Width-Independence Beyond Linear Objectives: 
Distributed Fair Packing and Covering Algorithms

Jelena Diakonikolas  
UC Berkeley  
jelena.d@berkeley.edu

Maryam Fazel  
University of Washington  
mfazel@uw.edu

Lorenzo Orecchia  
Boston University  
orecchia@bu.edu

Abstract

Fair allocation of resources has deep roots in early philosophy, and has been broadly studied in political science, economic theory, operations research, and networking. Over the past decades, an axiomatic approach to fair resource allocation has led to the general acceptance of a class of $\alpha$-fair utility functions parametrized by a single inequality aversion parameter $\alpha \in [0, \infty]$. In theoretical computer science, the most well-studied examples are linear utilities ($\alpha = 0$), proportionally fair or Nash utilities ($\alpha = 1$), and max-min fair utilities ($\alpha \to \infty$).

In this work, we consider general $\alpha$-fair resource allocation problems, defined as the maximization of $\alpha$-fair utility functions under packing constraints. We give improved distributed algorithms for constructing $\epsilon$-approximate solutions to such problems. Our algorithms are width-independent, that is, their running times depend only poly-logarithmically on the largest entry of the constraint matrix, and closely match the state-of-the-art guarantees for distributed algorithms for packing linear programs – the case $\alpha = 0$. The only previously known width-independent algorithms for $\alpha$-fair resource allocation, by Marasevic, Stein, and Zussman [29], have convergence times with much worse dependence on $\epsilon$ and $\alpha$. Our analysis leverages the Approximate Duality Gap framework of Diakonikolas and Orecchia [15] to obtain better algorithms with a more streamlined analysis.

Finally, we introduce a natural counterpart of $\alpha$-fairness for minimization problems and motivate its usage in the context of fair task allocation. This generalization yields $\alpha$-fair covering problems, for which we provide the first width-independent nearly-linear-time approximate solvers by reducing their analysis to the $\alpha < 1$ case of the $\alpha$-fair packing problem.

*Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing and while JD was a postdoctoral researcher at Boston University. It was partially supported by NSF grant #CCF-1718342 and by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant #CCF-1740425. MF was partially supported by NSF grant CCF-1409836 and ONR grant N000141612789.
1 Introduction

The question of how to split a set of limited resources is a fundamental question studied since antiquity, with the earliest theory of justice stated as Aristotle’s equity principle, whereby resources should be allocated in proportion to some pre-existing claims. This debate was formalized by political and moral philosophers, with two extremes being Bentham’s utilitarian principle [6], according to which the resources should be allocated so that a global utility is maximized, and the Rawlsian theory of justice [35], according to which the allocation should favor the least advantaged.

These two extremes are reflected in the most common algorithms within theoretical computer science. Most classical algorithms in the literature can be seen as utilitarian, with the goal of maximizing a (typically linear) utility, or minimizing cost; algorithms on the other extreme are known as max-min fair algorithms (e.g., [23]), in which resources are assigned according to the most egalitarian Pareto-optimal allocation. Notable exceptions to these two extremes are algorithms for computing market equilibria [14, 19].

The question of what constitutes “the best” allocation of resources was never settled. Indeed, there is a natural dichotomy between utility and fairness and one generally encounters tradeoffs between the two [8, 9]. In a recent survey in which people were asked to choose among different allocations, the participants naturally clustered into two groups, one favoring utility and the other favoring fairness [20]. Hence, it is a meaningful requirement to choose the fairness-utility tradeoff according to the application.

The study of fairness in economic theory, operations research, and networking culminated in a single class of utility functions known as $\alpha$-fair utilities:

$$f_{\alpha}(x) = \begin{cases} x^{1-\alpha} / (1-\alpha), & \text{if } \alpha \geq 0, \alpha \neq 1, \\ \log(x), & \text{if } \alpha = 1, \end{cases}$$

(1.1)

These functions were originally introduced in economic theory by Atkinson [2], for the purpose of ranking income distributions according to inequality aversion defined by the parameter $\alpha$. As $\alpha$ is increased from zero to infinity, fairness becomes more important than utility. When $\sum_j f_{\alpha}(x_j)$ is maximized over a convex set, with $x_j$ corresponding to the share of the resource to party $j$, the resulting solution is equivalent to the $\alpha$-fair allocation as defined by Mo and Walrand in the context of network bandwidth allocation [31]. This class of problems is well-studied in the literature, and axiomatically justified in several works [8, 25, 20].

Notable special cases of $\alpha$-fair allocations include: (i) utilitarian allocations (with linear objectives), for $\alpha = 0$, (ii) proportionally fair allocations that correspond to Nash bargaining solutions [32], for $\alpha = 1$, (iii) TCP-fair objectives that correspond to bandwidth allocations in the Internet TCP [22], for $\alpha = 2$, and (iv) max-min fair allocations that correspond to Kalai-Smorodinsky solutions in bargaining theory [21], for $\alpha \to \infty$. Up to constant multiplication and addition, these are the only utilities with the scale-invariance property [2, 12]: if all the resources are scaled by the same factor, the optimal allocations according to these utilities scale by that factor. This is one of the intuitive justifications for the resulting fairness properties.

In this paper, we consider the maximization of $\alpha$-fair utilities subject to positive linear (packing) constraints, to which we refer as the $\alpha$-fair packing problems. Specifically, given a non-negative matrix $A$ and a parameter $\alpha \geq 0$, $\alpha$-fair packing problems are defined as [29]:

$$\max \left\{ f_{\alpha}(x) \defeq \sum_{j=1}^{n} f_{\alpha}(x_j) : \ A x \leq \mathbb{1}, \ x \geq 0 \right\},$$

($\alpha$P)

where $\mathbb{1}$ is an all-ones vector, $\mathbf{0}$ is an all-zeros vector. Packing constraints are natural in many applications, including Internet congestion control [26], rate control in software defined networks [30], scheduling in wireless networks [36], multi-resource allocation and scheduling in datacenters [11, 20, 18], and a variety of applications in operations research, economics, and game theory [9, 19]. When $\alpha = 0$, the problem is equivalent to solving a packing linear program (LP).
We are interested in solving \((\alpha P)\) in a distributed model of computation. In such a model, each coordinate \(j\) of the allocation vector \(x\) is updated according to global problem parameters (e.g., \(m, n\)), relevant local information for coordinate \(j\), (namely, the \(j^{\text{th}}\) column of the constraint matrix \(A\)), and local information received in each round. The local per-round information for coordinate \(j\) is the slack \(1 - (Ax)_i\) of all the constraints \(i\) in which \(j\) participates. Such a model is natural for networking applications [22]. Further, as resource allocation problems frequently arise in large-scale settings in which results must be provided in real-time (e.g., in datacenter resource allocation [18, 20, 11]), the design of distributed solvers that can efficiently compute approximately optimal solutions to \(\alpha\)-fair allocation problems is of crucial importance. Note that in the case of LPs, the approaches based on the use of softmax objectives [37, 28] are parallel but not distributed, as they require access to full vector \(Ax\) to compute the allocation vector updates.

Our interest is specialized to first-order width-independent solvers. The use of first-order solvers is particularly relevant in this context, since these are the only solvers known to be implementable in the distributed model of computation and that lead to efficient, near-linear-time computation. Width-independent algorithms are of great theoretical interest: algorithms that are not width-independent cannot in general be considered polynomial time. These algorithms have primarily been studied in the context of packing and covering LPs. By contrast, width-independence is not possible for general LPs, for which first-order methods yield a linear dependence on the width \(p\) at best. More generally, the question of how the non-negativity of the constraints allows us to design width-independent algorithms is still an active area of research, with important connections to the design of scalable graph algorithms [33] and to the optimization of diminishing-returns functions [17]. From the optimization perspective, width-independence is surprising, as black-box application of any of the standard first-order methods does not lead to width-independent algorithms.

1.1 Our Contributions

We obtain improved distributed algorithms for constructing \(\epsilon\)-approximate solutions to \(\alpha\)-fair packing problems. As in [29], our specific convergence results depend on the regime of the parameter \(\alpha\), where each iteration takes linear work in the number of non-zero elements of \(A\):

- For \(\alpha \in [0, 1)\), Theorem C.4 leads to an \(\epsilon\)-approximate convergence in \(O\left(\frac{\log(n\rho)\log(mn\rho/\epsilon)}{(1-\alpha)^2\epsilon^2}\right)\) iterations. This bound matches the best known results for parallel and distributed packing LP solvers for \(\alpha = 0\) [1, 28], and improves the dependence on \(\epsilon\) compared to [29] from \(\epsilon^{-5}\) to \(\epsilon^{-2}\) for \(\alpha \in (0, 1)\).

- For \(\alpha = 1\), Theorem C.8 yields \(\epsilon\)-approximate convergence in \(O\left(\frac{\log^3(mn\rho/\epsilon)}{\epsilon^2}\right)\) iterations. The dependence on \(\epsilon\) compared to [29] is improved from \(\epsilon^{-5}\) to \(\epsilon^{-2}\).

- For \(\alpha > 1\), Theorem C.14 shows that \(O\left(\max\left\{\frac{\alpha^3\log(n\rho/\epsilon)}{\epsilon}, \frac{\log(\frac{1}{(\alpha-1)\epsilon})\log(mn\rho/\epsilon)}{\epsilon}\right\}\right)\) iterations suffice to obtain an \(\epsilon\)-approximation. This can be extended to the max-min-fair case by taking \(\alpha\) sufficiently large, as noted in [29]. The dependence on \(\epsilon\) compared to [29] is improved from \(\epsilon^{-4}\) to \(\epsilon^{-1}\).

While the analysis for each of these cases is somewhat involved, the algorithms we propose are extremely simple, as described in Algorithm 1 of Section 3. Moreover, our dependence on \(\epsilon\) is improved by a factor \(\epsilon^{-3}\) and the analysis is simpler than the one from [29], as it leverages the Approximate Duality Gap Technique (ADGT) of Diakonikolas and Orecchia [15], which makes it easier to follow and reconstruct.

Our final contribution is to introduce a natural counterpart of \(\alpha\)-fairness for minimization problems, which we use to study \(\beta\)-fair covering problems:\n
\[
\min \left\{ g_\beta(y) \define \sum_{i=1}^m \frac{y_i^{1+\beta}}{1+\beta} : \quad Ax \geq 1, \quad y \geq 0 \right\}. \tag{\beta C}
\]

\(^1\)I.e., their running times scale poly-logarithmically with the matrix width, defined as the maximum ratio of \(A\)'s non-zero entries.

\(^2\)As in previous work [29], the approximation is multiplicative for \(\alpha \neq 1\) and additive for \(\alpha = 1\) (see Theorem 3.3).

