

Approximating the Solution to Mixed Packing and Covering LPs in parallel $\tilde{O}(\epsilon^{-3})$ time

Michael W. Mahoney¹, Satish Rao², Di Wang², and Peng Zhang³

1 International Computer Science Institute and Department of Statistics, UC Berkeley, Berkeley, USA

`mmahoney@stat.berkeley.edu`

2 Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, USA

`wangd@eecs.berkeley.edu, satishr@berkeley.edu`

3 Department of Computer Science, Georgia Tech, Atlanta, USA

`pzhang60@gatech.edu`

Abstract

We study the problem of approximately solving positive linear programs (LPs). This class of LPs model a wide range of fundamental problems in combinatorial optimization and operations research, such as many resource allocation problems, solving non-negative linear systems, computing tomography, and single/multi commodity flows on graphs etc. For the special cases of pure packing or pure covering LPs, recent result by Allen-Zhu and Orecchia [2] gives $\tilde{O}(\frac{1}{\epsilon^3})$ -time parallel algorithm, which breaks the longstanding $\tilde{O}(\frac{1}{\epsilon^4})$ running time bound by the seminal work of Luby and Nisan [10].

We present new parallel algorithm with running time $\tilde{O}(\frac{1}{\epsilon^3})$ for the more general mixed packing and covering LPs, which improves upon the $\tilde{O}(\frac{1}{\epsilon^4})$ -time algorithm of Young [18, 19]. Our work leverages the ideas from both the optimization oriented approach [2, 17], as well as the more combinatorial approach with phases [18, 19]. In addition, our algorithm, when directly applied to pure packing or pure covering LPs, gives a improved running time of $\tilde{O}(\frac{1}{\epsilon^2})$.

1998 ACM Subject Classification F.2 ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY

Keywords and phrases Mixed packing and covering, Linear program, Approximation algorithm, Parallel algorithm

1 Introduction

Mixed packing and covering linear programs (LPs) are LPs formulated with non-negative coefficients, non-negative constraints and non-negative variables. They model a wide range of fundamental problems in combinatorial optimization and operations research, thus have long drawn interest in theoretical computer science [1, 2, 10, 18]. Notable special cases include pure packing LPs and pure covering LPs, which apply to most resource allocation problems, and can be formulated respectively as $\max_{x \geq 0} \{c^T x : Ax \leq b\}$ and $\min_{x \geq 0} \{c^T x : Ax \geq b\}$ where $c, A, b \geq 0$. More general than the pure packing and covering LPs, the mixed packing and covering LPs further capture problems requiring both packing and covering constraints, including solving non-negative linear systems, computing tomography, and single/multi commodity flows on graphs.

Formally, the mixed packing and covering LP is the optimization problem

$$\min\{\lambda : Px \leq \lambda p, Cx \geq c, x \geq 0\} \tag{1}$$

where P, C, p, c all have non-negative entries. A $(1 + \epsilon)$ -approximation is a feasible solution λ, x achieving $\lambda \leq (1 + \epsilon)\lambda_{\text{OPT}}$.



© Michael Mahoney and Satish Rao and Di Wang and Peng Zhang;
licensed under Creative Commons License CC-BY

Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Although one can use general LP solvers such as interior point method to solve packing and covering with convergence rate of $\log(\frac{1}{\epsilon})$ [5, 8, 9], such algorithms usually have very high per-iteration cost, as methods such as the computation of the Hessian and matrix inversion are involved. With the abundance of large-scale datasets, as well as the growing reliance on multiprocessors and cloud computing, low precision iterative solvers that can be highly parallelized are often more popular choices. Such parallel solvers compute approximate solutions usually in time with a poly-log dependence on the problem size, and nearly-linear total work, but they have $\text{poly}(\frac{1}{\epsilon})$ dependence on the approximation parameter ϵ .

Based on whether the running time depends on the width ρ , a parameter which typically depends on the dimension and the largest entry of A , these algorithms can be divided into width-dependent solvers and width-independent solvers. Width-dependent solvers are usually pseudo-polynomial, as the running time depends at least linearly on ρOPT , which itself can be large, while width-independent solvers are more efficient in the sense that they provide truly polynomial-time approximation solvers.

In this paper we focus on width-independent algorithms that produce $1 + \epsilon$ approximations in in $\text{poly}(\log n, \frac{1}{\epsilon})$ time and nearly-linear work. Time and work are standard notions from parallel algorithms that correspond to the longest chain of dependent operations and the total operations performed. In particular, time has a natural correspondence with iteration count, and these two measures have been used to study the performance of packing/covering LPs before [19]. Since the focus of this line of work is not on optimizing the log factors, we will follow the standard practice of using \tilde{O} to hide poly-log factors (which are at most $\log^3 n$) in our discussion ¹

1.1 Previous work

Most of the works on mixed packing and covering LPs, as well as the works on the special cases of pure packing and pure covering LPs, take one of the two approaches. The first approach is based on turning the constrained LPs into convex and smooth objective functions with trivial or no constraints. Approximately solving the LP is then reduced to approximately optimizing the smoothed function (See [13]), and general-purpose optimization schemes are usually applied directly. This approach traditionally gives algorithms that work in more general settings, but are width-dependent in the case of packing and covering (e.x. [3, 12, 13, 16]). Recent breakthroughs in [1, 2] leverage the insight from optimization ([21]) and the special structure of packing and covering problems, and get the first width-independent algorithms using first order optimization methods. In particular, [2] gives a parallel algorithm that takes $\tilde{O}(\frac{1}{\epsilon^3})$ time and $\tilde{O}(\frac{N}{\epsilon^3})$ work. Here N is the size of the input, i.e. the total number of non-zeros in the constraint matrices P and C . The result can be improved to run in $\tilde{O}(\frac{1}{\epsilon^2})$ time and $\tilde{O}(\frac{N}{\epsilon^2})$ work with simple modifications [17]. As to sequential algorithm, the remarkable result in [1] combines width-independence with Nesterov-like acceleration ([13, 14]), and gets a randomized algorithm with running time $\tilde{O}(\frac{N}{\epsilon})$. However, both results in [1, 2] are limited to pure packing and pure covering problems, and prior to our work no methods are able to obtain a time of $\tilde{O}(\frac{1}{\epsilon^3})$ or better for mixed covering and packing LPs [19].

The other approach is based on the Lagrangian-relaxation framework, where, similar to the optimization approach, certain hard constraints are replaced by a soft scalar-value penalty function, and the partial solution is iteratively updated to satisfy the remaining constraints while minimizing the increase of the penalty function. The analysis of the Lagrangian-relaxation algorithms are more of a combinatorial flavor comparing the the optimization schemes. Examples include the seminal work of Luby and Nisan [10], which gives the first width-independent algorithm for packing and covering, as

¹ $\tilde{O}(f(n))$ is often used to denote $O(f(n) \log^{O(1)}(f(n)))$, we modify it to include an additional factor of $\log^{O(1)} n$

well as subsequent works which improve the Luby and Nisan result in various ways [4,6,7,11,18,19]. Among them, only [11, 18, 19] works with mixed packing and covering LPs, while others only work in the pure packing and pure covering setting. In terms of parallel running time, the performance hasn't been improved since the result of Luby and Nisan, which gives $\tilde{O}(\frac{1}{\epsilon^4})$ depth. For algorithms working on mixed packing and covering, the results in [18, 19] have the best theoretical guarantee, with $\tilde{O}(\frac{1}{\epsilon^4})$ running time and $\tilde{O}(\frac{N}{\epsilon^2})$ total work. The result in [11] has worse bounds, but is stateless, which is a computational model on distributed algorithms with more restrictions on the processors (See [4]).

1.2 Our results

In this paper, we present a parallel algorithm that, given a mixed packing and covering LP with m variables and n total constraints, in $\tilde{O}(\frac{1}{\epsilon^3})$ iterations computes a $(1 + \epsilon)$ -approximate solution, or correctly reports the original mixed packing and covering LP is infeasible. The algorithm is deterministic and width-independent.

