l} > pa^{\tau_j^l}$. It then follows from these observations we can express the set P_j^l using linear inequalities $P_j^l = \{p \in \mathbb{R}_+^m\colon \pi_{\tau_j^l} > pa^{\tau_j^l}, \pi_{\tau_j^{l+1}} \leq pa^{\tau_j^{l+1}}\}$. Since $P^v = \bigcap_j P_j^{v_j}$, we have that P^v is given by the intersection of half-spaces defined by hyperplanes of the form $\pi_{\tau_j^l} = pa^{\tau_j^l}$ and $p_i = 0$.

So consider the arrangement given by all hyperplanes $\{\pi_{\tau_j^l} = pa^{\tau_j^l}\}_{j\in[K],\, l\in L}$ and $\{p_i = 0\}_{i=1}^m$. Given a face F in this arrangement and a set P^v, either F is contained in P^v or these sets are disjoint. Since the faces of the arrangement cover $\mathbb{R}^m$, it follows that each nonempty P^v contains at least one these faces.

Notice that the arrangement is defined by $|K||L|^K m \leq O((Km/\epsilon)\log(K/\epsilon))$ hyperplanes, where the last inequality uses the fact that $\log(1+\epsilon/4) \geq \epsilon\log(1+1/4)$ holds (by concavity) for $\epsilon \in [0,1]$. It is known that an arrangement with $h \geq m$ hyperplanes in $\mathbb{R}^m$ has at most $((eh)/m)^m$ faces (see section 6.1 of Matoušek [23] and p. 82 of Matoušek and Nešetřil [24]). Using the conclusion of the previous paragraph, we get that there are at most $(O((K/\epsilon)\log(K/\epsilon)))^m$ nonempty P^v's and the result follows.

This lemma implies that $\mathscr{W}_i^+$ and $\mathscr{W}_i^-$ each has size at most $(O((K/\epsilon)\log(K/\epsilon)))^m$, which then proves Lemma 5. Finally, applying Lemma 4 we conclude the proof of Lemma 1.

3. Robust OLA. In this section we consider (LP) with columns that may not belong to few one-dimensional subspaces. Given the results of the previous section, the idea is clear: we would like to perturb the columns of this LP so that it belongs to few one-dimensional subspaces and such that an approximate solution for this perturbed LP is also an approximate solution for the original one. More precisely, we will obtain a set of vectors $Q \subseteq \mathbb{R}^m$ and transform the vector a^t into $\tilde{a}^t$, which is a scaling of a vector in Q, and we let the rewards π_t remain unchanged.

A basic but crucial observation is that solutions to an LP are robust to slight changes in the constraint matrix. The following lemma makes this precise and will guide us to obtaining the desired set Q.

Lemma 9. *Consider real numbers $\pi_1, \ldots, \pi_n$ and vectors $a^1, \ldots, a^n$ and $\tilde{a}^1, \ldots, \tilde{a}^n$ in $\mathbb{R}_+^m$ such that $\|\tilde{a}^t - a^t\|_\infty \leq (\epsilon/(m+1))\|a^t\|_\infty$. If x is an ϵ-approximate solution for (LP) with columns $(\pi_t, \tilde{a}^t)$ and right-hand side $(1-\epsilon)B$, then x is a 2ϵ-approximate solution for the LP (LP).*

Proof. Let LP1 denote the LP with columns $(\pi_t, \tilde{a}^t)$ and right-hand side $(1-\epsilon)B$ and LP2 denote the LP with columns (π_t, a^t) and right-hand side B. Let x be an ϵ-approximate solution for LP1. Notice that we can upper bound $\|a^t - \tilde{a}^t\|_\infty$ as a function of $\|\tilde{a}^t\|_\infty$:

$$\|\tilde{a}^t\|_\infty \geq \|a^t\|_\infty - \|a^t - \tilde{a}^t\|_\infty \geq \frac{m}{\epsilon}\|a^t - \tilde{a}^t\|_\infty,$$

where the first inequality follows from triangle inequality. That is, we have $\|a^t - \tilde{a}^t\|_\infty \leq (\epsilon/m)\|\tilde{a}^t\|_\infty$.

Given this bound, it is easy to see that x is feasible for LP2:

$$\sum_t a_i^t x_t \leq \sum_t (\tilde{a}_i^t + \|a_i^t - \tilde{a}_i^t\|)x_t \leq (1-\epsilon)B + \sum_t \|a^t - \tilde{a}^t\|_\infty x_t \leq (1-\epsilon)B + \frac{\epsilon}{m}\sum_t \|\tilde{a}^t\|_\infty x_t \leq B,$$

where the last inequality uses the fact that $\sum_t \|\tilde{a}^t\|_\infty x_t \leq \|\tilde{a}^t\|_1 x_t \leq mB$, since x is a feasible solution and the $\tilde{a}^t$'s are nonnegative.

To show that x is a 2ϵ-approximate solution for LP2, it suffices to show that the optimum of LP1 is at least $1/(1+\epsilon)$ times the optimum of the LP2, since then x will be within a factor of $(1-\epsilon)/(1+\epsilon) \geq (1-2\epsilon)$ the optimum of LP2. So let x^* be an optimal solution for LP2. Using the same argument as before, it is easy to see that $x^*/(1+\epsilon)$ is feasible for LP1; this concludes the proof of the lemma.

3.1. Perturbing the columns. To simplify the notation, set $\delta = \epsilon/(m+1)$; also, for simplicity of exposition we assume that $1/\delta$ is integral.