\(^3\)We use \(\beta\) instead of \(\alpha\) to distinguish between the different parameters in the convergence analysis.
As for packing problems, the $\beta$-covering formulation can be motivated by the goal to produce an equitable allocation. In the work allocation example of covering problems, assigning all work to a single worker may provide a solution that minimizes total work. For instance, this may be the case if a single worker is greatly more efficient than other workers. However, a fair solution would allocate work so that each worker gets some and no worker gets too much. This is captured in the previous program by the fact that the objective quickly grows to infinity for $\beta > 0$, as the amount of work $y_i$ given to worker $i$ increases.

This generalization yields $\beta$-fair covering problems, for which we provide the first width-independent nearly-linear-time approximate solver that converges in $O(\frac{1+\beta}{\beta} \log(mn\rho))$ iterations, by reducing the analysis to the $\alpha < 1$ case of the $\alpha$-fair packing problems, as shown in Section 4.

1.2 Our Techniques

There are several difficulties that arise when considering cases $\alpha > 0$ compared to the linear case ($\alpha = 0$). Unlike linear objectives, which are 1-Lipschitz, $\alpha$-fair objectives for $\alpha > 0$ lack any good global properties typically used in convex optimization, such as e.g., Lipschitz-continuity of the function or its gradient. In particular, when $x_j \to 0$, $\frac{d f_\alpha(x_j)}{dx_j} \to \infty$. While it is possible to prune the feasible region to guarantee positivity of the vector $x$, any pruning that provably retains $\epsilon$-approximate solutions would in general require that the point $\frac{1}{n\rho} \mathbb{1}$ is contained in the pruned set, leading to the Lipschitz constants of the order $(n\rho)\alpha$ and $(n\rho)\alpha+1$. This makes it hard to directly apply arguments relying on gradient truncation used in the packing LP case [1]. To circumvent this issue, we use a change of variables, which reduces the objective to a linear one, but makes the constraints more complicated. Further, in the case $\alpha > 1$, the truncated gradient has the opposite sign compared to the $\alpha \leq 1$ cases. Though this change in the sign may seem minor, it invalidates the arguments that are typically used in analyzing distributed packing LP solvers [1, 16], which is one of the main reasons why in the linear case the solution to the covering LP is obtained by solving its dual – a packing LP. Unfortunately, in the case $\alpha > 1$, solving the dual problem seems no easier than solving the primal – in terms of truncation, the gradients have the same structure as in the covering LP.

Similar to the linear case [1], we use regularization of the constraint set to turn the problem into an unconstrained optimization problem over the non-negative orthant. The regularizing function is different from the standard generalized entropy typically used for LPs, and belongs to the same class of functions considered in the fair covering problem. The use of these regularizers seems more natural than entropy, as local smoothness properties used in the analysis of the algorithms hold regardless of whether the point at which local smoothness is considered is feasible according to the packing constraints, which is not true for entropic regularization. Furthermore, the use of these particular regularizers is crucial for reducing the fair covering problems to solving $\alpha$-fair packing problems with $\alpha < 1$ (see Section 4 and Appendix D).

While the analysis of the case $\alpha \in [0, 1)$ is similar to the analysis of packing LPs from our unpublished note [16], it is not clear how to generalize this argument to the cases $\alpha = 1$ and $\alpha > 1$, with these techniques or any others developed for packing LPs (see Sections 2.2 and 3 for more details). The analysis of the case $\alpha = 1$ is relatively simple, and can be seen as a generalization of the gradient descent analysis.

However, the case $\alpha > 1$ is significantly more challenging. First, ADGT [15] cannot be applied directly, for a number of reasons: (i) it is hard to argue that any naturally chosen initial solution is a constant-approximation-factor away from the optimum; this is because when $\alpha > 1$, $f_\alpha(\frac{1}{n\rho}) = -\frac{(n\rho)^{\alpha-1}}{\alpha-1}$, which may be much smaller than $-n/(\alpha - 1)$, while the optimal solution can be as large as $-n/(\alpha - 1)$, (ii) gradient truncation cannot be applied to the approximate gap constructed by ADGT (see Section 3), and (iii) without the gradient truncation, it is not clear how to argue that the approximate gap from ADGT decreases at the right rate (or at all), which is crucial for the ADGT argument to apply. One of the reasons for (iii) is that the homotopic approximation of the objective used in constructing an approximate dual in ADGT is a crude approximation of the true dual when $\alpha > 1$. 

3
Our main idea is to use the Lagrangian dual of the original, non-regularized problem, with two different arguments. The first argument is local and relies on similar ideas as [29]: it uses only the current iterate to argue that if certain regularity conditions do not hold, the regularized objective must decrease by a sufficiently large multiplicative factor. Compared to [29], we require much looser regularity conditions, which eventually leads to a much better dependence of the convergence time on $\epsilon$: the dependence is reduced from $1/\epsilon^4$ to $1/\epsilon$ without incurring any additional logarithmic factors in the input size, and even improving the dependence on $\alpha$. This is achieved through the use of the second argument, which relies on the aggregate history of the iterates that satisfy the regularity conditions. This argument is more similar to ADGT, though as noted above and unlike in the standard ADGT [15], the approximate gap is constructed from the Lagrangian dual of the original problem. To show that the approximate gap decreases at the right rate, we use a careful coupling of the regularity conditions with the gradient truncation (see Appendix C.3.2).

1.3 Related Work

Related Work in the Offline Setting A long line of work on packing and covering LPs has resulted in width-independent distributed algorithms [24, 27, 5, 37, 34, 3]. This has culminated in recent results that ensure convergence to an $\epsilon$-approximate optimal solution in $O(\log n/\epsilon^2)$ rounds of computation [1, 28]. However, when it comes to the general $\alpha$-fair resource allocation, only [29] provides a width-independent algorithm. The algorithm of [29] works in a very restrictive setting of stateless distributed computation, which leads to convergence times that are poly-logarithmic in the problem parameters, but have high dependence on the approximation parameter $\epsilon$ (namely, the dependence is $\epsilon^{-5}$ for $\alpha \leq 1$ and $\epsilon^{-4}$ for $\alpha > 1$).

Related Work in the Online Setting A related topic is that of online algorithms for budgeted allocation, including online packing LPs. In online problems, at each time step $t$ new problem data (i.e., coefficients of variable $x_t$ in the cost function and in the packing constraints) are revealed, and the algorithm needs to make an irrevocable decision that fixes the variable $x_t$. The goal is to find algorithms with guarantees on the competitive ratio under an adversarial input sequence. For online packing LPs, [13] proposed a continuous update algorithm that achieves a constant competitive ratio if the budget constraint is allowed to be violated by a factor that depends only logarithmically on the width, thus their algorithm is width-independent. For more general nonlinear cost functions, several recent extensions exist: [4] gives width-independent algorithms for online covering/packing for nonlinear costs under certain assumptions; [17] bounds the competitive ratio for packing problems that satisfy a "diminishing returns" property. For monotone functions, the simultaneous update algorithm in [17] is width-independent. However, the algorithms and the analysis used in this literature are not suitable for our purposes: given the restricted setting they are designed for, the algorithms’ performance with respect to problem size and accuracy $\epsilon$ is not as good as what we are able to obtain here. In addition, online algorithms such as in [4, 17] need to monotonically increase the dual variables in each iteration, while such assumptions are not necessary for our algorithms. As a result, competitive ratios for online algorithms can be studied by coarser arguments than what we need here.

1.4 Organization of the Paper

Section 2 introduces the necessary notation, together with the useful definitions and facts subsequently used in the convergence analysis. Section 3 provides the statement of the algorithm for $\alpha$-fair packing and overviews the main technical ideas used in its convergence analysis. The full technical argument is deferred to Appendix C. Section 4 overviews the results for $\beta$-fair covering, while the full convergence analysis is provided in Appendix D. We conclude in Section 5 and highlight several interesting open problems.
2 Notation and Preliminaries

We assume that the problems are expressed in their standard scaled form \([29, 27, 3, 1]\), so that the minimum non-zero entry of the constraint matrix \(A\) equals one. Note that even weighted versions of the problems can be expressed in this form through rescaling and the change of variables. Observe that under this scaling, maximum element of \(A\) is equal to the matrix width, denoted by \(\rho\) in the rest of the paper. The dimensions of the constraint matrix \(A\) are \(m \times n\).

2.1 Notation and Useful Definitions and Facts

We use boldface letters to denote vectors and matrices, and italic letters to denote scalars. We let \(x^a\) denote the vector \([x_1^a, x_2^a, ..., x_n^a]^T\), \(\exp(x)\) denote the vector \([\exp(x_1), \exp(x_2), ..., \exp(x_n)]^T\). Inner product of two vectors is denoted as \(\langle \cdot, \cdot \rangle\), while the matrix/vector transpose is denoted by \((\cdot)^T\). The gradient of a function \(f\) at coordinate \(j\), i.e., \(\frac{\partial f}{\partial x_j}\), is denoted by \(\nabla_j f(\cdot)\). We use the following notation for the truncated (and scaled) gradient \([1]\), for \(\alpha \neq 1\):

\[
\nabla_j f(x) = \begin{cases} 
(1 - \alpha) \nabla_j f(x), & \text{if } (1 - \alpha) \nabla_j f(x) \in [-1, 1], \\
1, & \text{otherwise}.
\end{cases}
\]  

(2.1)

As we will see later, the only relevant case for us will be the functions whose gradient coordinates satisfy \((1 - \alpha) \nabla_j f(x) \in [-1, \infty)\). Hence, the gradient truncation is irrelevant for \((1 - \alpha) \nabla_j f(x) < -1\). The definition of the truncated gradient for \(\alpha = 1\) is equivalent to the definition (2.1) with \(\alpha = 0\).