The bottleneck of each iteration is a matrix-vector multiplication, and can be implemented in $O(\log N)$ depth, in which case the running time of our algorithm is $\tilde{O}(\frac{1}{\epsilon^3})$. The total work of the algorithm we present in the paper is $\tilde{O}(\frac{N}{\epsilon^3})$. In particular, our result improves upon the current fastest parallel algorithm of mixed packing and covering LPs in [18, 19], where the running time is $\tilde{O}(\frac{1}{\epsilon^4})$. The work of the parallel algorithm in [18, 19] is $\tilde{O}(\frac{1}{\epsilon^2})$. We note that using a simple lazy update modification on the algorithm, which is the same technique used in [18, 19], we can reduce the work of our algorithm to $\tilde{O}(\frac{1}{\epsilon^2})$. Same as in [18, 19], this comes at the cost of requiring a centralized step in the parallel algorithm. Since the iteration count is the more interesting side of this line of work, we will not incorporate the lazy update in our algorithm for simplicity.

Furthermore, in the case of pure packing problem or pure covering problem, our algorithm allows a similar but simplified analysis, and will converge in $\tilde{O}(\frac{1}{\epsilon^2})$ iterations. This matches the running time achieved by the line of work [2, 17], but has the advantage of being deterministic and without centralized steps. We note that the technique we use in the analysis of the potential function in this work can be applied in a straightforward way to improve the result of [2] to have $\tilde{O}(\frac{1}{\epsilon^2})$ running time.

2 Technical overview

To compute $(1 + \epsilon)$ -approximation of a mixed packing and covering LP in the optimization form (1), via standard reduction and scaling (e.x. See [18]), it suffices to solve a $(1 + O(\epsilon))$ -feasibility problem as specified in (2) and (3).

At a high level, our work follows the Lagrangian-relaxation approach as in [18, 19]. In particular, we replace the hard packing and covering constraints with a scalar-valued potential function, which is continuous and smooth. The potential function measures how far away the current solution is from satisfying all the captured constraints. We then start with a \mathbf{x} of very small values, and iteratively increase \mathbf{x} by a carefully chosen increment vector Δ , which is guided by the gradients of the potential function. The increment vector Δ will keep driving down the potential function, until the function gets small enough, at which point we terminate, and conclude we either have an approximate solution or the original LP is infeasible.

Same as in [18, 19], each iteration a subset of the variables are picked based on the gradients of the potential function, and are increased within a local smooth region so we can bound the change of the potential function. However, unlike the parallel algorithm in [18, 19] that multiplicatively updates all variables in the subset with a uniform step size, we further incorporate the gradients into step sizes of individual variables' updates. This discriminative multiplicative step size allows more

aggressive updates on average, and is directly motivated by the line of works using gradient based optimization methods [1, 2, 17]. However, we move away from the optimization oriented view of these update steps in favor of more localized and adhoc analyses, which was developed to analyze direct adaptations of Young’s algorithm for purely-packing SDPs [15], leading to bounds similar to the optimization based approaches [20].

In particular, for each variable there will be a packing gradient capturing the packing part of the potential function, and there will be a covering gradient for the covering part of the potential function. The packing gradient captures the overall impact of a variable on the packing constraints, so a large packing gradient suggests the variable should not be increased. The covering gradient, on the other hand, gauges the impact of the variable on the covering constraints, and when large, it will be advantageous to increase the variable. To decide whether or not to increase a variable, we look at both gradients, and update only when the covering gradient is larger. Furthermore, to determine a variable’s update step size, we use the ratio of its two gradients.

Similar to Young’s algorithm [18], the overall progress of the algorithm is captured by how large the constraints become, which in turn depend on how large the variables are. A sufficient condition for the algorithm to terminate is when a variable is increased by more than a certain amount. To bound the number of iterations before any variable gets too large, we combine the notion of phases from [18] with the more refined analysis of gradient updates from more recent works [17]. This is owing to the clearer combinatorial structure of our interpretation of discriminative multiplicative steps

On a high level, the algorithm demonstrates the behavior that it will update a subset of variables for a while, until certain global progress is made, before shifting to other variables. The phases are created to capture these local windows, such that within a phase the algorithm only makes limited global progress. Since the update only increases variables, thus only increasing the values of the constraints, this translate to certain monotonicity-like property on the gradients within the interactions captured by a single phase. In [18], the phases are defined so that any variable being increased at the last iteration of a phase must have been increased in every iteration of the phase, which, coupled with a lower bound on each increase, gives a bound of the number of iterations in a phase. In our analysis, we significantly expand the phases to capture larger global progress, leading to a smaller number of phases.

The larger phases lead to a weaker monotonicity property on the gradients within a phase. To bound the number of iterations in our phase, we divide the iterations into two groups: some initial warm-up *bad* iterations followed by subsequent *good* iterations containing more interesting segments of the path to convergence (See Definition 1 for formal definitions). The bad iterations are ones where packing gradients are much smaller than covering gradients. Such steps create difficulties in the analysis, but an analysis identical to Young’s algorithm [18] shows that they must occur near the very start of a phase. In the subsequent good iterations, the packing gradients for all variables are relatively large comparing to their covering counterparts, so we only get weak signals as to which variables to move, and we can only move them by small steps. To bound the number of good iterations, we construct a dual solution from a suitable average of the packing and covering gradients over all the good iterations in a phase. Intuitively, if the primal LP is feasible, the dual solution certifies that there must be some key variable(s) we increase during the phase to achieve the fixed global progress. Particularly, we know that there is at least one variable that on average has smaller packing gradients than covering gradients. Moreover, since in the good iterations, the packing gradients are all at least on the same scale as the covering gradients, an argument in the spirit of Markov’s inequality then implies that the corresponding variable was increased by a large amount, which in turn leads to a bound on the number of iterations of a phase.

2.1 Remarks

We note that the phases in our result are virtual: we only need them in the analysis, but not in the actual execution of the algorithm. In particular, this modification to Young's algorithm [18] removes the dependency of updates on the phase of the overall algorithm as well as the current gradient. We believe this direct removal of phases also apply to other variants of Young's algorithm [19].

Furthermore, we believe that there is a more natural variant of the analysis that does not rely on phases, and treats all the iterations in a completely symmetric manner. Such an analysis is likely crucial for extending our results to the SDP setting, where the gradients exhibit much weaker monotonicity behaviors [20]. We are optimistic that it will lead to an $\tilde{O}(\frac{1}{\epsilon^2})$ bound for the mixed packing-covering case, which we believe is the more likely asymptotic behaviors of phase-less, gradient update variants of Young's algorithm [18, 19].

3 Parallel Algorithm for Mixed Packing and Covering LPs

3.1 Preliminaries

To compute $(1 + \epsilon)$ -approximation of a mixed packing and covering LP in the optimization form (1), via standard reduction and scaling (e.x. See [18]), it suffices to solve the following $(1 + O(\epsilon))$ -feasibility problem, that is, either find $\mathbf{x} \geq 0$ such that

$$0 < \max \mathbf{P}\mathbf{x} \leq (1 + \epsilon) \min \mathbf{C}\mathbf{x} \quad (2)$$

or conclude the following LP is infeasible

$$\begin{aligned} \mathbf{C}\mathbf{x} &\geq \mathbf{1} \\ \mathbf{P}\mathbf{x} &\leq (1 - 10\epsilon)\mathbf{1} \\ \mathbf{x} &\geq 0 \end{aligned} \quad (3)$$

We present our parallel $\tilde{O}(1/\epsilon^3)$ routine in Algorithm 1 for solving the $(1 + O(\epsilon))$ -feasibility problem above, that is, either find $\mathbf{x} \geq 0$ satisfying (2), or certify the infeasibility of (3). The input contains a packing constraint matrix $\mathbf{P} \in \mathbb{R}_{\geq 0}^{n_P \times m}$, a covering constraint matrix $\mathbf{C} \in \mathbb{R}_{\geq 0}^{n_C \times m}$, and an error parameter $0 < \epsilon$. That is, there are m variables, n_P packing constraints, and n_C covering constraints. We also use $n = n_P + n_C$ to denote the total number of constraints.