When constructing Q, we want the rays spanned by each of its vectors to be "uniform" over $\mathbb{R}^m_+$. Naturally, we focus on the intersection of these rays and the unit l_∞ sphere: we set Q to be a δ-net of the latter. More explicitly, we take Q to be the vectors in $\{\delta/2, \delta, 3\delta/2, \ldots, 1\}^m$ that have l_∞ norm 1. Note that all vectors in Q have *all* components strictly positive; also note that $|Q| = (O(m/\epsilon))^m$.

Given a vector a^t, we would like to set the transformed vector to be $q^t\|a^t\|_\infty$, where q^t is the vector in Q closest to $a^t/\|a^t\|_\infty$. However, the vectors obtained would not satisfy Assumption 1. Therefore, we actually set the transformed vector to be $\tilde{a}^t = q^t(\|a^t\|_\infty + \eta_t)$, where η_t is any continuous random variable with sufficiently small range (in particular, it is smaller than $\delta\|a^t\|_\infty/2$); also, we require these random variables to be mutually independent across all t's. Using the fact that $pq^t > 0$ for all $p \in \mathbb{R}^m_+$ different from 0, it follows from standard arguments that with probability 1 these transformed vectors satisfy Assumption 1 (Agrawal et al. [1], Devenur and Hayes [10]).

In addition, by definition of Q, for every vector $v \in \mathbb{R}^m$ of unit l_∞-norm, there is a vector $q \in Q$ with $\|v - q\|_\infty \leq \delta/2$. Using this observation, it follows that the vectors $\tilde{a}^t$ satisfy the property required in Lemma 9:

$$\|a^t - \tilde{a}^t\|_\infty \leq \|(a^t - q^t\|a^t\|_\infty)\|_\infty + |\eta_t|\|q^t\|_\infty = \|a^t\|_\infty \left\| \frac{a^t}{\|a^t\|_\infty} - q^t \right\|_\infty + |\eta_t|\|q^t\|_\infty \leq \delta\|a^t\|_\infty.$$

3.2. Algorithm robust OLA**.** One way to think of the algorithm robust OLA is that it works in two phases. First, it transforms the vectors a^t into $\tilde{a}^t$ as described above. Then it returns the solution obtained by running the algorithm OLA over the LP with columns $(\pi_t, \tilde{a}^t)$ and right-hand side $(1-\epsilon)B$. Notice that this algorithm can indeed be implemented to run in an online fashion.

Putting together the discussion in the previous paragraphs and the guarantee of OLA for almost one-dimensional columns given by Theorem 1 with $K = |Q| = (O(m/\epsilon))^m$, we obtain the following theorem.

Theorem 3. *Fix* $\epsilon \in (0, 1]$ *and suppose that* $B \geq \Omega((m^2/\epsilon^3)\log(m/\epsilon))$. *Then the algorithm robust* OLA *returns a solution to the online* (LP) *with expected value at least* $(1 - 10\epsilon)OPT$.

4. Algorithm (s, δ)**-OLA.** Our final algorithm robust DPA (as the algorithm DPA) can be seen as a combination of solutions to multiple sampled LPs, obtained via a modification of OLA denoted by (s, δ)-OLA. In this section we describe and analyze the algorithm (s, δ)-OLA.

This algorithm aims at solving the program $(2s, 1)$-LP and can be described as follows: it finds an optimal dual solution (p, α) for $(s, (1-\delta))$-LP and sets $x_{\sigma(t)} = x(p)_{\sigma(t)}$ for $t = s+1, s+2, \ldots, t' \leq 2s$ such that t' is the maximum one guaranteeing $\sum_{t=s+1}^{2s} a^{\sigma(t)} x_{\sigma(t)} \leq (s/n)B$ (for all other t's it sets $x_{\sigma(t)} = 0$).

The analysis of (s, δ)-OLA is similar to the one employed for OLA. The main difference is that this algorithm tries to approximate the value of the *random* LP $(2s, 1)$-LP. This requires a partition of the bad classifications, which is more refined than simply splitting into $\mathscr{X}_i^+$ and $\mathscr{X}_i^-$, and witness sets need to be redefined appropriately. Again let $S = \{\sigma(1), \sigma(2), \ldots, \sigma(s)\}$ be the random index set of the first s columns of the LP, let $T = \{\sigma(s+1), \sigma(s+2), \ldots, \sigma(2s)\}$ and $U = S \cup T$. We use π_U to denote the vector $(\pi_t)_{t \in U}$.

Lemma 10. *Fix an integer s and a real number $\delta \in (0, 1/10)$. Suppose that* (i) *Assumption 1 holds*; (ii) *there are $K \geq m$ one-dimensional subspaces of $\mathbb{R}^m$ containing the columns a^t's*; (iii) $\delta^2 sB/n \geq \Omega(m \ln(K/\delta))$. *Then algorithm (s, δ)-OLA returns a solution x satisfying $a_i^T(x) \leq B$ for all $i \in [m]$ with probability 1 and with expected value $\mathbb{E}[\pi_U x] \geq (1 - 3\delta)\mathbb{E}[OPT(2s)] - \mathbb{E}[OPT(s)] - \delta^2 OPT$.*

In the remaining part of the section we prove Lemma 10. Again we use p^S to denote the dual vector used by (s, δ)-OLA for its classification, and set $x^S = x(p^S)$. With slight abuse in the notation, we often see x^S as a (possibly infeasible) solution for $(2s, 1)$-LP, which means that we truncate the vector x^S to the first $2s$ coordinates $x^S_{\sigma(1)}, \ldots, x^S_{\sigma(2s)}$.