Most functions we will work with are convex differentiable functions defined on \(\mathbb{R}^n_+\). Thus, we will be stating all definitions assuming that the functions are defined on \(\mathbb{R}^n_+\). A useful definition of convexity of a function \(f : \mathbb{R}^n_+ \to \mathbb{R}\) is:

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathbb{R}^n_+.
\]  

(2.2)

Useful for our analysis is the concept of convex conjugates:

**Definition 2.1.** Convex conjugate of function \(\psi : \mathbb{R}^n_+ \to \mathbb{R}\) is defined as:

\[
\psi^*(z) \overset{\text{def}}{=} \sup_{x \geq 0} \{ \langle z, x \rangle - \psi(x) \}.
\]

Since we will only be considering continuous functions on the closed set \(\mathbb{R}^n_+\), \(\sup\) from the previous definition can be replaced by \(\max\). The following is a standard fact about convex conjugates:

**Fact 2.2.** Convex conjugate \(\psi^*\) of a function \(\psi\) is convex. Moreover, if \(\psi\) is strictly convex, \(\psi^*\) is differentiable, and the following holds:

\[
\nabla \psi^*(z) = \arg\max_{x \geq 0} \{ \langle z, x \rangle - \psi(x) \}.
\]

2.2 Fair Packing and Covering

**Fair Packing Problems.** Recall that \(\alpha\)-fair packing problems were defined by \((\alpha P)\). In the analysis, there are three regimes of \(\alpha\) that are handled separately: \(\alpha \in [0, 1)\), \(\alpha = 1\), and \(\alpha > 1\). In these three regimes, the \(\alpha\)-fair utilities \(f_\alpha\) exhibit very different behaviors, as illustrated in Fig 1. When \(\alpha = 0\), \(f_\alpha\) is just the linear utility, and \((\alpha P)\) is a packing LP. As \(\alpha\) increases from zero to one, \(f_\alpha\) remains non-negative, but its shape approaches the shape of the natural logarithm. When \(\alpha = 1\), \(f_\alpha\) is simply the natural logarithm. When \(\alpha > 1\), \(f_\alpha\) is non-positive and its shape approaches the shape of the natural logarithm as \(\alpha\) tends to 1. As \(\alpha\) increases, \(f_\alpha\) bends and becomes steeper, approaching the negative indicator of the interval \([0, 1]\) as \(\alpha \to \infty\).
It will be convenient to perform the following change of variables:

\[ x = F_\alpha(\hat{x}) \overset{\text{def}}{=} \begin{cases} \hat{x}^{1/\alpha}, & \text{if } \alpha \geq 0, \alpha \neq 1, \\ \exp(\hat{x}), & \text{if } \alpha = 1. \end{cases} \tag{2.3} \]

Let \( S_\alpha = \mathbb{R}^n_+ \), \( \hat{f}_\alpha(x) = \langle \frac{n}{1-\alpha} x, \alpha \rangle \) for \( \alpha \neq 1, \alpha \geq 0 \) and \( S_\alpha = \mathbb{R}^n \), \( \hat{f}_\alpha(x) = \langle 1, x \rangle \) for \( \alpha = 1 \). The problem \((\alpha P)\) can then equivalently be written (with the abuse of notation) as:

\[ \max \{ \hat{f}_\alpha(x) : A F_\alpha(x) \leq 1, x \in S_\alpha \}. \tag{\alpha P_c} \]

To bound the optimality gap in the analysis, it is important to bound the optimum objective function values, as in the following proposition.

**Proposition 2.3.** Let \( x^* \) be (any) optimal solution to \((\alpha P_c)\). Then:

- If \( \alpha \geq 0 \text{ and } \alpha \neq 1 \), \( \frac{n}{1-\alpha} (n\|A\|_\infty)^{\alpha-1} \leq \hat{f}_\alpha(x^*) \leq \frac{n}{1-\alpha} \).
- If \( \alpha = 1 \), \( -n \log(n\|A\|_\infty) \leq \hat{f}_\alpha(x^*) \leq 0 \).

**Proof.** The proof is based on the following simple argument. When \( F_\alpha(x) = \frac{1}{n\|A\|_\infty} \), \( x \) is feasible and we get a lower bound on the optimal objective value. On the other hand, if \( F_\alpha(x) > 1 \), then (as the minimum non-zero entry of \( A \) is at least 1), all constraints are violated, which gives an upper bound on the optimal objective value. The details are omitted. \( \square \)

It is possible to equivalently write \((\alpha P_c)\) as the following (non-smooth) saddle-point problem, similar to the case of the packing LP (see, e.g., [1]):

\[ \min_{x \in S_\alpha} -\hat{f}_\alpha(x) + \max_{y \geq 0} \langle AF_\alpha(x) - 1, y \rangle. \tag{2.4} \]

The main reason for considering the saddle-point formulation of \((\alpha P) - (\alpha P_c)\) is that after regularization (or smoothing) it can be turned into an unconstrained problem over the positive orthant, without losing much in the approximation error, under some mild regularity conditions on the steps of the algorithm. In particular, let \((x^*, y^*)\) be the optimal primal-dual pair in (2.4). Then, by Fenchel’s Duality (see, e.g., Proposition 5.3.8 in [7]), we have that \( -\hat{f}_\alpha(x^*) = \min_{x \in S_\alpha} \{ -\hat{f}_\alpha(x) + \langle AF_\alpha(x) - 1, y^* \rangle \} \). Hence, \( \forall x \geq 0 \):

\[ -\hat{f}_\alpha(x^*) \leq -\hat{f}_\alpha(x) + \langle AF_\alpha(x) - 1, y^* \rangle - \psi(y^*) + \psi(y^*) \]

\[ \leq -\hat{f}_\alpha(x) + \max_{y \geq 0} \{ \langle AF_\alpha(x) - 1, y \rangle - \psi(y) \} + \psi(y^*) \]

\[ = f_r(x) + \psi(y^*), \tag{2.5} \]
where \( f_r(x) = -f_\alpha(x) + \psi^*(A F_\alpha(x) - 1) \). The main idea is to show that we can choose the function \( \psi \) so that \( f_c \) closely approximates \(-f_\alpha\) around the optimal point \( x^* \), and, further, we can recover a \((1 + O(\epsilon))\)-approximate solution to \((\alpha P)\) from a \((1 + \epsilon)\)-approximate solution to \(\min_{x \geq 0} f_r(x)\). This will allow us to focus on the minimization of \( f_r \), without the need to worry about satisfying the packing constraints from \((\alpha P)\) in each step of the algorithm. The following proposition formalizes this statement and introduces the missing parameters. In the choice of \( \psi(\cdot) \), the factor \( C^{-\beta} \) ensures that the algorithm maintains (strict) feasibility of running solutions. The case \( C = 1 \) would allow violations of the constraints by a factor \((1 + \epsilon)\).

**Proposition 2.4.** Let \( \psi(y) = \sum_{i=1}^{m} \left( \frac{y_i}{C_{\alpha}^{1+\beta}(1+\beta)} - y_i \right) \), where \( \beta = \frac{\epsilon^4/4}{(1+\alpha)\log(4m\rho/\epsilon)} \), \( C = (1 + \epsilon/2)^{1/\beta} \), and \( \epsilon \in (0, \min\{1/2, 1/(10|\alpha - 1|)\}) \) is the approximation parameter. Then:

1. \( f_r(x) = -f_\alpha(x) + \frac{C\beta}{1+\beta} \sum_{i=1}^{m} (A F_\alpha(x))_{i}^{1+\beta} \).

2. Let \( x_*^r = \arg\min_{x \in S_{\alpha}} f_r(x) \), \( x_*^\alpha \) be a solution to \((\alpha P)\), and \( x_*^r = F_\alpha(x_*^r) \). Then \( x_*^r \) is \((\alpha P)\)-feasible and:

\[-f_\alpha(x_*^r) + f_\alpha(x_*^\alpha) \leq f_r(x_*^r) + f_\alpha(x_*^\alpha) \leq 2\epsilon f \equiv 2 \begin{cases} \epsilon n, & \text{if } \alpha = 1; \\ \epsilon(1-\alpha)f_\alpha(x_*^\alpha), & \text{if } \alpha \neq 1. \end{cases} \]

The proof is provided in Appendix A.

**Fair Covering Problems** A natural counterpart to \( \alpha \)-fair packing problems is the \( \beta \)-fair covering, defined in \((\beta C)\). Similar as in the case of \( \alpha \)-fair packing, when \( \beta = 0 \), the problem reduces to the covering LP. It is not hard to show (using similar arguments as in [31]) that when \( \beta \rightarrow \infty \), the optimal solutions to \((\beta C)\) converge to the min-max fair allocation.

For our analysis, it is convenient to work with the Lagrangian dual of \((\beta C)\) (this is also the Fenchel dual of the sum of \( g_\beta(y) \) and the indicator of the packing polytope), which is:

\[ \max_{x \geq 0} \langle 1, x \rangle - \frac{\beta}{1+\beta} \sum_{i=1}^{m} (Ax)_i^{1+\beta} \],

i.e., solving the dual of \((\beta C)\) is the same as minimizing \( f_r(x) \) from the packing problem, with \( \alpha = 0 \) and \( \beta \) from the fair covering formulation \((\beta C)\).

Similar as for the packing, it is useful to bound the optimal objective values.

**Proposition 2.5.** Let \( y_*^\beta \) be an optimal solution to \((\beta C)\). Then: \( \left( \frac{1}{m\rho} \right)^{1+\beta} \frac{m}{1+\beta} \leq \sum_{i=1}^{m} \frac{(y_*^\beta)_i^{1+\beta}}{1+\beta} \leq \frac{m}{1+\beta} \).

The following proposition is a simple corollary of Lagrangian duality.

**Proposition 2.6.** Let \( (y_*^\beta, x_*^\alpha) \) be the optimal primal-dual pair for \((\beta C)\). Then \( \langle 1, x_*^\alpha \rangle = (1 + \beta)g_\beta(y_*^\beta) \).

**Proof.** By strong duality, \( \langle 1, x_*^\alpha \rangle = \frac{\beta}{1+\beta} \sum_{i=1}^{m} (Ax)_i^{1+\beta} = g_\beta(y_*^\beta) \) and \( y_*^\beta = (Ax)_i^{1/\beta} \). \( \square \)

3 Fair Packing: Algorithm and Convergence Analysis Overview

The algorithm pseudocode is provided in Algorithm 1 (FAIRPACKING). All the parameter choices will become clear from the analysis.

We start by characterizing the “local smoothness” of \( f_r \), which will be crucial for the analysis.
Algorithm 1 FAIRPACKING($A$, $\epsilon$, $\alpha$)

1: $\beta = \frac{\epsilon / 4}{(1+\alpha)\log(4n\rho/\epsilon)}$
2: Initialize: $x_j^{(0)} = ((1-\epsilon)/(n\rho))^{1-\alpha} \mathbf{1}$.
3: if $\alpha < 1$ then
4: $z^{(0)} = \exp(\epsilon/4)\mathbf{1}$, $\beta' = \frac{(1-\alpha)\epsilon / 4}{\log(n\rho/(1-\epsilon))}$, $h = \frac{(1-\alpha)\beta'}{16\alpha(1+\alpha \beta)}$
5: for $k = 1$ to $K = \left\lceil 2/(1-\alpha)h\epsilon \right\rceil$ do
6: $x^{(k)} = (1 + z^{(k-1)})^{-1/\beta'}$
7: $z^{(k)} = z^{(k-1)} + \epsilon h \nabla f_r(x^{(k)})$
8: end for
9: else if $\alpha = 1$ then
10: for $k = 1$ to $K = \left\lceil 10 \log^2(8\rho nn/\epsilon) / \epsilon \beta \right\rceil$ do
11: $x^{(k)} = x^{(k-1)} - \frac{\beta}{4(1+\beta)} \nabla f_r(x^{(k-1)})$
12: end for
13: else
14: for $k = 1$ to $K = \left\lceil 800 (1+\alpha)^2 \log(n\rho/\epsilon \min\{\alpha - 1, 1\}) / \beta \min\{\alpha - 1, 1\} \right\rceil$ do
15: $x^{(k)} = (1 - \frac{\beta}{4(1+\alpha \beta)} \nabla f_r(x^{(k-1)}) \mathbf{x}^{(k-1)}$
16: end for
17: end if
18: return $F_\alpha(x^{(K)})$

3.1 Local Smoothness and Feasibility

The following lemma characterizes the step sizes that are guaranteed to decrease the function value. Since the algorithm makes multiplicative updates for $\alpha \neq 1$, we will require that $x > 0$, which will hold throughout, due to the particular initialization and the choice of the steps.