To certify that (3) is infeasible, we rely on the dual LP of (3).

► **Lemma 1.** *By duality, (3) is infeasible if there exists $\mathbf{y}, \mathbf{z} \geq 0$ s.t.*

$$(1 - 10\epsilon) \frac{\mathbf{C}^T \mathbf{z}}{\mathbf{1}^T \mathbf{z}} < \frac{\mathbf{P}^T \mathbf{y}}{\mathbf{1}^T \mathbf{y}}. \quad (4)$$

Proof. Eqn. (4) is a direct reformulation of the dual LP of (3). Since we only need the sufficient condition, the result is by weak duality. If there exists any $\mathbf{x} \geq 0$ satisfying (3), we have

$$\begin{aligned} (1 - 10\epsilon) \frac{\mathbf{x}^T \mathbf{C}^T \mathbf{z}}{\mathbf{1}^T \mathbf{z}} &\geq (1 - 10\epsilon) \frac{\mathbf{1}^T \mathbf{z}}{\mathbf{1}^T \mathbf{z}} = 1 - 10\epsilon, \\ \frac{\mathbf{x}^T \mathbf{P}^T \mathbf{y}}{\mathbf{1}^T \mathbf{y}} &\leq (1 - 10\epsilon) \frac{\mathbf{1}^T \mathbf{y}}{\mathbf{1}^T \mathbf{y}} = 1 - 10\epsilon. \end{aligned}$$

Together they give $1 < 1$, contradiction. ◀

As in [18], we use the soft-max $\text{lmax}(P\mathbf{x})$ and soft-min $\text{lmin}(C\mathbf{x})$ in our potential function

$$\text{lmax}(\mathbf{P}\mathbf{x}) = \ln\left(\sum_j \exp(\mathbf{P}\mathbf{x})_j\right) = \ln\left(\sum_j \exp(\mathbf{P}_j^T \mathbf{x})\right)$$

and

$$\text{lmin}(\mathbf{C}\mathbf{x}) = -\ln\left(\sum_j \exp(-\mathbf{C}\mathbf{x})_j\right) = -\ln\left(\sum_j \exp(-\mathbf{C}_j^T \mathbf{x})\right),$$

where $\mathbf{P}_j^T, \mathbf{C}_j^T$ are the j -th row of \mathbf{P}, \mathbf{C} respectively. In particular, these functions give smooth approximation to $\max \mathbf{P}\mathbf{x}$ and $\min \mathbf{C}\mathbf{x}$:

$$\begin{aligned} \max \mathbf{P}\mathbf{x} &\leq \text{lmax}(\mathbf{P}\mathbf{x}) \leq \max \mathbf{P}\mathbf{x} + \ln n \\ \min \mathbf{C}\mathbf{x} &\geq \text{lmin}(\mathbf{C}\mathbf{x}) \geq \min \mathbf{C}\mathbf{x} - \ln n. \end{aligned} \quad (5)$$

3.2 Algorithm

We start with small $\mathbf{x}_i^{(0)} = \frac{1}{m\|\mathbf{P}_{:,i}\|_\infty}, \forall i \in [m]$, and keep increasing \mathbf{x} properly, until it reaches the terminate condition in line 5, that is, $\max\{\max \mathbf{P}\mathbf{x}, \min \mathbf{C}\mathbf{x}\} \geq K = \frac{10\ln n}{\epsilon}$. The reason of the chosen K value is stated in Lemma 5. Roughly, when the difference of $\min \mathbf{C}\mathbf{x}$ and $\max \mathbf{P}\mathbf{x}$ becomes ϵ factor smaller than $\max\{\max \mathbf{P}\mathbf{x}, \min \mathbf{C}\mathbf{x}\}$, we know that $\max \mathbf{P}\mathbf{x} \leq (1+O(\epsilon)) \min \mathbf{C}\mathbf{x}$.

In each iteration of the while-loop, we first delete all covering constraints which has already reached K . Since \mathbf{x} never decreases, we know that once a row is deleted, we no longer need to look at it. Note the covering matrix cannot be empty, since we enter the iteration with $\min \mathbf{C}\mathbf{x} < K$. We compute the vectors \mathbf{y}, \mathbf{z} , which are exponentials of the values of the packing and covering constraints respectively. We then compute \mathbf{a} and \mathbf{b} , which can be considered as gradients of $\text{lmax}(\mathbf{P}\mathbf{x})$ and $\text{lmin}(\mathbf{C}\mathbf{x})$ respectively, and use them to guide our update on \mathbf{x} . In particular, we update x_i if $\mathbf{a}_i \leq (1 - \epsilon/50)\mathbf{b}_i$ (i.e. $i \in B$). Furthermore, we update x_i multiplicatively by a factor depends on the ratio of $\frac{\mathbf{a}_i}{\mathbf{b}_i}$, as specified in Eqn. (6) and line 13. Note that the smallest update in our algorithm is by a factor of $(1 + \Omega(\frac{\epsilon^2}{\ln n}))$, which is the same as the fixed update step size in [19], and in general our updates take larger steps.

Note that in our analysis, we equivalently view \mathbf{z} as the full n_C -dimensional vector, where the coordinates corresponding to deleted constraints are filled by 0's. In particular, the matrix-vector product of the original \mathbf{C} with the n_C -dimensional \mathbf{z} will be the same as the product of the reduced covering matrix $\mathbf{C}^{(t)}$ and reduced \mathbf{z} .

3.3 Proof of Correctness

In this section we will show Algorithm 1 will terminate, and output the correct answer.

Lemma 2 shows that empty B certifies the infeasibility of the input instance (3), which proves the correctness if we end up in the case of line 11.

► **Lemma 2.** *If the problem instance (3) is feasible, then*

$$\forall \mathbf{x} \geq 0, B = \{i : \mathbf{a}_i \leq (1 - \frac{\epsilon}{50})\mathbf{b}_i\} \neq \emptyset.$$

Proof. Let \mathbf{x}^* be a feasible solution of (3). Assume by contradiction, $\exists \mathbf{x} \geq 0, \forall i \in [m], \mathbf{a}_i^{(t)} > (1 - \frac{\epsilon}{50})\mathbf{b}_i^{(t)}$. By definition of \mathbf{a}, \mathbf{b} , it is equivalent to $\exists \mathbf{y}, \mathbf{z} \geq 0$ such that

$$\left(1 - \frac{\epsilon}{50}\right) \frac{\mathbf{C}^T \mathbf{z}}{\mathbf{1}^T \mathbf{z}} < \frac{\mathbf{P}^T \mathbf{y}}{\mathbf{1}^T \mathbf{y}}.$$

Then the result follows directly from Lemma 1. ◀

Algorithm 1 Parallel algorithm for mixed packing and covering LPs

Input: P, C, ϵ **Output:** "infeasible" or $\mathbf{x} \geq 0$ s.t. $P\mathbf{x} \leq (1 + \epsilon)C\mathbf{x}$

- 1: Let $K = \frac{10 \ln n}{\epsilon}$, $\alpha = \frac{1}{K}$, where n is the number of constraints.
 - 2: Initialize $\mathbf{x}_i^{(0)} = \frac{1}{m \|\mathbf{P}_{:,i}\|_\infty}$, $\forall i \in [m]$, where m is the number of variables.
 - 3:
 - 4: Let $t = 0$.
 - 5: **while** $\max P\mathbf{x} < K$ and $\min C\mathbf{x} < K$ **do**
 - 6: Let $C^{(t)}$ be C with rows j such that $(C\mathbf{x}^{(t)})_j \geq K$ deleted.
 - 7: Let $\mathbf{y}^{(t)} = \exp(P\mathbf{x}^{(t)})$, $\mathbf{z}^{(t)} = \exp(-C^{(t)}\mathbf{x}^{(t)})$.
 - 8:
 - 9: $\mathbf{a}^{(t)} = \frac{P^T \mathbf{y}^{(t)}}{\mathbf{1}^T \mathbf{y}^{(t)}}$, $\mathbf{b}^{(t)} = \frac{(C^{(t)})^T \mathbf{z}^{(t)}}{\mathbf{1}^T \mathbf{z}^{(t)}}$.
 - 10: Define $B^{(t)} = \{i : \mathbf{a}_i^{(t)} \leq (1 - \frac{\epsilon}{50})\mathbf{b}_i^{(t)}\}$.
 - 11: If $B^{(t)} = \emptyset$, then return "infeasible".
 - 12: Let

$$\Delta_i^{(t)} = \begin{cases} \frac{1}{2}(1 - \frac{\mathbf{a}_i^{(t)}}{\mathbf{b}_i^{(t)}}) \in [\epsilon/100, \frac{1}{2}] & \text{if } i \in B^{(t)} \\ 0 & \text{if } i \notin B^{(t)} \end{cases} \quad (6)$$
 - 13: $\mathbf{x}_i^{(t+1)} \leftarrow \mathbf{x}_i^{(t)}(1 + \alpha\Delta_i^{(t)})$.
 - 14: $t \leftarrow t + 1$.
 - 15: **end while**
 - 16: **return** $\mathbf{x} = \frac{\mathbf{x}^{(t)}}{K}$.
-