As before, we focus on proving the following lemma; the proof that this lemma implies Lemma 10 is presented at the end of this section.

Lemma 11. *Fix an integer s and a real number $\delta \in (0, 1/10)$. Suppose that* (i) *Assumption 1 holds*; (ii) *there are $K \geq m$ one-dimensional subspaces of $\mathbb{R}^m$ containing the columns a^t's*; (iii) $\delta^2 sB/n \geq \Omega(m \ln(K/\delta))$. *Then with probability at least $(1 - \delta^2)$, x^S satisfies $a_i^T(x^S) \leq B$ for all $i \in [m]$ and has value $\pi_U x^S \geq (1 - 3\delta)OPT(2s)$.*

In a given scenario, we now say that x^S is *bad* if $a_i^T(s^S) > B$ for some $i \in [m]$ or if $\pi_U x^S (1 - 3\delta)OPT(2s)$. In this scenario, now a classification $x \in \mathcal{X}$ can be *badly learned for budget i due to infeasibility* if $a_i^S(x) \leq (1 - \delta)B$ and $a_i^T(x) > B$; x can be *badly learned for budget i due to value* if $a_i^S(x) \geq (1 - 2\delta)B$ and $a_i^U(x) < (1 - 3\delta)B$. Then x can be *badly learned for budget i* if it falls into any of the above cases. The following is the appropriate modification of Lemma 2 for our current setting, and can be proved exactly in the same way.

Lemma 12. *Consider a scenario where x^S satisfies the following*: (i) *for all $i \in [m]$, $a_i^T(x^S) \leq B$ and* (ii) *for all $i \in [m]$ with $p_i^S > 0$, $a_i^U(x^S) \geq (1 - 3\delta)B$. Then x^S is good.*

Because of our definitions, this lemma implies that inequality (2) still hold.

4.1. Witness sets. In the analysis of OLA, each $x \in \mathcal{X}$ could be badly learned for budget i due to either infeasibility or (exclusively) due to value, which motivated the definitions of $\mathcal{X}_i^+$ and $\mathcal{X}_i^-$. Now the same x can be badly learned for budget i due to both conditions. Therefore, we introduce two different partitions of $\mathcal{X}$, which tells *why* a classification is unlikely to be badly learned because of the appropriate condition. That is, we define $\mathcal{X}_i^+ = \{x \in \mathcal{X}\colon a_i(x) > (1 - \delta)B + \delta B/2\}$ and $\mathcal{Y}_i^+ = \{x \in \mathcal{X}\colon a_i(x) \leq (1 - \delta)B + \delta B/2\}$ as the partition associated to the infeasibility condition and $\mathcal{X}_i^- = \{x \in \mathcal{X}\colon a_i(x) < (1 - 2\delta)B - \delta B/2\}$ and $\mathcal{Y}_i^- = \{x \in \mathcal{X}\colon a_i(x) \geq (1 - 2\delta)B - \delta B/2\}$ as the partition associated to the value condition. For example, $\mathcal{X}_i^-$ is the set of classifications that are unlikely to be infeasible because of a small $a_i(\cdot)$ value. Also, note that these classifications are all based on the total budget occupation rather than on the budget occupation in the first $2s$ columns only.

Given this more refined tagging of elements in $\mathcal{X}$, we also need to redefine witness sets. We say that $(\mathcal{W}_i^+, \mathcal{W}_i^-, \mathcal{Z}_i^+, \mathcal{Z}_i^-)$ are *witness sets* for $(\mathcal{X}_i^+, \mathcal{X}_i^-, \mathcal{Y}_i^-, \mathcal{Y}_i^+)$, respectively, if they satisfy the following:

$$
\begin{aligned}
w \in \mathcal{W}_i^+ &\;\Rightarrow\; a_i(w) \geq (1 - \delta)B + \frac{\delta B}{4}, & x \in \mathcal{X}_i^+ &\;\Rightarrow\; \exists\, w \in \mathcal{W}_i^+\colon w \subseteq x \\
w \in \mathcal{Z}_i^+ &\;\Rightarrow\; a_i(w) \geq (1 - 2\delta)B - \frac{3\delta B}{4}, & x \in \mathcal{Y}_i^- &\;\Rightarrow\; \exists\, w \in \mathcal{Z}_i^+\colon w \subseteq x \\
w \in \mathcal{W}_i^- &\;\Rightarrow\; a_i(w) \leq (1 - 2\delta)B - \frac{\delta B}{4}, & x \in \mathcal{X}_i^- &\;\Rightarrow\; \exists\, w \in \mathcal{W}_i^+\colon x \subseteq w \\
w \in \mathcal{Z}_i^- &\;\Rightarrow\; a_i(w) \leq (1 - \delta)B + \frac{3\delta B}{4}, & x \in \mathcal{Y}_i^+ &\;\Rightarrow\; \exists\, w \in \mathcal{W}_i^+\colon x \subseteq w.
\end{aligned}
$$