Lemma 3.1. 1. Let $x > 0$. If $\alpha \neq 1$, $\Gamma = \text{diag}(\gamma)$, and $\gamma_j = -\frac{c_j}{4} \cdot (1-\alpha)\beta/(1+\alpha \beta) \nabla f_j(x)$ for $c_j \in [0, 1]$ then:

$$f_r(x + \Gamma x) - f_r(x) \leq -\frac{\beta(1-\alpha)}{1+\alpha \beta} \sum_j \frac{c_j}{4} \left(1 - \frac{c_j}{2}\right) x_j \nabla f(x) \nabla f_j(x).$$

2. If $\alpha = 1$ and $\Delta x \geq 0$ is such that $\Delta x_j = \frac{c_j \beta}{4(1+\beta)} \nabla f_j(x)$ for $c_j \in [0, 1]$, then:

$$f_r(x + \Delta x) - f_r(x) \leq -\frac{\beta}{1+\beta} \sum_{j=1}^n \frac{c_j}{4} \left(1 - \frac{c_j}{2}\right) \nabla f_j(x) \nabla f_j(x).$$

Lemma 3.1 also allows us to guarantee that the algorithm always maintains feasible solutions, as follows.

Proposition 3.2. Solution $x^{(k)}$ held by FAIRPACKING at any iteration $k \geq 0$ is $(\alpha P_\epsilon)$-feasible.

Proof. By the initialization and steps of FAIRPACKING, $x^{(k)} \in S_\alpha$, $\forall k$. It remains to show that it must be $A F_\alpha(x^{(k)}) \leq 1$, $\forall k$. Observe that $A x^{(0)} \leq (1-\epsilon) \mathbf{1}$. Suppose that in some iteration $k$, $\exists i$ such that $(A F_\alpha(x))_i \geq 1 - \epsilon/8$. Fix one such $i$ and let $\tilde{k}$ be first such iteration. We provide the proof for the case when $\alpha < 1$. The cases $\alpha = 1$ and $\alpha > 1$ follow by similar arguments.

Assume that $\alpha < 1$. Then for all $j$ such that $A_{ij}(x_j^{(k)})^{1-\alpha} \geq \frac{1}{4n}$ (there must exist at least one such $j$, as $(A F_\alpha(x))_i \geq 1 - \frac{\epsilon}{8} \geq \frac{3}{8}$), we have $(x_j^{(k)})^{1-\alpha} \geq \frac{1}{4n} \rho$ and $\nabla f_j(x^{(k)}) \geq \frac{1}{1-\alpha}(1 + (\frac{1}{4n})^\alpha (\alpha + 2/3)) > \frac{1}{1-\alpha}$.
Hence, using Lemma 3.1, \((A(x^{(k+1)}) \sqrt[1-\alpha]{\epsilon}), \leq (A(x^{(k)}) \sqrt[1-\alpha]{\epsilon}),\). As the maximum increase in \((A(x^{(k)}) \sqrt[1-\alpha]{\epsilon}),\) in any iteration is by factor less than \(1 + \frac{\epsilon}{k},\) it follows that it must be \(A(x^{(k)}) \sqrt[1-\alpha]{\epsilon}, \leq 1, \forall k.\)

\[\square\]

### 3.2 Main Theorem

Our main results are summarized in the following theorem. The theorem is proved through Theorems C.4, C.8, and C.14, which can be found in Appendix C.

**Theorem 3.3. (Main Theorem)** Given \(A, \alpha \geq 0,\) and \(\epsilon \in (0, \min\{1/2, 1/(10|\alpha - 1|)\}],\) let \(x^{(k)}_\alpha = F_\alpha(x^{(k)})\) be the solution produced by FAIRPACKING and let \(x^*_\alpha\) be the optimal solution to \((\alpha P)\). Then \(x^{(k)}_\alpha\) is \((\alpha P)-\)feasible and \(f_\alpha(x^{(k)}_\alpha) - f_\alpha(x^{(k)}_\alpha) = O(\epsilon),\) where:

\[
\epsilon_f = \begin{cases} 
\epsilon n, & \text{if } \alpha = 1, \\
\epsilon(1 - \alpha)f_\alpha(x^{(k)}_\alpha), & \text{if } \alpha \neq 1.
\end{cases}
\]

The total number of iterations taken by the algorithm, each requiring linear work in the number of non-zero elements of \(A,\) is:

\[
K = \begin{cases} 
O\left(\frac{\log(np) \log(mnp/\epsilon)}{(1-\alpha)^2\epsilon^2}\right), & \text{if } \alpha \in [0, 1), \\
O\left(\frac{\log^3(mnp/\epsilon)}{\epsilon^2}\right), & \text{if } \alpha = 1, \\
O\left(\max\left\{\frac{\alpha^3 \log(1/\epsilon) \log(mnp/\epsilon)}{\epsilon}, \frac{\log(\frac{1}{(\alpha-1)\epsilon}) \log(mnp/\epsilon)}{\epsilon(\alpha-1)}\right\}\right), & \text{if } \alpha > 1.
\end{cases}
\]

### 3.3 Approximate Duality Gap

The proof relies on the construction of an approximate duality gap, similar to our general technique [15]. The idea is to construct an estimate of the optimality gap for the running solution. Namely, we want to show that an estimate of the true optimality gap \(-f_\alpha(x^{(k)}_\alpha) + f_\alpha(x^*_\alpha)\) decreases as the function of the iteration count \(k,\) where \(x^{(k)}_\alpha = F_\alpha(x^{(k)})\) (recall that, by Proposition 3.2, \(x^{(k)}_\alpha\) is \((\alpha P)-\)feasible). By construction of \(f_r,\) we have that \(f_r(x^{(k)}_\alpha) \geq -f_\alpha(x^{(k)}_\alpha),\) hence it is an upper bound on \(-f_\alpha(x^{(k)}_\alpha).\) In the analysis, we will use \(U_k = f_r(x^{(k+1)}_\alpha)\) as the upper bound. The lower bound \(L_k\) needs to satisfy \(L_k \geq -f_\alpha(x^*_\alpha).\) The approximate optimality (or duality, see [15]) gap at iteration \(k\) is defined as \(U_k - L_k.\)

The main idea in the convergence argument is to show that the duality gap \(G_k\) decreases at rate \(H_k;\) namely, the idea is to show that \(H_k G_k \leq H_{k-1} G_{k-1} + E_k\) for an increasing sequence of positive numbers \(H_k\) and some “sufficiently small” \(E_k.\) This argument is equivalent to saying that \(-f_\alpha(x^{(k+1)}_\alpha) + f_\alpha(x^*_\alpha) \leq U_k - L_k = G_k \leq H_k G_0 + \sum_{i=1}^{k} E_i / H_{k_i},\) which gives the standard form of convergence for first-order methods. Observe that for this argument to lead to the convergence times of the form poly-log(input-size)/poly(\(\epsilon\)), the initial gap should correspond to a constant (or poly-log) optimality gap. This will be achievable through the appropriate initialization for \(\alpha \leq 1.\) For \(\alpha > 1,\) it is unclear how to initialize the algorithm to guarantee small initial gap (and the right change in the gap in general). Instead, we will couple this gap argument with another argument, so that the gap argument is valid on some subsequence of the iterates. We will argue that in the remaining iterations \(f_r\) must decrease by a sufficiently large multiplicative factor, so that either way we approach a \((1 + \epsilon)\) -approximate solution at the right rate.

**Local Smoothness and the Upper Bound** As already mentioned, our choice of the upper bound will be \(U_k = f_r(x^{(k+1)}_\alpha).\) The reason that the upper bound “looks one step into the future” is that it will hold a sufficiently lower value than \(f_r(x^{(k)}_\alpha)\) (and it will always decrease, due to Lemma 3.1) to compensate for any decrease in the lower bound \(L_k\) and yield the desired change in the gap: \(H_k G_k \leq H_{k-1} G_{k-1} + E_k.\)
Lower Bound  Let \( \{h_{\ell}\}_{\ell=0}^k \) be a sequence of positive real numbers such that \( H_k = \sum_{\ell=0}^k h_{\ell}. \)

The simplest lower bound is just a consequence of convexity of \( f_r \) and the fact that it closely approximates \( -f_\alpha \) (due to Proposition 2.4):

\[
-f_\alpha(x^*_\alpha) \geq f_r(x^*_\alpha) - 2\epsilon_f \geq \frac{\sum_{\ell=0}^k h_{\ell} (f_r(x^{(\ell)}) + \langle \nabla f_r(x^{(\ell)}), x^*_\alpha - x^{(\ell)} \rangle)}{H_k}.
\]

(3.1)

Even though simple, this lower bound is generally useful for the analysis of gradient descent, and we will show that it can be extended to the analysis of the \( \alpha = 1 \) case. However, this lower bound is not useful in the case of \( \alpha \neq 1 \). The reason comes as a consequence of the "gradient-descent-type" decrease from Lemma 3.1. While for \( \alpha = 1 \), the decrease can be expressed solely as the function of the gradient \( \nabla f(x^{(k)}) \) (and global problem parameters), when \( \alpha = 1 \), the decrease is also a function of the current solution \( x^{(k)} \). This means that the progress made by the algorithm (in the primal space) and consequently the approximation error would need to be measured w.r.t. to the norm \( \| \cdot \|_{1/x^{(k)}} \). In other words, we would need to be able to relate \( \sum_{j=1}^n (x^{(k)}_j - x^*_j)^2 / x^{(k)}_j \) to the value of \( f_r(x^*) \), which is not even clear to be possible (see the convergence argument from Section C.2 for more information).