If Algorithm 1 doesn't terminate with line 11, it must increase at least one variable by at least a factor of $(1 + \frac{\epsilon^2}{10 \ln n})$ each iteration, so the algorithm must reach the termination condition of the while loop at some point, and we need to show the output \mathbf{x} satisfies (2). We consider the following potential function,

$$f(\mathbf{x}^{(t)}) = \text{lmax}(P\mathbf{x}^{(t)}) - \text{lmin}(C^{(t)}\mathbf{x}^{(t)}) = \ln(\mathbf{1}^T \mathbf{y}^{(t)}) + \ln(\mathbf{1}^T \mathbf{z}^{(t)}).$$

We first quantify the changes of lmax and lmin when we update the variables. This type of smoothness analysis is standard in analyzing algorithms that make updates using gradient information. Similar results are derived in other works on packing and covering (See [2, 18]). The particular analysis we develop can deal with larger gradient steps. In particular, the approach of our analysis allows updates that may move the gradients of some variables out of their respective coordinate-wise smooth regions, as long as we can still bound the combined impact on the potential function from updates of all variables. This approach can extend straightforwardly to show larger updates also work in [2], and improve their pure packing algorithm to run in $\tilde{O}(\frac{N}{\epsilon^2})$ iterations. Since the proof is technically tedious, we include the proof in Appendix A.

► **Lemma 3.** *At each iteration t ,*

$$\text{lmax}(P\mathbf{x}^{(t+1)}) \leq \text{lmax}(P\mathbf{x}^{(t)}) + \alpha \langle \mathbf{a}^{(t)}, (1 + \Delta^{(t)}) \cdot \Delta^{(t)} \cdot \mathbf{x}^{(t)} \rangle$$

and

$$\text{lmin}(C^{(t+1)}\mathbf{x}^{(t+1)}) \geq \text{lmin}(C^{(t)}\mathbf{x}^{(t)}) + \alpha \langle \mathbf{b}^{(t)}, (1 - \Delta^{(t)}) \cdot \Delta^{(t)} \cdot \mathbf{x}^{(t)} \rangle,$$

where $\Delta \cdot \mathbf{x}$ is the entry-wise product vector, i.e., $(\Delta \cdot \mathbf{x})_i = \Delta_i x_i$.

With the above bounds on the changes of the two components $\text{lmax}(\mathbf{P}\mathbf{x})$ and $\text{lmin}(\mathbf{C}\mathbf{x})$, we can show how our updates move the potential function $f(\mathbf{x})$.

► **Lemma 4.** *Given $\max \mathbf{P}\mathbf{x}^{(t)} < \frac{10 \ln n}{\epsilon}$ and $\min \mathbf{C}\mathbf{x}^{(t)} < \frac{10 \ln n}{\epsilon}$, we always have $f(\mathbf{x}^{(t)}) \leq 2 \ln n$ during the execution of Algorithm 1.*

Proof. Initially, $\mathbf{x}_i^{(0)} = \frac{1}{m \|\mathbf{P}_{:,i}\|_\infty}$, we have $\mathbf{P}\mathbf{x}^{(0)} \leq \mathbf{1}$ and $\mathbf{C}\mathbf{x} \geq 0$, thus $f(\mathbf{x}^{(0)}) \leq 2 \ln n$. To show $f(\mathbf{x}) \leq 2 \ln n$ for all iterations t before terminate, it suffices to show that $f(\mathbf{x})$ is non-increasing during the process. From Lemma 3,

$$\begin{aligned} f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)}) &\leq \alpha \langle \mathbf{a}, (1 + \Delta) \cdot \Delta \cdot \mathbf{x}^{(t)} \rangle - \alpha \langle \mathbf{b}, (1 - \Delta) \cdot \Delta \cdot \mathbf{x}^{(t)} \rangle \\ &= \sum_i \alpha \Delta_i x_i (\mathbf{a}_i (1 + \Delta_i) - \mathbf{b}_i (1 - \Delta_i)). \end{aligned}$$

For each $i \in [m]$, by our update rule (6), either Δ_i , or $\Delta_i = \frac{1}{2}(1 - \frac{\mathbf{a}_i}{\mathbf{b}_i})$, in which case

$$\mathbf{a}_i (1 + \Delta_i) - \mathbf{b}_i (1 - \Delta_i) = \frac{3\mathbf{a}_i \mathbf{b}_i - \mathbf{a}_i^2}{2\mathbf{b}_i} - \frac{\mathbf{a}_i + \mathbf{b}_i}{2} = \frac{2\mathbf{a}_i \mathbf{b}_i - \mathbf{a}_i^2 - \mathbf{b}_i^2}{2\mathbf{b}_i} \leq 0,$$

so all the summands are non-positive, thus $f(\mathbf{x})$ is non-increasing. ◀

The above lemma guarantees that the difference between $\text{lmax}(\mathbf{P}\mathbf{x})$ and $\text{lmin}(\mathbf{C}\mathbf{x})$ is bounded by $2 \ln n$, which by Eqn. (5) suggests $\max \mathbf{P}\mathbf{x} \leq \min \mathbf{C}\mathbf{x} + O(\ln n)$. Then when the two terms are large at termination, we are approximately feasible as the difference is a factor of ϵ smaller.

► **Lemma 5.** *If Algorithm 1 terminates with line 16, then it returns an $\mathbf{x} \geq 0$ with $0 < \max \mathbf{P}\mathbf{x} \leq (1 + \epsilon) \min \mathbf{C}\mathbf{x}$.*

Proof. Suppose the algorithm terminates at iteration T , that is, $\max \mathbf{P}\mathbf{x}^{(T)} \geq \frac{10 \ln n}{\epsilon}$ or $\min \mathbf{C}\mathbf{x}^{(T)} \geq \frac{10 \ln n}{\epsilon}$. Consider iteration $T - 1$, the covering matrix is not empty (otherwise, the algorithm terminates before iteration T). Since $\mathbf{x}^{(T)} = \mathbf{x}^{(T-1)} \cdot (1 + \alpha \Delta^{(T-1)}) \leq (1 + \frac{5\epsilon}{\ln n}) \mathbf{x}^{(T-1)}$, we have $\max \mathbf{P}\mathbf{x}^{(T-1)} \geq \frac{5 \ln n}{\epsilon}$ or $\min \mathbf{C}\mathbf{x}^{(T-1)} \geq \frac{5 \ln n}{\epsilon}$.