Again to simplify the notation, given a set x we define $\text{skewm}_i^S(\delta, x)$ to be the event that $a_i^S(x) \leq (1 - \delta)B$, $\text{skewp}_i^S(\delta, x)$ to be the event that $a_i^S(x) \geq (1 - \delta)B$, and similarly replacing the set S by the sets T and U. The following expression, which is the analogous to (3)–(4), establishes the connection between the events where classifications can be badly learned and witness sets:

$$
\begin{aligned}
\bigvee_{x \in \mathcal{X}} \{x \text{ can be badly learned for budget } i\} \subseteq &\left(\bigvee_{w \in \mathcal{W}_i^+} \text{skewm}_i^S(\delta, w)\right) \vee \left(\bigvee_{w \in \mathcal{Z}_i^+} \text{skewm}_i^U(3\delta, w)\right) \\
&\vee \left(\bigvee_{w \in \mathcal{W}_i^-} \text{skewp}_i^S(2\delta, w)\right) \vee \left(\bigvee_{w \in \mathcal{Z}_i^-} \text{skewp}_i^T(0, w)\right). \qquad (6)
\end{aligned}
$$

To see that this expression holds, take $x \in \mathscr{X}$. Suppose that $x \in \mathscr{X}_i^+$ and let $w \in \mathscr{W}_i^+$ be contained in x. Then the event $\{x$ can be badly learned for budget i due to infeasibility$\}$ is contained in $\text{skewm}_i^S(\delta, w)$. Similarly, if $x \in \mathscr{Y}_i^+$ let $w \in \mathscr{Z}_i^-$ contain x; then the event $\{x$ can be badly learned for budget i due to infeasibility$\}$ is contained in $\text{skewm}_i^T(0, w)$. The reasoning for the event $\{x$ can be badly learned for budget i due to value$\}$ is similar.

The following is analogous to Lemma 4 and is proved in the Appendix D.

LEMMA 13. *Suppose that, for all $i \in [m]$, there are witness sets for $(\mathscr{X}_i^+, \mathscr{X}_i^-, \mathscr{Y}_i^+, \mathscr{Y}_i^-)$ of size at most M. Then* $\Pr(x^S \text{ is bad }) \le 8mM \exp(-\delta^2 sB/(136n))$.

4.2. Good witness sets. We now construct witness sets of size at most $(O((K/\delta)\log(K/\delta)))^m$, so Lemma 11 will follow directly from Lemma 13. The development mirrors that of §2.4. Let $C_1, C_2, \ldots, C_K$ be a partition of the index set $[n]$ such that for all j, the columns $\{a^t\}_{t \in C_j}$ belong to the same one-dimensional subspace.

Cover the interval $[0, B+m]$ with intervals $\{I_l\}_{l \in L}$, where $I_0 = [0, \delta B/(8K))$ and $I_l = [(\delta B/(8K))(1+\delta/8)^{l-1}, (\delta B/(8K))(1+\delta/8)^l)$ for $l > 0$ and $L = \{0, \ldots, \lceil \log_{1+\delta/8}(16K)/\delta \rceil + 1\}$. Define $\mathscr{B}_{i,j}^l$ as the set of classifications $x \in \mathscr{X}|_{C_j}$ whose occupation $a_i(x)$ lies in the interval I_l. Finally, for $v \in L^K$, define the family of classifications $\mathscr{B}_i^v = \{(y^1, y^2, \ldots, y^K)\colon y^j \in \mathscr{B}_{i,j}^{v_j}\}$. Given $v \in L$, let $\underline{w}_i^v$ be the inclusion-wise smallest element in $\mathscr{B}_i^v$ and let $\bar{w}_i^v$ be the inclusion-wise largest element in $\mathscr{B}_i^v$.

Now we construct the witness sets as before. Set $\mathscr{W}_i^+ = \{\underline{w}_i^v\colon\ a_i(\underline{w}_i^v) \ge (1-\delta)B + \delta B/4, \mathscr{B}_i^v \cap \mathscr{X} \ne \varnothing\}$, set $\mathscr{Z}_i^+ = \{\underline{w}_i^v\colon\ a_i(\underline{w}_i^v) \ge (1-2\delta)B - 3\delta B/4, \mathscr{B}_i^v \cap \mathscr{X} \ne \varnothing\}$, set $\mathscr{W}_i^- = \{\bar{w}_i^v\colon\ a_i(\bar{w}_i^v) \le (1-2\delta)B - \delta B/4, \mathscr{B}_i^v \cap \mathscr{X} \ne \varnothing\}$, and finally set $\mathscr{Z}_i^- = \{\bar{w}_i^v\colon a_i(\bar{w}_i^v) \le (1-\delta)B + (3\delta B)/4, \mathscr{B}_i^v \cap \mathscr{X} \ne \varnothing\}$.

Following the same steps as in the proof of Lemma 7, one can check that $(\mathscr{W}_i^+, \mathscr{W}_i^-, \mathscr{Z}_i^+, \mathscr{Z}_i^-)$ are *witness sets* for $(\mathscr{X}_i^+, \mathscr{X}_i^-, \mathscr{Y}_i^+, \mathscr{Y}_i^-)$. Moreover, the proof of Lemma 8 can be used to show that, for a fixed $i \in [m]$, at most $(e(K/\delta)\log(K/\delta))^m$ of the $\mathscr{B}_i^v$'s contain an element of $\mathscr{X}$, which then imposes the same upper bound on the size of the witness sets. This concludes the proof of Lemma 11.