However, for \( \alpha < 1 \), it is possible to obtain a useful lower bound from (3.1) after performing gradient truncation and regularization, similar as in our note on packing and covering LP [16]. Let \( \phi : \mathbb{R}_{+}^n \rightarrow \mathbb{R} \) be a convex function (that will be specified later). Then, denoting \( x^* = x^*_\alpha = \arg \min_u f_r(u) \), we have:

\[
f_r(x^*) \geq \frac{\sum_{\ell=0}^k h_{\ell} (f_r(x^{(\ell)}) - \langle \nabla f_r(x^{(\ell)}), x^{(\ell)} \rangle) + \sum_{\ell=0}^k h_{\ell} \langle \nabla f_r(x^{(\ell)}), x^*_\alpha \rangle + \frac{1}{1-\alpha} \phi(x^*) - \frac{1}{1-\alpha} \phi(x^*)}{H_k}
\]

\[
\geq \frac{\sum_{\ell=0}^k h_{\ell} (f_r(x^{(\ell)}) - \langle \nabla f_r(x^{(\ell)}), x^{(\ell)} \rangle) + \frac{1}{1-\alpha} \min_{u \geq 0} \{ \sum_{\ell=0}^k h_{\ell} \langle \nabla f_r(x^{(\ell)}), u \rangle + \phi(u) \}}{H_k}
\]

\[
- \frac{1}{1-\alpha} \phi(x^*)
\]

\[
= L_{k<1}^{\alpha} + 2\epsilon_f.
\]

Note that the same lower bound cannot be derived for \( \alpha \geq 1 \). The reason is that we cannot perform gradient truncation, as for \( \alpha > 1 \) (resp. \( \alpha = 1 \)), \( \langle \nabla f_r(x), x^* \rangle \geq \frac{1}{1-\alpha} \langle \nabla f_r(x), x^* \rangle \) (resp. \( \langle \nabla f_r(x), x^* \rangle \geq \langle \nabla f_r(x), x^* \rangle \) does not hold. For \( \alpha > 1 \), we will make use of the Lagrangian dual of \((\alpha P_c)\), which takes the following form for \( y \geq 0 \):

\[
g(y) = - \langle 1, y \rangle - \frac{\alpha}{1-\alpha} \sum_{j=1}^n (A^T y)_j \frac{1}{\alpha}. \quad (3.2)
\]

Finally, we note that it is not clear how to make use of the Lagrangian dual in the case of \( \alpha \leq 1 \). When \( \alpha < 1 \), the terms \( -\frac{1}{1-\alpha} (A^T y)_j \frac{1}{\alpha} \) approach \(-\infty\) as \((A^T y)_j\) approaches zero. A similar argument can be made for \( \alpha = 1 \), in which case the Lagrangian dual is \( g(y) = - \langle 1, y \rangle + n + \sum_{j=1}^n \log(A^T y)_j \). In [29], this was handled by ensuring that \((A^T y)_j\) never becomes "too small," which requires step sizes that are smaller by a factor \( \epsilon \) and generally leads to much slower convergence.

4 Fair Covering

In this section, we show how to reduce the fair covering problem to the \( \alpha < 1 \) case from Section C.1. We will be assuming throughout that \( \beta \geq \frac{\epsilon/4}{\log(mnp/\epsilon)} \), as otherwise the problem can be reduced to the linear
covering (see, e.g., [16]). Note that the only aspect of the analysis that relies on $\beta$ being “sufficiently small” in the $\alpha \in [0, 1)$ case is to ensure that $f_c$ closely approximates $-f_\alpha$ around the optimum of $(\alpha P), (\alpha P_c)$. Here, we will need to choose $\beta'$ to be “sufficiently small” to ensure that the lower bound from the $\alpha < 1$ case closely approximates $-g_\beta$ around the optimum $y^*$.

To apply the analysis from Section C.1 (setting $\alpha = 0$) and obtain an approximate solution to ($\beta C$), we need to ensure that: (i) the initial gap is at most $O(1)(1+\beta)g_\beta(y^*)$ and (ii) the solution $(1+\epsilon)y_\beta^{(K)}$ returned by the algorithm after $K$ iterations is ($\beta C$)-feasible – namely, $y_\beta^{(K)} \geq 0$ and $A^T y_\beta^{(K)} \geq (1+\epsilon)^{-1}1$. The algorithm pseudocode is provided in Algorithm 2 (FAIRCOVERING). The main convergence result for the fair covering is provided in the following theorem. Its proof can be found in Appendix D.

**Theorem 4.1.** The solution $y_\beta^{(K)}$ produced by FAIRCOVERING after $K = 1 + \lceil 2/(\epsilon h) \rceil$ iterations satisfies $A^T y_\beta^{(K)} \geq (1-\epsilon/2)1$ and $g_\beta(y_\beta^{(k)}) - g_\beta(y^*_\beta) \leq 3\epsilon(1 + \beta)g_\beta(y^*_\beta)$.

### 5 Conclusion

We presented efficient width-independent distributed algorithms for solving the class of $\alpha$-fair packing and covering problems. This class contains the unfair case of packing and covering LPs, for which we obtain convergence times that match that of the best known packing and covering LP solvers [1, 16, 28]. Our results greatly improve upon the only known width-independent solver for the general $\alpha$-fair packing [29], both in terms of simplicity of the convergence analysis and in terms of the resulting convergence time.

Nevertheless, several open problems merit further investigation. First, the phenomenon of width-independence is still not fully understood. Hence, understanding the problem classes on which obtaining width-independent solvers is possible is a promising research direction. The results in this context are interesting both from theoretical and practical perspectives: on one hand, width-independence is surprising as it cannot be obtained by applying results from first-order convex optimization in a black-box manner, while on the other, packing and covering constraints model a wide range of problems encountered in practice. Second, we believe that our results for the $\alpha \leq 1$ cases are not completely tight – given the results for the fair covering, we would expect the convergence time to scale as $O((1+\beta)\log(mn\rho))$, where $O$ hides the poly-logarithmic terms. However, obtaining such results will require new ideas, possibly relying on the Lagrangian duality. We remind the reader of the obstacles in using Lagrangian duality in these cases discussed at the end of Section 3. Finally, we expect that fair packing and covering solvers will be useful as primitives in solving more complex optimization problems, and may even have non-trivial connections to fairness in machine learning, due to the similar connections to political philosophy [10].
A Omitted Proofs from Section 2

Proof of Proposition 2.4. The proof of the first part follows immediately by solving:

\[ \psi^*(AF_\alpha(x) - 1) = \max_{y \geq 0} \{ \langle AF_\alpha(x) - 1, y \rangle - \psi(y) \} = \max_{y \geq 0} \left\{ \{ \langle AF_\alpha(x), y \rangle - \frac{1}{C^{\beta}} \sum_{i=1}^{m} y_i^{1+\beta} \right\} \]

which is solved for \( y_i = C(AF_\alpha(x))^i_1^{1/\beta} \), and leads to \( \psi^*(AF_\alpha(x) - 1) = \frac{C^{\beta}}{1+\beta} \sum_{i=1}^{m} (AF_\alpha(x))^i_1^{1+\beta} \).

Hence, using Proposition 2.3, \( \frac{\beta}{1+\beta} \sum_{i=1}^{m} (AF_\alpha(x))^i_1^{1+\beta} \leq \frac{\epsilon}{2} (1 - \alpha) f_\alpha(x^*_\alpha) \). As \( f_\alpha((1 - \epsilon)x^*_\alpha) = (1 - \epsilon)^{1-\alpha} f_\alpha(x^*_\alpha) \geq (1 - 3\alpha(1-\alpha)/2) f_\alpha(x^*_\alpha) \) for \( \alpha \neq 1 \) and \( f_\alpha((1 - \epsilon)x^*_\alpha) = n \log(1 - \epsilon) + f_\alpha(x^*_\alpha) \geq -\frac{3}{2} \epsilon n + f_\alpha(x^*_\alpha) \), it follows that \( f_r(x^*_\alpha) \leq f_r(x) = -f_\alpha((1 - \epsilon)x^*_\alpha) + \frac{\beta}{1+\beta} \sum_{i=1}^{m} (AF_\alpha(x))^i_1^{1+\beta} \leq -f_\alpha(x^*_\alpha) + 2\epsilon f_r \).

Finally, the \((\alpha P)\)-feasibility of \( x^*_\alpha \) (and, by the change of variables, \((\alpha P)\)-feasibility of \( x^*_\alpha \)) follows from Proposition 3.2 and Lemma 3.1 that are stated and proved in the next section. \( \square \)

B Omitted Proofs from Section 3

Proof of Lemma 3.1.

Proof of Part 1. Writing a Taylor approximation of \( f_r(x + \Gamma x) \), we have:

\[ f_r(x + \Gamma x) = f_r(x) + \langle \nabla f_r(x), \Gamma x \rangle + \frac{1}{2} \langle \nabla^2 f_r(x + t\Gamma x) \Gamma x, \Gamma x \rangle, \]  

for some \( t \in [0, 1] \). The gradient and the Hessian of \( f_r \) are given by:

\[ \nabla_j f_r(x) = \frac{1}{1-\alpha} \left( -1 + \sum_i A_{ij} x_j \frac{1}{1-\alpha} - 1 C(AF_\alpha(x))^i_1^{1/\beta} \right) \]  

(B.2)

\[ \nabla^2_{jk} f_r(x) = \frac{\alpha}{(1-\alpha)^2} \sum_i A_{ij} x_j \frac{1}{1-\alpha} - 1 C(AF_\alpha(x))^i_1^{1/\beta} \]

(B.3)

To have the control over the change in the function value, we want to enforce that the Hessian of \( f_r \) does not change by more than a factor of two in one step. To do so, let \( \gamma_m \) be the maximum (absolute) multiplicative update. Then, to have \( \nabla^2_{jk} f_r(x + \Gamma x) \leq 2 \nabla^2_{jk} f_r(x) \), it is sufficient to enforce: (i) \( (1 + \gamma_m) \frac{1}{1-\alpha} - 2^{1/\beta} \frac{1}{(1-\alpha)} \leq 2 \) (from the first term in (B.3)) and (ii) \( (1 + \gamma_m) \frac{1}{1-\alpha} - 2^{1/\beta} \frac{1}{(1-\alpha)} \leq 2 \) (from the second term in (B.3)). Combining (i) and (ii), it is not hard to verify that it suffices to have:

\[ \gamma_m \leq \frac{\beta |1-\alpha|}{2(1+\alpha/\beta)}. \]  

(B.4)
Assume from now on that $|γ_j| ≤ γ_m ≤ \frac{β(1-α)}{2(1+αβ)}$, ∀j. Then, we have:

$$\frac{1}{2} \langle ∇^2 f_r(x + tΓx)Γx, Γx \rangle ≤ \langle ∇^2 f_r(x)Γx, Γx \rangle$$

$$= \sum_j \frac{α}{(1-α)^2} \sum_i \mathcal{γ}_j^2 A_{ij} x_j \frac{1}{1-α} C(\mathbf{A} f(\mathbf{x})) \frac{β}{i} + \frac{1}{1-α} \sum_i \mathcal{C}(\mathbf{A} f(\mathbf{x})) \frac{β}{i} (\mathbf{A}Γx \frac{1}{1-α})^2.$$

Observe that, by Cauchy-Schwartz Inequality,

$$\langle ∇ f_r(x), Γx \rangle^2 = (\sum_j A_{ij} x_j \frac{1}{1-α} γ_j)^2 ≤ (∑ \mathbf{A} f(\mathbf{x}))^2 \sum_i A_{ij} x_j \frac{1}{1-α} γ_j^2.$$

Therefore:

$$\frac{1}{2} \langle ∇^2 f_r(x + tΓx)Γx, Γx \rangle \leq \frac{α + 1/β}{(1-α)^2} C \sum_i (∑ \mathbf{A} f(\mathbf{x}))^2 \frac{β}{i} \sum_j A_{ij} x_j \frac{1}{1-α} γ_j^2$$

$$= \frac{1 + αβ}{(1-α)^2} \sum_j γ_j^2 x_j ((1-α)∇_j f_r(x) + 1).$$

Since $⟨∇ f_r(x), Γx⟩ = \sum_j γ_j x_j ∇_j f_r(x)$ and $|∇_j f_r(x)| ≤ 2 \left| \frac{(1-α)∇_j f_r(x)}{1+αβ} \right|$, choosing $γ_j = -\frac{c_j}{4}$.

$$\frac{β(1-α)}{1+αβ} ∇_j f_r(x)$$

and combining (B.5) and (B.1):

$$f_r(x + Γx) - f_r(x) ≤ \frac{β(1-α)}{1+αβ} \sum_j \frac{c_j}{4} (1 - \frac{c_j}{2}) x_j ∇_j f_r(x) ∇_j f_r(x),$$

as claimed.