By Lemma 4,

$$\max \mathbf{P}\mathbf{x}^{(T-1)} \leq \text{lmax}(\mathbf{P}\mathbf{x}^{(T-1)}) \leq \text{lmin}(\mathbf{C}^{(T-1)} \mathbf{x}^{(T-1)}) + 2 \ln n \leq \min \mathbf{C}\mathbf{x}^{(T-1)} + 2 \ln n.$$

Since $2 \ln n \leq \epsilon \cdot \frac{5 \ln n}{\epsilon}$, we have

$$\max \mathbf{P}\mathbf{x}^{(T-1)} \leq (1 + \epsilon) \min \mathbf{C}\mathbf{x}^{(T-1)}.$$

This also gives $\max \mathbf{P}\mathbf{x}^{(T)} \leq (1 + \epsilon) \min \mathbf{C}\mathbf{x}^{(T)}$, since \mathbf{x}^T is within in multiplicative factor $1 + \frac{\epsilon}{10 \ln n}$ of \mathbf{x}^{T-1} . Since we start with $\mathbf{x} > 0$, and only increase \mathbf{x} , we also have $\max \mathbf{P}\mathbf{x} > 0$. So the \mathbf{x} we return at the end satisfies (2). ◀

3.4 Analysis of Convergence

So far we have proved that Algorithm 1 will terminate, and will either output \mathbf{x} satisfying (2) at the end, or terminate earlier and correctly certify (3) is infeasible. In this section we show that if (3) is feasible, Algorithm 1 must finish with the first case in $\tilde{O}(\frac{1}{\epsilon^3})$ iterations, so if the algorithm takes more than $\frac{1000 \ln n \ln(\frac{m}{\epsilon})}{\epsilon^3}$ iterations to complete, we can terminate it, and correctly output that (3) is infeasible.

We adapt the concept *phase* from Young's algorithm. Note phase is only used in our analysis, and our algorithm does not contain phase. Formally, phase s contains the iterations t such that

$$\frac{n_P}{n_C} \cdot 2^s \leq \frac{\mathbf{1}^T \mathbf{y}^{(t)}}{\mathbf{1}^T \mathbf{z}^{(t)}} < \frac{n_P}{n_C} \cdot 2^{s+1}$$

where n_P is the number of packing constraints and n_C is the number of covering constraints.

Since we only increase \mathbf{x} , $\frac{\mathbf{1}^T \mathbf{y}}{\mathbf{1}^T \mathbf{z}}$ is monotonically increasing, so each phase covers a consecutive sequence of iterations. Furthermore, as $\ln(\frac{\mathbf{1}^T \mathbf{y}}{\mathbf{1}^T \mathbf{z}}) = \ln \max(\mathbf{P}\mathbf{x}) + \ln \min(\mathbf{C}\mathbf{x})$ measures global progress towards termination, each phase captures a fixed amount of progress. From our definition of phases, and the termination condition, we have

► **Lemma 6.** *The total number of phases in Algorithm 1 is $O(\frac{\log n}{\epsilon})$.*

Proof. Since \mathbf{x} is monotonically increasing, $\mathbf{y} = \exp(\mathbf{P}\mathbf{x})$ and $\mathbf{z} = \exp(-\mathbf{C}\mathbf{x})$ are monotonically increasing and decreasing respectively, which implies that the quantity $\frac{\mathbf{1}^T \mathbf{y}}{\mathbf{1}^T \mathbf{z}}$ is monotonically increasing. Initially $\mathbf{P}\mathbf{x}^{(0)} \geq 0$, $\mathbf{C}\mathbf{x}^{(0)} \geq 0$, we know $\frac{\mathbf{1}^T \mathbf{y}^{(0)}}{\mathbf{1}^T \mathbf{z}^{(0)}} \geq \frac{n_P}{n_C}$. By the termination condition in Algorithm 1, the ratio never goes beyond $n_P \exp(\frac{10 \log n}{\epsilon})$. Therefore, the total number of phases is $O(\frac{\log n}{\epsilon})$. ◀

We now bound the number of iterations in a single phase. The iterations of a phase are divided into two groups, the bad iterations and the good iterations, formally defined as follows.

► **Definition 1.** If in an iteration t , we have for all i

$$\frac{\mathbf{a}_i^{(t)}}{\mathbf{b}_i^{(t)}} > \frac{1}{3}, \tag{7}$$

then we call it a *good* iteration. Otherwise we call it a *bad* iteration.

Note a phase may contain only bad iterations or only good iterations. We bound the total number of iterations in the two groups separately.

As discussed earlier, the bad iterations capture the initial warm-up iterations of a phase, where in any bad iteration, we can identify some variable x_i with a strong signal (i.e. $\frac{\mathbf{a}_i}{\mathbf{b}_i} \leq \frac{1}{3}$), so we can increase the variable by a lot. This restricts the warm-up sequence from getting too long, and we formalize the intuition in the following lemma.

► **Lemma 7.** *In a single phase, the number of bad iterations is at most $O(\ln n \ln(\frac{m}{\epsilon})/\epsilon)$.*

Proof. We will prove the result by showing that there cannot be any bad iteration after the initial $100 \ln n \ln(\frac{m}{\epsilon})/\epsilon$ iterations of a phase. By contradiction, if for any variable i , after $\Omega(\ln n \ln(\frac{m}{\epsilon})/\epsilon)$ iterations of a phase, we have at iteration t such that for some i ,

$$\frac{\mathbf{a}_i^{(t)}}{\mathbf{b}_i^{(t)}} = \frac{(\mathbf{P}^T \mathbf{y}^{(t)})_i}{\mathbf{1}^T \mathbf{y}^{(t)}} \frac{\mathbf{1}^T \mathbf{z}^{(t)}}{(\mathbf{C}^T \mathbf{z}^{(t)})_i} \leq \frac{1}{3},$$

then this ratio is at most $\frac{2}{3}$ in all previous iterations of this phase, since $\frac{(\mathbf{P}^T \mathbf{y})_i}{(\mathbf{C}^T \mathbf{z})_i}$ is monotonically increasing, and $2^s \leq \frac{\mathbf{1}^T \mathbf{y}}{\mathbf{1}^T \mathbf{z}} < 2^{s+1}$ in this phase. Equivalently, this is saying $\mathbf{a}_i \leq \frac{2}{3} \mathbf{b}_i$, so $i \in B$ in all previous $\Omega(\ln n \ln(\frac{m}{\epsilon})/\epsilon)$ iterations of the phase, and $\Delta_i \geq \frac{1}{6}$ in all those iterations.

Each iteration the multiplicative update on x_i is $(1 + \alpha \Delta_i)$, which is $(1 + \Theta(\frac{\epsilon}{10 \ln n}))$ since $\Delta_i \geq \frac{1}{6}$. As x_i starts with $\frac{1}{m \|P_{:,i}\|_\infty}$, after $100 \ln n \ln(\frac{m}{\epsilon})/\epsilon$ updates, we have $x_i \gg \frac{10 \ln n}{\epsilon \|P_{:,i}\|_\infty}$, which gives $\max \mathbf{P}\mathbf{x} \gg \frac{10 \ln n}{\epsilon}$, so the algorithm must have terminated. ◀

The above lemma guarantees that all iterations after the first $100 \ln n \ln(\frac{m}{\epsilon})/\epsilon$ must be good iterations, so we proceed to bound the number of these good iterations in a single phase. Without loss of generality, we index these good iterations in a phase as $1, \dots, T$ by shifting t .

We first identify one variable that must be updated extensively in these iterations.