PROOF OF LEMMA 10. Let x be the solution returned by (s,δ)-OLA and let $\mathscr{E}$ denote the event that x^S is good. For any scenario in $\mathscr{E}$, we have $x_{\sigma(t)} = x^S_{\sigma(t)}$ for all $t = s+1, s+2, \ldots, 2s$. Therefore, we get that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{2s} \pi_{\sigma(t)} x_{\sigma(t)}\right] &\ge \mathbb{E}\left[\sum_{t=1}^{2s} \pi_{\sigma(t)} x_{\sigma(t)} \,\middle|\, \mathscr{E}\right]\Pr(\mathscr{E}) \ge \mathbb{E}\left[\sum_{t=1}^{2s} \pi_{\sigma(t)} x^S_{\sigma(t)} \,\middle|\, \mathscr{E}\right]\Pr(\mathscr{E}) - \mathbb{E}[OPT(s) \mid \mathscr{E}]\Pr(\mathscr{E}) \\ &\ge \mathbb{E}\left[\sum_{t=1}^{2s} \pi_{\sigma(t)} x^S_{\sigma(t)} \,\middle|\, \mathscr{E}\right]\Pr(\mathscr{E}) - \mathbb{E}[OPT(s)]. \qquad (7)\end{aligned}$$

To lower bound the first term in the right-hand side we again use the definition of $\mathscr{E}$:

$$\mathbb{E}\left[\sum_{t=1}^{2s} \pi_{\sigma(t)} x^S_{\sigma(t)} \,\middle|\, \mathscr{E}\right]\Pr(\mathscr{E}) \ge (1-3\delta)\,\mathbb{E}[OPT(2s) \mid \mathscr{E}]\Pr(\mathscr{E})$$

and

$$\mathbb{E}[OPT(2s)] = \mathbb{E}[OPT(2s) \mid \mathscr{E}]\Pr(\mathscr{E}) + \mathbb{E}[OPT(2s) \mid \bar{\mathscr{E}}]\Pr(\bar{\mathscr{E}}) \le \mathbb{E}[OPT(2s) \mid \mathscr{E}]\Pr(\mathscr{E}) + \delta^2 OPT,$$

where the last inequality uses Lemma 11. Combining the previous two inequalities give that $\mathbb{E}[\sum_{t=1}^{2s} \pi_{\sigma(t)} x^S_{\sigma(t)} \mid \mathscr{E}]$ $\Pr(\mathscr{E}) \ge (1-3\delta)\,\mathbb{E}[OPT(2s)] - \delta^2 OPT$, and the result follows from Equation (7).

5. Robust DPA. In this section we describe our final algorithm, which has an improved dependence on $1/\epsilon$. Following Agrawal et al. [1], the idea is to update the dual vector used in the classification as new columns arrive. More precisely, we use the first $2^i \epsilon n$ columns to classify columns $2^i \epsilon n + 1, \ldots, 2^{i+1}\epsilon n$. This leads to improved generalization bounds, which in turn give the reduced dependence on $1/\epsilon$.

5.1. Algorithm robust DPA. To simplify the description of the algorithm, we assume in this section that $\log(1/\epsilon)$ is an integer.

As in the algorithm robust OLA, robust DPA may be thought of in two phases. In the first phase it converts the vectors a^t into $\tilde{a}^t$, just as in the first phase of robust OLA. In the second phase, for $i = 0, \ldots, \log(1/\epsilon) - 1$, it runs $(\epsilon 2^i n, \sqrt{\epsilon/2^i})$-OLA over (LP) with columns $(\pi_t, \tilde{a}^t)$ and right-hand side $(1-\epsilon)B$ to obtain the solution x^i. The algorithm finally returns the solution x consisting of the "union" of x^i's: $x = \sum_i x^i$.

Note that the second phase corresponds exactly to using the first $\epsilon 2^i n$ columns to classify the columns $\epsilon 2^i n + 1, \ldots, \epsilon 2^{i+1} n$. This relative increase in the size of the training data for each learning problem allows us

to reduce the dependence of B on ϵ in each of the iterations, whereas the error from all the iterations telescope and are still bounded as before. Furthermore, notice that robust DPA can be implemented to run online.

The analysis of robust DPA reduces to that of (s, δ)-OLA. That is, using the definition of the parameters of (s, δ)-OLA used in robust DPA and Lemma 10, it is routine to check that the algorithm produces a feasible solution that has expected value $(1 - \epsilon)OPT$. The next theorem formally states the guarantees of our final algorithm DPA (a complete proof is presented in Appendix E).

THEOREM 4. *Fix* $\epsilon \in (0, 1/100)$ *and suppose that* $B \geq \Omega((m^2/\epsilon^2)\ln(m/\epsilon))$. *Then the algorithm robust* DPA *returns a solution to the online LP* (LP) *with expected value at least* $(1 - 50\epsilon)OPT$.

6. Open problems. A very interesting open question is whether the techniques introduced in this work can be used to obtain improved algorithms for generalized allocation problems (Feldman et al. [15]). The difficulty in this problem is that the classifications of the columns are not linear anymore; they essentially come from a conjunction of linear classifiers. Given this additional flexibility, having the columns in few one-dimensional subspaces does not seem to impose strong enough properties in the classifications. It would be interesting to find the appropriate geometric structure of the columns in this case.

Of course a direct open question is to improve the lower or upper bound on the dependence on the right-hand side B to obtain $(1 - \epsilon)$-competitive algorithms.

Appendix A. Bernstein inequality for sampling without replacement.