**Proof of Part 2.** The proof follows the same line of argument as in the case of $α ≠ 1$ above. Recall that when $α = 1$, $f_r(x) = −⟨\mathbf{I}, \mathbf{x}⟩ + \frac{β}{1+β} \sum m_{i=1}^m (\mathbf{A} \exp(\mathbf{x}))_{i}^{β}$. Hence:

$$∇_j f_r(x) = -1 + \exp(x_j) \sum m_{i=1}^m A_{ij} C(\mathbf{A} \exp(\mathbf{x}))_{i}^{1/β},$$

$$∇_j f_r(x) = \sum_{(j=k)} \exp(x_j) \sum m_{i=1}^m A_{ij} C(\mathbf{A} \exp(\mathbf{x}))_{i}^{1/β}$$

$$+ \frac{1}{β} \exp(x_j) \exp(x_k) \sum m_{i=1}^m A_{ij} A_{ik} C(\mathbf{A} \exp(\mathbf{x}))_{i}^{1/β-1}.$$

It is not hard to verify that when $Δx_j ≤ \frac{β}{2}$, ∀j, then $∇_j f_r(x + Δx) ≤ 2∇_j f_r(x), ∀j, k$. Hence, the Taylor approximation of $f_r(x + Δx)$ gives:

$$f_r(x + Δx) ≤ f_r(x) + ∇_j f_r(x) Δx_j + ∇_j f_r(x) Δx_j Δx_j .$$

Let us bound $⟨∇^2 f_r(x)Δx, Δx⟩$, as follows:

$$⟨∇^2 f_r(x)Δx, Δx⟩ = \sum m_{j=1}^n Δx_j^2 \exp(x_j) \sum m_{i=1}^m A_{ij} C(\mathbf{A} \exp(\mathbf{x}))_{i}^{1/β}$$

$$+ \frac{1}{β} \sum m_{i=1}^m C(\mathbf{A} \exp(\mathbf{x}))_{i}^{1/β-1} \left( ∑ m_{j=1}^n A_{ij} \exp(x_j) Δx_j \right)^2$$

$$≤ \left( 1 + \frac{1}{β} \right) \sum m_{j=1}^n Δx_j^2 \exp(x_j) \sum m_{i=1}^m A_{ij} C(\mathbf{A} \exp(\mathbf{x}))_{i}^{1/β}. (B.9)$$

13
where we have used that, by Cauchy-Schwartz Inequality,
\[
\left( \sum_{j=1}^{n} A_{ij} \exp(x_j) \Delta x_j \right)^2 \leq (\mathbf{A} \exp(\mathbf{x}))_i \sum_{j=1}^{n} A_{ij} \exp(x_j) \Delta x_j^2.
\]

Observe (from (B.6)) that \( \exp(x_j) \sum_{i=1}^{n} A_{ij} (\mathbf{A} \exp(\mathbf{x}))_i^{1/\beta} = \nabla_j f_r(\mathbf{x}) \). Hence, combining (B.8) and (B.9):
\[
f_r(\mathbf{x} + \Delta \mathbf{x}) - f_r(\mathbf{x}) \leq \sum_{j=1}^{n} \Delta x_j \nabla_j f_r(\mathbf{x}) + \frac{1 + \beta}{\beta} \sum_{j=1}^{n} \Delta x_j^2 (1 + \nabla_j f_r(\mathbf{x})).
\]
As \( |\nabla_j f_r(\mathbf{x})| \leq 2 |\nabla_j f_r(\mathbf{x})| \), if \( \Delta x_j = c_j \frac{\beta}{4(1+\beta)} \) for some \( c_j \in (0, 1) \), then:
\[
f_r(\mathbf{x} + \Delta \mathbf{x}) - f_r(\mathbf{x}) \leq \sum_{j=1}^{n} \left( 1 - \frac{c_j}{2} \right) \Delta x_j \nabla_j f_r(\mathbf{x})
\leq - \frac{\beta}{1 + \beta} \sum_{j=1}^{n} \frac{c_j}{4} \left( 1 - \frac{c_j}{2} \right) \nabla_j f_r(\mathbf{x}) \nabla_j f_r(\mathbf{x}),
\]
as claimed.

\[\square\]

C Convergence Analysis for Fair Packing

C.1 Convergence Analysis for \(\alpha \in [0, 1)\)

To analyze the convergence of FAIRPACKING, we need to specify \(\phi(\cdot)\) from the lower bound \(L_k^{\alpha<1}\) introduced in Section 3. To simplify the notation, in the rest of the section, we use \(L_k\) to denote \(L_k^{\alpha<1}\). We define \(\phi\) in two steps, as follows:
\[
\phi(\mathbf{x}) \overset{\text{def}}{=} \psi(\mathbf{x}) - \left( \nabla \psi(\mathbf{x}^{(0)}) + h_0 \nabla f_r(\mathbf{x}^{(0)}), \mathbf{x} \right),
\psi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{\varepsilon} \sum_{j=1}^{n} \left( - \frac{x_j^{1-\beta'}}{1 - \beta'} + x_j \right),
\]
where \(\beta' = \frac{(1 - \alpha) \varepsilon / 4}{\log(n\varepsilon/(1 - \varepsilon))}\).

This particular choice of \(\phi\) is made for the following reasons. First, \(\epsilon \psi(\mathbf{x})\) closely approximates \(\langle \mathbf{I}, \mathbf{x} \rangle\) (up to an \(\epsilon\) multiplicative factor, unless \(\langle \mathbf{I}, \mathbf{x} \rangle\) is negligible). This will ensure that \(\frac{1}{1 - \alpha} \phi(\mathbf{x}^*)\) is within \(O(1 - \alpha)f_\alpha(\mathbf{x}^*)\), which will allow us to bound the initial gap by \(O(1 - \alpha)f_\alpha(\mathbf{x}^*)\). To understand the role of the term \(\langle \nabla \psi(\mathbf{x}^{(0)}) + h_0 \nabla f_r(\mathbf{x}), \mathbf{x} \rangle\), notice that the steps of FAIRPACKING are defined as:
\[
\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{u} \geq 0} \left\{ \sum_{\ell=0}^{k} h_\ell \left( \nabla f_r(\mathbf{x}^{(\ell)}), \mathbf{u} \right) + \phi(\mathbf{u}) \right\}.
\]
The role of the term \(\langle \nabla \psi(\mathbf{x}^{(0)}) + h_0 \nabla f_r(\mathbf{x}^{(0)}), \mathbf{x} \rangle\) is to ensure that \(\mathbf{x}^{(1)} = \mathbf{x}^{(0)}\), which will allow us to properly initialize the gap. Finally, the scaling factor \(\frac{1}{\varepsilon}\) ensures that \(\mathbf{z}^{(k)} \leq 1 + \varepsilon/2\) (see the proof of
Lemma C.13), which will allow us to argue that the steps satisfy the assumptions of Lemma 3.1. Throughout the analysis, it will be crucial to ensure that the upper bound can compensate for any decrease in the lower bound. Some of these statements are formalized in the following proposition.

**Proposition C.1.** Let \( z^{(k)} \), \( \psi(\cdot) \), and \( \phi(\cdot) \) be defined as in Equations (C.1), (C.2). Then:

1. \( \min_{\alpha \geq 0} \left\{ \sum_{\ell=0}^{k} h(\nabla J(\alpha^{(\ell)}), u) + \phi(u) \right\} = \arg \min_{u \geq 0} \left\{ \psi(u) - \langle \nabla \psi(x^{(0)}), u \rangle \right\} = x^{(0)} \).
2. \( x^{(1)} = \arg \min_{u \geq 0} \left\{ h(\nabla J(\alpha^{(0)}), u) + \epsilon \phi(u) \right\} = x^{(0)} \) and \( x^{(k+1)} = \epsilon \nabla \psi(z^{(k)}) \).
3. \( \epsilon/4 \leq z^{(0)} \leq \epsilon/2 \).

**Proof:** The first part follows directly from the definitions of \( z^{(k)} \) and \( \phi \), using the first-order optimality condition to solve the minimization problem that defines \( \psi \).

For the second part, using the definition of \( \psi \) and the first-order optimality condition:

\[
\arg \min_{u \geq 0} \left\{ h(\nabla J(\alpha^{(0)}), u) + \phi(u) \right\} = \arg \min_{u \geq 0} \left\{ \psi(u) - \langle \nabla \psi(x^{(0)}), u \rangle \right\} = x^{(0)}.
\]

Similarly, for \( x^{(k+1)} \), we have \( x^{(k+1)} = \arg \min_{u \geq 0} \left\{ z^{(k)} + \epsilon \psi(u) \right\} \). It is not hard to verify (using the first-order optimality condition) that \( x^{(k+1)} = (1 + z_{j}^{(k)})^{-\beta} = \epsilon \nabla \psi(z^{(k)}) \).

For the last part, recall that \( x^{(0)} = \left( \frac{1-\epsilon}{n-p} \right)^{\frac{1}{1-\epsilon}} \) and observe that \( \nabla \psi(x) = \frac{1}{\epsilon} (1 - x^{-\beta}) \). Hence:

\[
z_{j}^{(0)} = \left( \frac{n-p}{1-\epsilon} \right)^{\frac{\beta}{1-\epsilon}} - 1 = \left( \frac{n-p}{1-\epsilon} \right)^{\frac{\epsilon/4}{\epsilon}} - 1 = \exp(\epsilon/4) - 1.
\]

The rest of the proof follows by approximating \( \exp(\epsilon/4) \).