► **Claim 8.** *Suppose the instance (3) is feasible, then There exists $i \in [m]$ such that*

$$\sum_{t=1}^T \mathbf{b}_i^{(t)} - \mathbf{a}_i^{(t)} \geq 10\epsilon \sum_{t=1}^T \mathbf{b}_i^{(t)}. \quad (8)$$

Proof. Define $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ to be the sum of the normalized gradients of iterations $1, \dots, T$, that is,

$$\bar{\mathbf{y}} = \sum_{t=1}^T \frac{\mathbf{y}^{(t)}}{\mathbf{1}^T \mathbf{y}^{(t)}}, \quad \bar{\mathbf{z}} = \sum_{t=1}^T \frac{\mathbf{z}^{(t)}}{\mathbf{1}^T \mathbf{z}^{(t)}}.$$

Note $\mathbf{1}^T \bar{\mathbf{y}} = \mathbf{1}^T \bar{\mathbf{z}} = T$. Recall $\mathbf{a}_i^{(t)}$ and $\mathbf{b}_i^{(t)}$ are respectively $\frac{(\mathbf{P}^T \mathbf{y}^{(t)})_i}{\mathbf{1}^T \mathbf{y}^{(t)}}$ and $\frac{(\mathbf{C}^T \mathbf{z}^{(t)})_i}{\mathbf{1}^T \mathbf{z}^{(t)}}$, then

$$\sum_{t=1}^T \mathbf{a}_i^{(t)} = \frac{T(\mathbf{P}^T \bar{\mathbf{y}})_i}{\mathbf{1}^T \bar{\mathbf{y}}}, \quad \sum_{t=1}^T \mathbf{b}_i^{(t)} = \frac{T(\mathbf{C}^T \bar{\mathbf{z}})_i}{\mathbf{1}^T \bar{\mathbf{z}}}.$$

Assume by contradiction, $\forall i \in [m], \sum_{t=1}^T \mathbf{a}_i^{(t)} > (1 - 10\epsilon) \sum_{t=1}^T \mathbf{b}_i^{(t)}$, that is,

$$\frac{\mathbf{P}^T \bar{\mathbf{y}}}{\mathbf{1}^T \bar{\mathbf{y}}} > (1 - 10\epsilon) \frac{\mathbf{C}^T \bar{\mathbf{z}}}{\mathbf{1}^T \bar{\mathbf{z}}}.$$

By Lemma 1, $\bar{\mathbf{y}}, \bar{\mathbf{z}}$ certify infeasibility of the instance (3), which contradicts the assumption. ◀

The above claim gives us a variable that on average has smaller packing gradients than covering gradients in this iteration. Together with the property we have on the good iterations (7), we can bound the number of good iterations.

► **Lemma 9.** *In a single phase, the number of good iterations is at most $O(\ln n \ln(\frac{m}{\epsilon})/\epsilon^2)$.*

Proof. Let x_i be a variable satisfying Eqn. (8). We want to turn Eqn. (8) into some lower bound on the total multiplicative update on x_i through these iterations. Intuitively, a bad case is that in some iteration t , $\mathbf{a}_i^{(t)}, \mathbf{b}_i^{(t)}$ are much larger than the values in other iterations, since they can dominate the terms from other iterations in Eqn. (8), but not much to the total update of x_i , since their ratio is what matters to the update. However, since we are inside one single phase, and only looking at good iterations, we can show the bad scenario will not show up.

Formally, let $l = \mathbf{a}_i^{(1)}$ and $u = \mathbf{b}_i^{(1)}$. Since $(\mathbf{P}^T \mathbf{y})_i$ monotonically increases, and $\mathbf{1}^T \mathbf{y}$ will increase but not by more than a factor of 2 in a phase, we have

$$\mathbf{a}_i^{(t)} = \frac{(\mathbf{P}^T \mathbf{y}^{(t)})_i}{\mathbf{1}^T \mathbf{y}^{(t)}} \geq l/2 \quad \forall t = 1, \dots, T \quad (9)$$

Similarly, we have

$$\mathbf{b}_i^{(t)} = \frac{(\mathbf{C}^T \mathbf{z}^{(t)})_i}{\mathbf{1}^T \mathbf{z}^{(t)}} \leq 2u \quad \forall t = 1, \dots, T \quad (10)$$

Furthermore, since we are looking at the good iterations, we have

$$l \geq \frac{1}{3}u. \quad (11)$$

The inequalities above allow us to turn the difference-based guarantee from Eqn. (8) into lower bounds on ratios we need.

By the update (6), we have

$$\Delta_i^{(t)} \geq \frac{(1 - \frac{\epsilon}{50})\mathbf{b}_i^{(t)} - \mathbf{a}_i^{(t)}}{2\mathbf{b}_i^{(t)}}.$$

So we can lower bound the total update on \mathbf{x}_i as follows

$$\begin{aligned} \mathbf{x}_i^{(T)} &\geq \mathbf{x}_i^{(1)} \exp\left(\frac{\alpha \sum_t \Delta_i^{(t)}}{2}\right) \\ &= \mathbf{x}_i^{(1)} \exp\left(\frac{\alpha}{4} \sum_t \frac{(1 - \frac{\epsilon}{50})\mathbf{b}_i^{(t)} - \mathbf{a}_i^{(t)}}{\mathbf{b}_i^{(t)}}\right) \\ &\geq \mathbf{x}_i^{(1)} \exp\left(\frac{\alpha}{4} \sum_t \left(\frac{\mathbf{b}_i^{(t)} - \mathbf{a}_i^{(t)}}{2u} - \frac{\epsilon}{50}\right)\right) \end{aligned}$$

where we used (10) in the last line.

From Eqn. (8), we have

$$\begin{aligned} \sum_t \mathbf{b}_i^{(t)} - \mathbf{a}_i^{(t)} &\geq 10\epsilon \sum_i \mathbf{b}_i^{(t)} \\ &\geq \frac{10\epsilon}{1 - 10\epsilon} \sum_t \mathbf{a}_i^{(t)} \\ &\geq \frac{\epsilon u T}{1 - 10\epsilon} \geq \epsilon u T. \end{aligned}$$

The first two lines both follow from Eqn. (8), the next line follows from $\mathbf{a}_i^{(t)} \geq l/2 \geq u/6$.

Thus

$$\mathbf{x}_i^{(T)} \geq \mathbf{x}_i^{(1)} \exp\left(\frac{\epsilon \alpha T}{8} - \frac{\epsilon \alpha T}{200}\right) \geq \frac{1}{m \|\mathbf{P}_{:,i}\|_\infty} \exp\left(\frac{\epsilon \alpha T}{10}\right).$$

If $T \geq \frac{100 \ln n \ln \frac{m}{\epsilon}}{\epsilon^2} \geq \frac{100 \ln \frac{m}{\epsilon}}{\epsilon \alpha}$, we have $\mathbf{x}_i^{(T)} \gg \frac{10 \ln n}{\epsilon \|\mathbf{P}_{:,i}\|_\infty}$. So the algorithm must have terminated since $\max \mathbf{P} \mathbf{x} \gg \frac{10 \ln n}{\epsilon}$. \blacktriangleleft

Lemma 7 and Lemma 9 bound the total number of iterations in a phase by $\tilde{O}(\frac{1}{\epsilon^2})$, together with the bound on the number of phases, which is $\tilde{O}(\frac{1}{\epsilon})$, we guarantee the total number of iterations in Algorithm 1 is $\tilde{O}(\frac{1}{\epsilon^3})$ if the LP in (3) is feasible.

► Theorem 10. *Algorithm 1 solves the $(1+\epsilon)$ -feasibility problem correctly. It runs in parallel time $\tilde{O}(1/\epsilon^3)$ with the total work $\tilde{O}(N/\epsilon^3)$, where N is the number of non-zero entries in the constraint matrix.*

Proof. The correctness and convergence follows from the lemmas in the prior sections. We only need to look at the running time and total work.

At each iteration, we compute all updated values in $O(\log N)$ parallel time. Since the total number of iterations is $\tilde{O}(\frac{1}{\epsilon^3})$, the algorithm terminates in parallel time $\tilde{O}(\frac{1}{\epsilon^3})$.

To see the total work, consider the following implementation. For each $i \in [m]$, we maintain $\mathbf{P}_{ji} \mathbf{x}_i$ if $\mathbf{P}_{ij} \neq 0$; similarly we maintain $\mathbf{C}_{ji} \mathbf{x}_i$ if $\mathbf{C}_{ij} \neq 0$. Besides, we maintain the values of $\mathbf{y}, \mathbf{z}, \mathbf{P}^T \mathbf{y}, \mathbf{C}^T \mathbf{z}, \mathbf{1}^T \mathbf{y}$ and $\mathbf{1}^T \mathbf{z}$. When we update \mathbf{x}_i , we update these values accordingly, with

work proportional to the number of non-zero entries in the i th column of the constraint matrix. For each fixed variable x_i , the total time of updates is at most $\tilde{O}(\frac{1}{\epsilon^2})$. Thus, the work on this part is $\tilde{O}(\frac{N}{\epsilon^2})$. Additionally, we need to compute the $\mathbf{a}_i, \mathbf{b}_i$ for all variables at the beginning of each iteration to determine which variables to update, this takes $\tilde{O}(N)$ work each iteration, so the total work is $\tilde{O}(\frac{N}{\epsilon^3})$.