LEMMA 14 (THEOREM 2.14.19 IN VAN DER VAART AND WELLNER [26]). *Let* $Y = \{Y_1, \ldots, Y_n\}$ *be a set of real numbers in the interval* $[0, 1]$ *and let* $0 < \epsilon < 1$. *Let* S *be a random subset of* Y *of size* s *and let* $Y_S = \sum_{i \in S} Y_i$. *Setting* $\mu = (1/n)\sum_i Y_i$ *and* $\sigma^2 = (1/n)\sum_i (Y_i - \mu)^2$, *we have that for every* $\tau > 0$

$$\Pr(|Y_S - s\mu| \geq \tau) \leq 2\exp\left(-\frac{\tau^2}{2s\sigma^2 + \tau}\right).$$

Notice that, since the Y_i's belong to the interval $[0, 1]$, we can upper bound the variance by the mean as follows:

$$\sigma^2 \leq \frac{1}{n}\sum_i |Y_i - \mu| \leq \frac{1}{n}\left(\sum_i |Y_i| + \sum_i |\mu|\right) = 2\mu.$$

This gives the following corollary.

COROLLARY 1. *Consider the conditions of the previous lemma. Then for all* $\tau > 0$

$$\Pr(|Y_S - s\mu| \geq \tau) \leq 2\exp\left(-\frac{\tau^2}{4s\mu + \tau}\right)$$

Appendix B. Proof of Lemmas 2 and 3.

PROOF OF LEMMA 2. Fix a scenario σ for the duration of the proof. By assumption x^S is feasible for (LP), so it suffices to show that it attains a value of at least $(1 - 3\epsilon)OPT$. For that, consider (LP) with a modified right-hand side:

$$\max \sum_{t=1}^{n} \pi_t x_t$$

$$\sum_{t=1}^{n} a_i^t x_t \leq a_i(x^S) \quad \forall\, i \in [m] \tag{modLP}$$

$$x \in [0, 1]^n.$$

Consider the Lagrangian relaxation $L(p, x) = \sum_{t=1}^{\epsilon n} \pi_t x_t - \sum_{i=1}^{m} p_i(\sum_{t=1}^{\epsilon n} a_i^t x_t - a_i(x^S))$ and recall that $\inf_{p \in \mathbb{R}^m} \max_{x \in [0,1]^n} L(p, x)$ equals OPT(modLP), the optimum value of the LP (modLP). Notice that x^S is an optimal solution for $\max_{x \in [0,1]^n} L(p^S, x)$, which is then at least OPT(modLP). Since x^S is clearly feasible for (modLP), it follows that x^S is an optimal solution for the latter.

Now let x^* be an optimal solution for (LP). Since $a_i(x^S) \geq (1 - 3\epsilon)B$ for all i, and since $a^t \geq 0$ for all t, it follows that $(1 - 3\epsilon)x^*$ is feasible for (modLP). By linearity of the objective function we get that OPT(modLP) $\geq (1 - 3\epsilon)\sum_{t=1}^{n} \pi_t x_t^* = (1 - 3\epsilon)OPT$ and the result follows.

Proof of Lemma 3. Fix a scenario σ for the duration of the proof. Let x^* be an optimal solution for $(\epsilon n, (1-\epsilon))$-LP in complementary slackness with p^S. If $p^S a^t > \pi_t$, the corresponding constraint in the dual is loose and by complementary slackness we get $x_t^* = 0$. If $p^S a^t < \pi_t$, then for dual feasibility we have $\alpha_t^* > 0$ and by complementary slackness we have $x_t^* = 1$. Since we assumed the π_t's strictly positive, notice that $p^S = 0$ implies $x_t^* = 1$ for all t.

From the definition of x^S we get that $x^S \le x^*$ and, since the a^t's are nonnegative, the feasibility of x^* implies that $a_i^S(x^S) \le (1-\epsilon)B$ for all $i \in [m]$, which gives the first part of the lemma. To prove the second part, we claim that x^S and x^* differ in at most m positions: if $p^S = 0$ we get that $1 = x_t^S = x_t^*$ for all t; if $p^S \neq 0$, then Assumption 1 implies that there are at most m values of t such that $p^S a^t = \pi_t$, and the previous paragraph gives the claim. Therefore, from primal complementary slackness of the pair (x^*, p^S), we get that whenever $p^S > 0$, $a_i^S(x^S) \ge a_i^S(x^*) - m = (1-\epsilon)B - m \ge (1-2\epsilon)B$, where the last inequality follows from the fact that $B \ge 1/\epsilon$. This concludes the proof of the lemma.

Appendix C. Proof of Lemma 7. We prove that $\mathscr{W}_i^+$ is a witness set for $\mathscr{X}_i^+$; the proof that $\mathscr{W}_i^-$ is a witness set for $\mathscr{X}_i^-$ is analogous.

First, we claim that for all $x \in \mathscr{X}_i^+$, there is $x' \in \mathscr{X}$ such that $x' \subseteq x$ and $a_i(x') \in [B, B+m]$. To see this, let p be such that $x = x(p)$. For $\lambda \ge 0$, define $p^\lambda = p + \lambda e_i$, where e_i denotes the ith canonical vector. We have that $a_i(x(p^0)) > B$ (since $x(p) \in \mathscr{X}_i^+$) and $a_i(x(p^\infty)) = 0$ (since columns with $a_i^t > 0$ will have at some point $p^\lambda a^t \ge \pi_t$). Because of the assumption that the input is in general position, whenever $a_i(x(p^\lambda))$ is discontinuous (as a function of $\lambda \ge 0$) the right and the left limits differ by at most m. It then follows that there is $\lambda \ge 0$ such that $a_i(x(p^\lambda)) \in [B, B+m]$, and since $x(p^\lambda) \subseteq x$ for all $\lambda \ge 0$ the claim follows.