Using Proposition C.1, we can now bound the initial gap, as follows.

**Proposition C.2.** Let \( h_{0} = H = 1 \). Then \( H_{0}G_{0} - 2(1 - \alpha) f_{\alpha}(x^{*}) \leq 2 f_{\alpha}(x^{*}) \).

**Proof:** From Proposition C.1, \( U_{1} = f_{1}(x^{(1)}) = f_{1}(x^{(0)}) \) and thus: \( H_{0}G_{0} = -\widehat{\phi}(z^{(0)}) \leq 2(1 - \alpha) f_{\alpha}(x^{*}) \). The rest of the proof follows by bounding \( \widehat{\psi}(z^{(0)}) \) and \( \phi(x^{*}) \). For the former, it is not hard to verify that:

\[
\sum_{j=1}^{n} \frac{z_{j}^{(0)}}{z_{j}^{(0)} - \beta} \leq (1 + \epsilon/2) \langle 1, x^{(0)} \rangle.
\]

Hence, as \( x^{(1)} = x^{(0)} = (z_{j}^{(0)})^{-1/\beta} \), we have:

\[
-\widehat{\psi}(z^{(0)}) \leq \frac{\beta(1 + \epsilon/2)}{\epsilon} \langle 1, x^{(0)} \rangle \leq \frac{1}{2} (1 - \alpha) \langle 1, x^{(0)} \rangle \leq \frac{1}{2} (1 - \alpha)^{2} f_{\alpha}(x^{*}).
\]

For the latter, observe first that as \( x^{*} \leq 1 \) (by feasibility, Proposition 3.2), it must be \( \psi(x^{*}) \leq 0 \). Hence, we can finally bound \( \phi(x^{*}) \) as:

\[
\phi(x^{*}) \leq -\langle \nabla \psi(x^{*}) + h_{0} \nabla J_{\alpha}(x^{(0)}), x^{*} \rangle \leq \langle (-1/2 - 1) 1, x^{*} \rangle \leq \frac{3}{2} (1, x^{*}) \leq \frac{3}{2} (1 - \alpha) f_{\alpha}(x^{*}),
\]

as \( \nabla J_{\alpha}(x) \geq -1, \forall x \) and \( \nabla \psi(x^{(0)}) = z^{(0)}/\epsilon \geq - (1/2) 1 \) (due to Proposition C.1).
The crucial part of the convergence analysis is to show that for some choice of step sizes $h_k$, $H_k G_k \leq H_{k-1} G_{k-1} + 2 h_k \epsilon_f$. Note that to make the algorithm as fast as possible (since its convergence rate is proportional to $H_k$), we would like to set $h_k$ as large as possible. However, enforcing the condition $H_k G_k \leq H_{k-1} G_{k-1} + 2 h_k \epsilon_f$ will set an upper bound on the choice of $h_k$. We have the following lemma.

**Lemma C.3.** (Main Lemma.) If $G_{k-1} - 2 \epsilon_f \leq 2 f_{\alpha}(x_{n}^k)$ and $h_k \leq \frac{(1-\alpha)\beta}{\log(\epsilon_f \log(m \epsilon_f / 2))}$, then $H_k G_k \leq H_{k-1} G_{k-1} + 2 h_k \epsilon_f$, $\forall k \geq 1$.

**Proof.** The role of the assumption $G_{k-1} - 2 \epsilon_f \leq 2 (1-\alpha) f_{\alpha}(x_{n}^k)$ is to guarantee that $z_{j(k-1)} \geq - (\epsilon_f/2)$. Namely, if $z_{j(k-1)} < - \epsilon_f/2$, for any $j$, $\psi^x(z_{j(k-1)})$ blows up, making the gap $G_{k-1}$ much larger than $3 (1-\alpha) f_{\alpha}(x_{n}^k)$. This is not hard to argue (see also a similar argument in [16]) and hence we omit the details and assume from now on that $z_{j(k-1)} \geq - (\epsilon_f/2)$. Note that this assumption holds initially due to Proposition C.1. Observe that as $\epsilon < 1/H_k$ and $\nabla J_{f}(x_{\ell}) \leq 1$, $\forall \ell$, we also have $z_{j(k)} \leq 1 + \epsilon_f/2$.

To be able to apply Lemma 3.1, we need to ensure that $|x_{j(k+1)} - x_{j(k)}| \leq c_j \frac{\beta}{4(1+\beta)} \nabla J_{f}(x_{k})$, for all $j$ and for $c_j \in (0, 1]$. Recalling the definition of $x_{j(k+1)}$, $x_{j(k+1)} = \nabla \nabla \tilde{\psi}(z_{j(k)}) = (1 + z_{j(k)})^{-1/\beta}$. As $z_{j(k)} = z_{j(k-1)} + h_k \nabla J_{f}(x_{k})$, we have:

$$x_{j(k+1)} = (1 + z_{j(k-1)} + h_k \epsilon_0 \nabla J_{f}(x_{k}))^{-1/\beta} = x_{j(k)} \left( 1 + \frac{h_k \nabla J_{f}(x_{k})}{1 + z_{j(k-1)}} \right)^{-1/\beta'}.
$$

Suppose first that $\nabla J_{f}(x_{k}) \leq 0$. Then $\frac{\nabla J_{f}(x_{k})}{1 + z_{j(k-1)}} \leq \frac{\nabla J_{f}(x_{k})}{1 + z_{j(k-1)}} \leq \frac{\nabla J_{f}(x_{k})}{2 + \epsilon_f / 2}$. As $h_k \leq \frac{(1-\alpha)\beta}{8(1+\beta)}$ and $\nabla J_{f}(x_{k}) \geq 1$, we have:

$$1 - \frac{1 - \epsilon_f / 2}{2 - \epsilon_f / 2} \cdot \frac{(1-\alpha)\beta}{8(1+\beta)} \nabla J_{f}(x_{k}) \leq \left( 1 + \frac{h_k \nabla J_{f}(x_{k})}{1 + z_{j(k-1)}} \right)^{-1/\beta'} \leq 1 - \frac{(1-\alpha)\beta}{4(1+\beta)} \nabla J_{f}(x_{k}).
$$

Similarly, when $\nabla J_{f}(x_{k}) > 0$, $\frac{\nabla J_{f}(x_{k})}{1 + z_{j(k-1)}} \leq \frac{\nabla J_{f}(x_{k})}{1 - \epsilon_f / 2}$. As $h_k \leq \frac{(1-\alpha)\beta}{8(1+\beta)}$ and $\nabla J_{f}(x_{k}) \leq 1$, we have:

$$1 - \frac{(1-\alpha)\beta}{4(1+\beta)} \nabla J_{f}(x_{k}) \leq \left( 1 + \frac{h_k \nabla J_{f}(x_{k})}{1 + z_{j(k-1)}} \right)^{-1/\beta'} \leq 1 - \frac{1 - \epsilon_f / 2}{2 - \epsilon_f / 2} \cdot \frac{(1-\alpha)\beta}{8(1+\beta)} \nabla J_{f}(x_{k}).
$$

Either way, Lemma 3.1 can be applied with $c_j \geq \frac{1 - \epsilon_f / 2}{2(1 + \epsilon_f / 2)} \geq 1/10$, and we have:

$$H_k U_k - H_{k-1} U_{k-1} \leq h_k \nabla J_{f}(x_{k}) = H_k \beta \frac{(1-\alpha)\beta}{50(1+\alpha\beta)} \sum_{j=1}^{n} x_{j(k)} \nabla J_{f}(x_{k}) \nabla J_{f}(x_{k}).$$

(C.3)

On the other hand, the change in the lower bound is:

$$H_k L_k - H_{k-1} L_{k-1} = h_k \left( f_{\alpha}(x_{k}) - \left( \nabla J_{f}(x_{k}), x_{k} \right) \right) + \frac{1}{1 - \alpha} \left( \tilde{\psi}(z_{k}) - \tilde{\psi}(z_{k-1}) \right) + 2 h_k \epsilon_f.
$$

(C.4)

Using Taylor’s Theorem:

$$\tilde{\psi}(z_{k}) - \tilde{\psi}(z_{k-1}) = \left( \nabla \tilde{\psi}(z_{k-1}), z_{k} - z_{k-1} \right) + \frac{1}{2} \left( \nabla^2 \tilde{\psi}(\tilde{z}) (z_{k} - z_{k-1}), z_{k} - z_{k-1} \right),
$$

(C.5)
where \( \hat{z} = z^{(k-1)} + t(z^{(k)}) - z^{(k-1)} \), for some \( t \in [0, 1] \). Recall that \( \epsilon \nabla_j \hat{\psi}(z^{(k-1)}) = (1 + z_j^{(k-1)})^{-1/\beta'} = x_j^{(k)} \) and \( z^{(k)} - z^{(k-1)} = \epsilon_{hk} \nabla_j r(z^{(k)}) \). Observe that \( \nabla^2_{jj} \hat{\psi}(z) = -\frac{1}{\epsilon h} (1 + z_j)^{-1+(1+\beta)/\beta'} \), \( \nabla^2_{jk} \hat{\psi}(z) = 0 \), for \( j \neq k \). As \( z_j^{(k-1)} \geq -\epsilon/2 \) and \( z_j^{(k)} = z_j^{(k-1)} + \epsilon h_k \nabla_j r(x^{(k)}) \geq 1 - \epsilon h_k \), we have that:

\[
(1 + z_j^{(k)})^{-1/\beta'} \leq (1 - \epsilon h_k) - \frac{1}{\epsilon h} (1 + z_j^{(k-1)})^{-1/\beta'} < (1 + \epsilon/2)(1 + z_j^{(k-1)})^{-1/\beta'},
\]
as \( \frac{\epsilon h_k}{\beta'} \leq (1 - \epsilon/2) \frac{(1-\alpha)\beta}{8(1+\beta')} \). Further, as \( x_j^{(k)} = (1 + z_j^{(k-1)})^{-1/\beta'} \) and \( z_j^{(k)} \geq 1 - \epsilon/2 \), we have that \( (1 + z_j^{(k-1)})^{-1/\beta'} \leq x_j^{(k)}/(1 - \epsilon/2) \). Hence, (C.5) implies:

\[
\hat{\psi}(z^{(k)}) - \hat{\psi}(z^{(k-1)}) \geq \epsilon h_k \left( \nabla^2_{j} f_r(x^{(k)}) \right)^2 - \frac{3(\epsilon h_k)^2}{2\beta'} \sum_{j=1}^{n} x_j^{(k)} \left( \nabla^2_{jk} f_r(x^{(k)}) \right)^2. \tag{C.6}
\]

Combining (C.3), (C.4), and (C.6), to complete the proof, it suffices to show that, \( \forall j \),

\[
\xi_j \overset{\text{def}}{=} h_k \left( \nabla_j f_r(x^{(k)}) - \frac{1}{1 - \alpha} \nabla_j f_r(x^{(k)}) \right) + \frac{3(\epsilon h_k)^2}{2\beta'} \left( \nabla^2 f_r(x^{(k)}) \right)^2 - \frac{H_k \beta}{50(1 + \alpha \beta)} \nabla_j f_r(x^{(k)}) \nabla^2 f_r(x^{(k)}) \leq 0.
\]

Consider the following two cases for \( \nabla_j f_r(x^{(k)}) \):

**Case 1:** \( \nabla_j f_r(x^{(k)}) < 1 \). Then \( \nabla_j f_r(x^{(k)}) = (1 - \alpha) \nabla_j f_r(x^{(k)}) \), and we have:

\[
\xi_j = \frac{\left( \nabla^2 f_r(x^{(k)}) \right)^2}{1 - \alpha} \left( \frac{3(\epsilon h_k)^2}{2\beta'} \frac{H_k \beta}{50(1 + \alpha \beta)} \right) < 0,
\]
as \( \epsilon h_k \leq (1 - \epsilon/2) \frac{(1-\alpha)\beta}{8(1+\beta')} \).