We see the majority of the work is actually on computing the gradients for the variables we may not update. We point out that we can implement the same lazy update as in [19], which on a high level is just that if a variable has a large $\frac{a_i}{b_i}$ in an iteration, and is not updated, we don't recompute its gradients, until $\frac{1^T y}{1^T z}$ grows by more than a factor of $1 + \epsilon$. This can reduce the work to $\tilde{O}(\frac{N}{\epsilon^2})$, but requires a centralized step to control the phases. We omit the details as it is a straightforward adaptation. ◀

4 Pure Packing and Pure Covering LPs

We point out that in the case of pure packing or pure covering LPs, Algorithm 1 converges in $\tilde{O}(\frac{1}{\epsilon^2})$ iterations. This improves upon the result of [17], since our algorithm is deterministic, and doesn't need centralized steps. We include the details in Appendix B.

Acknowledgments. We thank Richard Peng for helpful discussions. DW was supported by ARO Grant W911NF-12-1-0541, SR was funded by NSF Grant CCF-1118083, and MM acknowledges the support of the NSF, AFOSR, and DARPA.

Appendix A Missing Proof

► **Lemma 3.** *At each iteration t ,*

$$\text{lmax}(\mathbf{P}\mathbf{x}^{(t+1)}) \leq \text{lmax}(\mathbf{P}\mathbf{x}^{(t)}) + \alpha \langle \mathbf{a}^{(t)}, (1 + \Delta^{(t)}) \cdot \Delta^{(t)} \cdot \mathbf{x}^{(t)} \rangle$$

and

$$\text{lmin}(\mathbf{C}^{(t+1)}\mathbf{x}^{(t+1)}) \geq \text{lmin}(\mathbf{C}^{(t)}\mathbf{x}^{(t)}) + \alpha \langle \mathbf{b}^{(t)}, (1 - \Delta^{(t)}) \cdot \Delta^{(t)} \cdot \mathbf{x}^{(t)} \rangle,$$

where $\Delta \cdot \mathbf{x}$ is the entry-wise product vector, i.e., $(\Delta \cdot \mathbf{x})_i = \Delta_i x_i$.

Proof. To simplify, we omit superscript (t) in the proof.

$$\begin{aligned} \text{lmax}(\mathbf{P}\mathbf{x}^{(t+1)}) &= \ln \sum_j \exp(\mathbf{P}_j^T(\mathbf{x} + \alpha \Delta \cdot \mathbf{x})) \\ &= \ln \sum_j \exp(\mathbf{P}_j^T \mathbf{x}) \cdot \exp(\alpha \mathbf{P}_j^T(\Delta \cdot \mathbf{x})) \\ &\leq \ln \sum_j \exp(\mathbf{P}_j^T \mathbf{x}) (1 + \alpha \mathbf{P}_j^T(\Delta \cdot \mathbf{x}) + \alpha^2 (\mathbf{P}_j^T(\Delta \cdot \mathbf{x}))^2). \end{aligned}$$

Recall \mathbf{P}_j^T is the j -th row of \mathbf{P} (i.e. the j -th packing constraint). The last inequality is by Taylor expansion, with $\epsilon \leq \Delta_i \leq \frac{1}{2}$, $\alpha = \epsilon/10 \log n$, and $\mathbf{P}_j^T \mathbf{x} \leq 10 \log n / \epsilon$ for all j , so $\alpha \mathbf{P}_j^T(\Delta \cdot \mathbf{x}) \leq \frac{1}{2}$.

We can control the second order term as follows

$$\begin{aligned}
\alpha^2(\mathbf{P}_j^T(\Delta \cdot \mathbf{x}))^2 &= \alpha^2 \left(\sum_i \mathbf{P}_{ji}(\Delta_i \mathbf{x}_i) \right)^2 \\
&\leq \alpha^2 \left(\sum_i \Delta_i \mathbf{P}_{ji}(\Delta_i \mathbf{x}_i) \right) \left(\sum_i \mathbf{P}_{ji} \mathbf{x}_i \right) \\
&\leq \alpha^2 \left(\sum_i \Delta_i \mathbf{P}_{ji}(\Delta_i \mathbf{x}_i) \right) \frac{10 \ln n}{\epsilon} \\
&= \alpha \left(\sum_i \Delta_i \mathbf{P}_{ji}(\Delta_i \mathbf{x}_i) \right) = \alpha \mathbf{P}_j^T(\Delta \cdot \Delta \cdot \mathbf{x}).
\end{aligned}$$

The first inequality is by the Cauchy-Schwarz inequality $\langle u, v \rangle^2 \leq \|u\|^2 \|v\|^2$, with $u_i = \sqrt{\Delta_i} \mathbf{P}_{ji}(\Delta_i \mathbf{x}_i)$ and $v_i = \sqrt{\mathbf{P}_{ji} \mathbf{x}_i}$. The second inequality is due to $\mathbf{P}_j^T \mathbf{x} \leq \frac{10 \ln n}{\epsilon}$ for all j . The last line is by our choice of $\alpha = \frac{\epsilon}{10 \ln n}$.

So far we have bounded the impact of the updates on each individual constraint, and we have

$$\text{lmax}(\mathbf{P}\mathbf{x}^{(t+1)}) \leq \ln \sum_j \exp(\mathbf{P}_j^T \mathbf{x}) (1 + \alpha \mathbf{P}_j^T(\Delta \cdot \mathbf{x}) + \alpha \mathbf{P}_j^T(\Delta \cdot \Delta \cdot \mathbf{x})).$$

We then translate the changes on each constraint to the combined change on $\text{lmax}(\mathbf{P}\mathbf{x})$. Intuitively, the combined change is a convex combination on the changes of each constraint, weighted by their exponential values $\mathbf{y}_j = \exp(\mathbf{P}_j^T \mathbf{x})$.

$$\begin{aligned}
&\ln \sum_j \exp(\mathbf{P}_j^T \mathbf{x}) (1 + \alpha \mathbf{P}_j^T(\Delta \cdot \mathbf{x}) + \alpha \mathbf{P}_j^T(\Delta \cdot \Delta \cdot \mathbf{x})) \\
&= \ln \sum_j \mathbf{y}_j (1 + \alpha \mathbf{P}_j^T(\Delta \cdot \mathbf{x}) + \alpha \mathbf{P}_j^T(\Delta \cdot \Delta \cdot \mathbf{x})) \\
&= \ln \left((\mathbf{1}^T \mathbf{y}) \left(1 + \sum_j \frac{\mathbf{y}_j}{\mathbf{1}^T \mathbf{y}} (\alpha \mathbf{P}_j^T(\Delta \cdot \mathbf{x}) + \alpha \mathbf{P}_j^T(\Delta \cdot \Delta \cdot \mathbf{x})) \right) \right) \\
&= \ln \left((\mathbf{1}^T \mathbf{y}) \left(1 + \alpha \left\langle \frac{\mathbf{y}}{\mathbf{1}^T \mathbf{y}}, \mathbf{P}((\mathbf{1} + \Delta) \cdot \Delta \cdot \mathbf{x}) \right\rangle \right) \right).
\end{aligned}$$

We can then write out the change of $\text{lmax}(\mathbf{P}\mathbf{x})$ explicitly as

$$\begin{aligned}
\text{lmax}(\mathbf{P}\mathbf{x}^{(t+1)}) &\leq \ln \left((\mathbf{1}^T \mathbf{y}) \left(1 + \alpha \left\langle \frac{\mathbf{y}}{\mathbf{1}^T \mathbf{y}}, \mathbf{P}((\mathbf{1} + \Delta) \cdot \Delta \cdot \mathbf{x}) \right\rangle \right) \right) \\
&= \ln \left((\mathbf{1}^T \mathbf{y}) \left(1 + \alpha \left\langle \frac{\mathbf{P}^T \mathbf{y}}{\mathbf{1}^T \mathbf{y}}, (\mathbf{1} + \Delta) \cdot \Delta \cdot \mathbf{x} \right\rangle \right) \right) \\
&= \ln \left((\mathbf{1}^T \mathbf{y}) (1 + \alpha \langle \mathbf{a}, (\mathbf{1} + \Delta) \cdot \Delta \cdot \mathbf{x} \rangle) \right) \\
&= \text{lmax}(\mathbf{P}\mathbf{x}) + \ln (1 + \alpha \langle \mathbf{a}, (\mathbf{1} + \Delta) \cdot \Delta \cdot \mathbf{x} \rangle) \\
&\leq \text{lmax}(\mathbf{P}\mathbf{x}) + \alpha \langle \mathbf{a}, (\mathbf{1} + \Delta) \cdot \Delta \cdot \mathbf{x} \rangle.
\end{aligned}$$

Recall $\mathbf{a} = \frac{\mathbf{P}^T \mathbf{y}}{\mathbf{1}^T \mathbf{y}}$ as defined in (9), and the last line is by $\ln(1 + x) \leq x$ for $x \geq 0$.