So take a classification $x \in \mathscr{X}_i^+$ and let x' be as above. The fact that $a_i(x') \le B + m$ and the nonnegativity of the a^t's implies that there is a $l \in L^K$ such that $x' \in \mathscr{B}_i^l$. Since $\underline{w}^l$ is the unique smallest set in $\mathscr{B}_i^l$, clearly $x' \subseteq \underline{w}^l$. To show that $\underline{w}^l \in \mathscr{W}_i^+$, it suffices to argue that $a_i(\underline{w}^l) \ge (1-\epsilon/2)B$.

Since $\underline{w}^l, x' \in \mathscr{B}_i^l$, for all j such that $l_j > 0$ we have $a_i(\underline{w}^l|_{C_j}) \ge a_i(x'|_{C_j})/(1+\epsilon/4)$. Moreover, for j such that $l = 0$ we have $a_i(x(p)|_{C_j}) < \epsilon B/(4K)$. Adding over all $j \in [K]$ gives

$$a_i(\underline{w}^l) \ge \frac{1}{1+\epsilon/4}\left[a_i(x(p)) - \sum_{j:\ l_j=0} a_i(x(p)|_{C_j})\right] \ge \frac{B}{1+\epsilon/4} - \frac{\epsilon B}{4} \ge \left(1 - \frac{\epsilon}{2}\right)B,$$

where the third inequality follows Equation (5). Thus, $\underline{w}^l \in \mathscr{W}_i^+$.

Since this property holds for all $x \in \mathscr{X}_i^+$, we conclude that $\mathscr{W}_i^+$ is a witness set for $\mathscr{X}_i^+$.

Appendix D. Proof of Lemma 13. For $w \in \mathscr{W}_i^+$, we can use Corollary 1 with $\tau = \delta s a_i(w)/(4n)$ to show that $\Pr(\text{skewm}^S(\delta, w)) \le 2\exp(-\delta^2 sB/(136n))$. For $w \in \mathscr{L}_i^+$, using this corollary with $\tau = \delta s a_i(w)/(2n)$ gives that $\Pr(\text{skewm}^U(3\delta, w)) \le 2\exp(-\delta^2 sB/(68n))$. For $w \in \mathscr{W}_i^-$, we can use this corollary with $\tau = \delta sB/(4n)$ to show that $\Pr(\text{skewp}^S(2\delta, w)) \le 2\exp(-\delta^2 sB/(68n))$. Finally, for $w \in \mathscr{L}_i^-$, using this corollary with $\tau = \delta sB/(4n)$ gives that $\Pr(\text{skewp}^T(0, w)) \le 2\exp(-\delta^2 sB/(68n))$.

Employing these bounds and union bounding over all terms in inequality (6) concludes the proof of the lemma.

Appendix E. Proof of Theorem 4. Let LP1 denote the LP with columns $(\pi_t, \tilde{a}^t)$ and right-hand side $\tilde{B} = (1-\epsilon)B$ and LP2 denote the LP with columns (π_t, a^t) and right-hand side B. We show that robust DPA returns a $(1 - 21.5\epsilon)$-approximation for LP1, and the theorem will follow from Lemma 9.

First we show that the returned solution x is feasible for LP1. By definition of the algorithm, $a_j(x^i) \le \epsilon 2^i \tilde{B}$ for all i, j. By linearity, $a_j(x) = \sum_i a_j(x^i) \le \epsilon \tilde{B} \sum_{i=0}^{\log(1/\epsilon)-1} 2^i \le \tilde{B}$.

To verify the value of the returned solution, we first show that $\delta^2 sB/n \ge \Omega(m \ln(K/\delta))$ in every call to (s, δ)-OLA made by robust DPA. As in §3, the columns $\tilde{a}^t$'s belong to at most $K = O(m/\epsilon)^m$ 1-dimensional subspaces. Since $B \ge \Omega((m^2/\epsilon^2)\ln(m/\epsilon))$, we have that for each $i = 0, \ldots, \log(1/\epsilon) - 1$ setting $s = \epsilon 2^i n$ and $\delta = \sqrt{\epsilon/2^i}$ satisfies the expression $\delta^2 sB/n \ge \Omega(m \ln(K/\delta))$.

Then applying Lemma 10 we get that for all $i = 0, \ldots, \log(1/\epsilon) - 1$, $\mathbb{E}[\pi x^i] \ge (1 - 3\sqrt{\epsilon/2^i})\mathbb{E}[OPT(\epsilon 2^{i+1} n)] - \mathbb{E}[OPT(\epsilon 2^i n)] - (\epsilon OPT)/2^i$. By linearity of the objective value and of expectations

$$\mathbb{E}[\pi x] = \sum_i \mathbb{E}[\pi x^i] \ge -\mathbb{E}[OPT(\epsilon n)] - \sum_{i=0}^{\log(1/\epsilon)-2} \left(3\sqrt{\frac{\epsilon}{2^i}}\right)\mathbb{E}[OPT(\epsilon n 2^{i+1})] + (1 - 3\sqrt{2}\epsilon - \epsilon)OPT.$$

Lemma 2.4 of Agrawal et al. [1] states that $\mathbb{E}[OPT(s)] \le (s/n)OPT$ for all $s \ge 0$. Employing this observation, we get

$$\mathbb{E}[\pi x] \ge OPT - \epsilon OPT\left[3\sqrt{2} + 2 + 3\sqrt{\epsilon} \sum_{i=0}^{\log(1/\epsilon)-2} 2^{i/2+1}\right].$$

Since the summation in the expression can be upper bounded by $((2\sqrt{2}^{\log(1/\epsilon)})/(\sqrt{2}-1)) \le 5/\sqrt{\epsilon}$, we get that $\mathbb{E}[\tilde{\pi} x] \ge (1 - 21.5\epsilon)OPT$. This concludes the proof of the theorem.