For the $\text{lmin}(\mathbf{C}^{(t)} \mathbf{x})$ part, we follow the same approach.

$$-\text{lmin}(\mathbf{C}^{(t+1)} \mathbf{x}^{(t+1)}) = \ln \sum_k \exp(-\mathbf{C}^{(t+1)} \mathbf{x}^{(t+1)})_k,$$

since $C^{(t+1)}$ can only have same or more rows dropped from $C^{(t)}$ due to the overly satisfied constraints, we know $\ln \sum_k \exp(-C^{(t+1)}\mathbf{x}^{(t+1)})_k \leq \ln \sum_k \exp(-C^{(t)}\mathbf{x}^{(t+1)})_k$. Again omitting the superscript (t) , we have

$$\begin{aligned} -\text{lmin}(C^{(t+1)}\mathbf{x}^{(t+1)}) &\leq \ln \sum_k \exp(-C(\mathbf{x} + \alpha\Delta \cdot \mathbf{x}))_k \\ &= \ln \sum_k \exp(-C\mathbf{x})_k \exp(-\alpha C(\Delta \cdot \mathbf{x}))_k \\ &\leq \ln \sum_k \exp(-C\mathbf{x})_k (1 - \alpha C(\Delta \cdot \mathbf{x}) + \alpha^2 (C(\Delta \cdot \mathbf{x}))^2)_k. \end{aligned}$$

The last inequality is due to $(\alpha C(\Delta \cdot \mathbf{x}))_k \leq \frac{1}{2}$ for all k , since in $C^{(t)}$ we only keep those constraints that are not above $\frac{10 \ln n}{\epsilon}$ yet, i.e. $(C\mathbf{x})_k \leq \frac{1}{\alpha}$ for all k . Again using Cauchy-Schwarz inequality as before and derivations similar to the $\text{lmax}(P\mathbf{x})$ case, we get

$$-\text{lmin}(C^{(t+1)}\mathbf{x}^{(t+1)}) \leq -\text{lmin}(C^{(t)}\mathbf{x}^{(t)}) - \alpha \langle \mathbf{b}, (1 - \Delta) \cdot \Delta \cdot \mathbf{x} \rangle.$$

◀

Appendix B Pure Packing and Pure Covering LPs

We will look at pure packing LPs, and the case for pure covering LPs will be symmetric.

Given packing LP in the optimization form

$$\max_{\mathbf{x} \geq 0} \{ \mathbf{1}^T \mathbf{x} : P\mathbf{x} \leq \mathbf{1} \}$$

via standard reduction and scaling, we need to solve a $(1 + \epsilon)$ -feasibility problem the same as in the mixed packing and covering case specified in (2) and (3). In the case of pure packing, we will have $C = c\mathbf{1}^T$ for some constant c .

The correctness proof follows from the mixed case, and we discuss how we get faster convergence for pure packing.

The special structure of C greatly simplifies the convergence analysis, as now z is a scalar, and $\mathbf{b}_i^{(t)} = c$ for all variables x_i across all iterations t . This allows us to aggregate the good iterations from all phases, and bound the total number of good phases by $\tilde{O}(\frac{1}{\epsilon^2})$, which will lead to $\tilde{O}(\frac{1}{\epsilon^2})$ total iterations.

In particular, now we look at all the good iterations across all phases together, and WLOG number them $1, \dots, T$. Claim 8 still holds, as it doesn't rely on phases. Then we can prove a stronger version of Lemma 9

► **Lemma 11.** *The total number of good iterations T is at most $O(\ln n \ln(\frac{m}{\epsilon})/\epsilon^2)$.*

The proof is a straightforward adaptation of the proof of Lemma 9. The proof is simpler, since now \mathbf{b}_i is a constant across all iterations, so the property (7) we have on the good iterations directly put all the values on the same scale, so we can lower bound the total update on x_i across all good iterations.

References

- 1 Zeyuan Allen-Zhu and Lorenzo Orecchia. Nearly-linear time positive LP solver with faster convergence rate. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, pages 229–236, 2015. Newer version available at <http://arxiv.org/abs/1411.1124>.

- 2 Zeyuan Allen-Zhu and Lorenzo Orecchia. Using optimization to break the epsilon barrier: A faster and simpler width-independent algorithm for solving positive linear programs in parallel. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '15*, pages 1439–1456, 2015.
- 3 Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012.
- 4 Baruch Awerbuch and Rohit Khandekar. Stateless distributed gradient descent for positive linear programs. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 691–700, 2008.
- 5 S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- 6 Lisa Fleischer. A fast approximation scheme for fractional covering problems with variable upper bounds. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004, New Orleans, Louisiana, USA, January 11-14, 2004*, pages 1001–1010, 2004.
- 7 Christos Koufogiannakis and Neal E. Young. A nearly linear-time PTAS for explicit fractional packing and covering linear programs. *Algorithmica*, 70(4):648–674, 2014.
- 8 Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in $\tilde{O}(\text{vrnk})$ iterations and faster algorithms for maximum flow. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 424–433, 2014.
- 9 Yin Tat Lee and Aaron Sidford. Efficient inverse maintenance and faster algorithms for linear programming. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 230–249, 2015.
- 10 Michael Luby and Noam Nisan. A parallel approximation algorithm for positive linear programming. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, May 16-18, 1993, San Diego, CA, USA*, pages 448–457, 1993.
- 11 Faraz Makari Manshadi, Baruch Awerbuch, Rainer Gemula, Rohit Khandekar, Julián Mestre, and Mauro Sozio. A distributed algorithm for large-scale generalized matching. *PVLDB*, 6(9):613–624, 2013. Available at <http://www.vldb.org/pvldb/vol6/p613-makarimanshadi.pdf>.
- 12 Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.
- 13 Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.
- 14 Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
- 15 Richard Peng and Kanat Tangwongsan. Faster and simpler width-independent parallel algorithms for positive semidefinite programming. In *Proceedings of the 24th ACM symposium on Parallelism in algorithms and architectures, SPAA '12*, pages 101–108, 2012. Available at <http://arxiv.org/abs/1201.5135>.
- 16 James Renegar. Efficient first-order methods for linear programming and semidefinite programming. *CoRR*, abs/1409.5832, 2014.
- 17 Di Wang, Michael W. Mahoney, Nishanth Mohan, and Satish Rao. Faster parallel solver for positive linear programs via dynamically-bucketed selective coordinate descent. *CoRR*, abs/1511.06468, 2015.
- 18 Neal E. Young. Sequential and parallel algorithms for mixed packing and covering. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 538–546, 2001.
- 19 Neal E. Young. Nearly linear-time approximation schemes for mixed packing/covering and facility-location linear programs. *CoRR*, abs/1407.3015, 2014.

- 20 Zeyuan Allen Zhu, Yin Tat Lee, and Lorenzo Orecchia. Using optimization to obtain a width-independent, parallel, simpler, and faster positive SDP solver. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1824–1831, 2016. Available at <http://arxiv.org/abs/1507.02259>.
- 21 Zeyuan Allen Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *CoRR*, abs/1407.1537, 2014.