References

[1] Agrawal S, Wang Z, Ye Y (2013) A dynamic near-optimal algorithm for online linear programming. http://arxiv.org/abs/0911.2974.
[2] Babaioff M, Immorlica N, Kempe D, Kleinberg R (2007) A knapsack secretary problem with applications. *Approximation, Randomization, and Combinatorial Optimization*, Lecture Notes in Computer Science, Vol. 4627 (Springer-Verlag, Berlin, Heidelberg), 16–28.
[3] Babaioff M, Immorlica N, Kempe D, Kleinberg R (2008) Online auctions and generalized secretary problems. *SIGecom Exchanges* 7(2):1–11.
[4] Babaioff M, Dinitz M, Gupta A, Immorlica N, Talwar K (2009) Secretary problems: Weights and discounts. *Proc. 20th SODA* (SIAM, Philadelphia), 1245–1254.
[5] Bateni M, Hajiaghayi M, Zadimoghaddam M (2010) Submodular secretary problem and extensions. *Approximation, Randomization, and Combinatorial Optimization*, Lecture Notes in Computer Science, Vol. 6302 (Springer-Verlag, Berlin, Heidelberg), 39–52.
[6] Birge J, Louveaux F (1997) *Introduction to Stochastic Programming* (Springer, New York).
[7] Borodin A, El-Yaniv R (1998) *Online Computation and Competitive Analysis* (Cambridge University Press, Cambridge, UK).
[8] Buchbinder N, Naor J (2009) Online primal-dual algorithms for covering and packing. *Math. Oper. Res.* 34(2):270–286.
[9] Cucker F, Zhou D (2007) *Learning Theory: An Approximation Theory Viewpoint* (Cambridge University Press, Cambridge, UK).
[10] Devenur NR, Hayes TP (2009) The AdWords problem: Online keyword matching with budgeted bidders under random permutations. Fortnow L, Pu P, eds. *Proc. ACM Conf. Electronic Commerce* (ACM, New York), 71–78.
[11] Devanur NR, Jain K, Sivan B, Wilkens CA (2011) Near optimal online algorithms and fast approximation algorithms for resource allocation problems. Chen Y, Roughgarden T, eds. *Proc. ACM Conf. Electronic Commerce* (ACM, New York), 29–38.
[12] Devroye L, Wagner T (1979) Distribution-free performance bounds for potential function rules. *IEEE Trans. Inform. Theory* 25:601–604.
[13] Dynkin EB (1963) The optimum choice of the instant for stopping a Markov process. *Soviet Math. Dokl* 150(4):238–240.
[14] Everett H (1963) Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Oper. Res.* 11(3):399–417.
[15] Feldman J, Henzinger M, Korula N, Mirrokni VS, Stein C (2010) Online stochastic packing applied to display ad allocation. *Algorithms–ESA*, Lecture Notes in Computer Science, Vol. 6346 (Springer-Verlag, Berlin, Heidelberg), 182–194.
[16] Gilbert JP, Mosteller F (1966) Recognizing the maximum of a sequence. *J. Amer. Statist. Assoc.* 61(313):35–73.
[17] Goel G, Mehta A (2008) Online budgeted matching in random input models with applications to adwords. *Proc. 19th SODA* (SIAM, Philadelphia), 982–991.
[18] Im S, Wang Y (2011) Secretary problems: Laminar matroid and interval scheduling. *Proc. 22nd SODA* (SIAM, Philadelphia), 1265–1274.
[19] Karp RM, Vazirani UV, Vazirani VV (1990) An optimal algorithm for on-line bipartite matching. *Proc. 22nd STOC* (ACM, New York), 352–358.
[20] Kenyon C (1996) Best-fit bin-packing with random order. *Proc. 7th SODA* (SIAM, Philadelphia), 359–364.
[21] Kleinberg R (2005) A multiple-choice secretary algorithm with applications to online auctions. *Proc. 16th SODA* (SIAM, Philadelphia), 630–631.
[22] Kutin S, Niyogi P (2002) Almost-everywhere algorithmic stability and generalization error. *Proc. 18th Conf. Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco), 275–282.
[23] Matoušek J (2002) *Lectures on Discrete Geometry* (Springer-Verlag, New York).
[24] Matoušek J, Nešetřil J (1998) *Invitation to Discrete Mathematics* (Oxford University Press, Cary, NC).
[25] Soto JA (2011) Matroid secretary problem in the random assignment model. *Proc. 22nd SODA* (SIAM, Philadelphia), 1275–1284.
[26] van der Vaart A, Wellner JA (1996) *Weak Convergence and Empirical Processes* (Springer-Verlag, New York).
[27] Vazirani V (2001) *Approximation Algorithms* (Springer-Verlag, Berlin, Heidelberg).