**Case 2:** \( \nabla_j f_r(x^{(k)}) = 1 \). Then \( \nabla_j f_r(x^{(k)}) \geq 1 \). Then:

\[
\xi_j = h_k \left( \frac{3(\epsilon h_k)^2}{\beta'} - 1 \right) + \nabla_j f_r(x^{(k)}) \left( h_k - \frac{H_k \beta}{50(1 + \alpha \beta)} \right) \leq 0,
\]
by the choice of \( h_k \).

We are now ready to bound the overall convergence of FAIRPACKING for \( \alpha < 1 \).

**Theorem C.4.** Let \( h_0 = 1 \), \( h_k = h = \frac{(1-\alpha)\beta}{16(1+\beta)} \) \( k \geq 1 \). Then, after at most \( K = \lceil \frac{2}{H(1-\alpha)\epsilon} \rceil = \theta \left( \frac{\log(h_0)}{\log(h_0/\epsilon)} \right) \) iterations of FAIRPACKING, we have that \( x^{(k+1)} = F_{\alpha}(x^{(k+1)}) \) is \( (\alpha P) \)-feasible and:

\[
f_{\alpha}(x^{(K+1)}) - f_{\alpha}(x^{*}) \geq -3\epsilon(1 - \alpha) f_{\alpha}(x^{*}).
\]

**Proof.** Feasibility of \( x^{(K+1)} \) follows from Proposition 3.2, as the steps of FAIRPACKING satisfy the conditions of Lemma 3.1.

Due to Proposition C.2, the assumptions of Lemma C.3 hold initially and hence they hold for all \( k \) (as Lemma C.3 itself when applied to iteration \( k \) implies that its assumptions hold at iteration \( k + 1 \)). Thus, we have: \( G_K \leq H_{0}G_{0} + \sum_{t=0}^{K-1} \frac{H_{2}\epsilon f}{H_{K}} = H_{0}G_{0} + 2\epsilon f \). As, from Proposition C.2, \( H_{0}G_{0} \leq 2f_{\alpha}(x^{*}) \) and \( H_{K} = Kh \geq \frac{2}{(1-\alpha)\epsilon} \), it follows that \( G_{K} \leq 3\epsilon(1 - \alpha) f_{\alpha}(x^{*}) \). Finally, recalling that by construction, \( -f_{\alpha}(x^{(K+1)}) + f_{\alpha}(x^{*}) \leq G_{K} \), the claimed statement follows. \( \square \)
C.2 Convergence Analysis for $\alpha = 1$

Let us start by bounding the coordinates of the running solutions $x^{(k)}$, for each iteration $k$. This will allow us to bound the initial-gap-plus-error $H_0 G_0 + \sum_{k} E_i$ in the convergence analysis.

**Proposition C.5.** In each iteration $k$, $-\log(2\rho m C) I \leq x^{(k)} \leq 0$.

**Proof.** Using Proposition 3.2, $x^{(k)} \leq 0$ follows immediately by $\min_{ij} A_{ij} \neq 0 A_{ij} = 1$.

Suppose that in some iteration $k$, $x^{(k)} \leq \log(\frac{1}{2\rho m C}) + \epsilon / 4$. Then, using Proposition 3.2:

$$\nabla_j f_r(x^{(k)}) \leq -1 + C \cdot \frac{1}{2\rho m C} \exp(\epsilon / 4) \rho m \leq -\frac{1 - \epsilon}{2}.$$  

Hence, $x^{(k)}$ must increase in iteration $k$. Since the maximum decrease in any coordinate and in any iteration is by less than $\epsilon / 4$, it follows that $x^{(k)} \geq -\log(2\rho m C) I$, as claimed. \hfill $\square$

Recall that $U_k = f_r(x^{(k+1)})$ and $L_k = \sum_{i=0}^{k} h_k (f(x^{(i)}) + \langle \nabla f_r(x^{(i)}), x^* - x^{(i)} \rangle)$. Using the Cauchy-Schwartz Inequality:

$$H_0 L_0 \geq h_0 f(x^{(0)}) - h_0 \|\nabla f_r(x^{(0)})\| \cdot \|x^* - x^{(0)}\| - 2H_0 \epsilon n,$$  

while, from Lemma 3.1,

$$H_0 U_0 \leq h_0 f(x^{(0)}) - H_0 \left( \frac{\beta}{8(1 + \beta)} \|\nabla f_r(x^{(0)})\|^2 \right).$$  

Combining (C.7) and (C.8) with $-a^2 + 2ab \leq b^2$, $\forall a, b$, and as $H_0 = h_0$, it follows that:

$$H_0 G_0 = H_0 (U_0 - L_0) \leq 2(1 + \beta) \cdot \frac{\beta}{h_0} \|x^* - x^{(0)}\|^2 + 2h_0 \epsilon n,$$  

as claimed. \hfill $\square$

The main part of the analysis is to show that for $k \geq 1$, $H_k G_k - H_{k-1} G_{k-1} \leq E_k$, which, combined with Proposition C.6 and the definition of the gap would imply $f(x^{(k+1)}) - f(x^*) \leq G_k \leq \sum_{k=0}^{n} E_i$, allowing us to bound the approximation error. This is done in the following lemma.

**Lemma C.7.** If, for $k \geq 1$, $\frac{h_k}{H_k} \leq \frac{\beta}{8(1 + \beta) \log(2\rho m C)}$, then $H_k G_k - H_{k-1} G_{k-1} \leq E_k$, where $E_k = \frac{2(1 + \beta)}{\beta} \cdot \frac{h_0^2}{H_0} \|x^* - x^{(k)}\|^2$.

**Proof.** By the definition of the lower bound and Cauchy-Schwartz Inequality:

$$H_k L_k - H_{k-1} L_{k-1} \geq h_k f_r(x^{(k)}) - h_k \sum_{j=1}^{n} \|\nabla_j f_r(x^{(k)})\| \cdot |x_j^* - x_j^{(k)}|,$$  

(C.9)
while, by Lemma 3.1,
\[ H_k U_k - H_{k-1} U_{k-1} \leq h_k f_r(x^{(k)}) - \frac{H_k \beta}{8(1 + \beta)} \sum_{j=1}^n \nabla_j f_r(x^{(k)}) \nabla_j f_r(x^{(k)}). \] (C.10)

Hence, combining (C.9) and (C.10):
\[ H_k G_k - H_{k-1} G_{k-1} \leq \sum_{j=1}^n \left( h_k |\nabla_j f_r(x^{(k)})| \cdot |x_j^* - x_j^{(k)}| - \frac{H_k \beta}{8(1 + \beta)} \nabla_j f_r(x^{(k)}) \nabla_j f_r(x^{(k)}) \right). \] (C.11)

Let \( e_j = h_k |\nabla_j f_r(x^{(k)})| \cdot |x_j^* - x_j^{(k)}| - \frac{H_k \beta}{8(1 + \beta)} \nabla_j f_r(x^{(k)}) \nabla_j f_r(x^{(k)}) \) be the \( j \)th term in the summation from the last equation, and consider the following two cases.

**Case 1:** \( \nabla_j f_r(x^{(k)}) \leq 1. \) Then \( \nabla_j f_r(x^{(k)}) = \nabla_j f_r(x^{(k)}) \) and using that \(-a^2 + 2ab \leq b^2, \forall a, b:
\[ e_j \leq \frac{2(1 + \beta) h_k^2}{\beta} (x_j^* - x_j^{(k)})^2. \] (C.12)

**Case 2:** \( \nabla_j f_r(x^{(k)}) > 1. \) Then \( \nabla_j f_r(x^{(k)}) = 1. \) From Proposition C.5, \(-log(2pmC) \leq x_j^{(k)} \leq 0 \) and similar bounds can be obtained for \( x_j^* \) (see [29]). It follows that:
\[ e_j \leq |\nabla_j f(x^{(k)})| \left( h_k \log(2pmC) - \frac{H_k \beta}{8(1 + \beta)} \right) \leq 0, \] (C.13)
as \( \frac{h_k}{H_k} \leq \frac{\beta}{8(1 + \beta) \log(2pmC)}. \)

Combining (C.11)-(C.13), it follows that \( H_k G_k - H_{k-1} G_{k-1} \leq \sum_{j=1}^n |\nabla_j f_r(x^{(k)})| \leq 1 \frac{2(1 + \beta) h_k^2}{\beta} (x_j^* - x_j^{(k)})^2 \leq E_k, \) as claimed.

We are now ready to obtain the final convergence bound for \( \alpha = 1 : \)

**Theorem C.8.** If \( k \geq 10 \log^2(2pmC) \frac{e^\beta}{e^{\beta}} = O\left( \frac{\log^3(4mn/\epsilon)}{\epsilon^4} \right), \) then \( x_\alpha^{(k+1)} = \exp(x^{(k+1)}) \) is \((\alpha P)\)-feasible and \( f_\alpha(x_\alpha^{(k+1)}) - f_\alpha(x_\alpha^{(k)}) \geq -3\epsilon \).

**Proof.** Combining Proposition C.6 and Lemma C.7, we have that if for \( \ell \geq 1, \frac{h_\ell}{H_\ell} \leq \lambda \) then \( G_k \leq \frac{2(1 + \beta)}{\beta \lambda} \sum_{\ell=0}^k \frac{h_\ell^2}{H_\ell} \|x^* - x^{(\ell)}\|^2 + 2n\epsilon. \) As discussed before, \( \|x^* - x^{(\ell)}\|^2 \leq n \log^2(2pmC), \) and so:
\[ G_k \leq \frac{2(1 + \beta)}{\beta} n \log^2(2pmC) \frac{1}{H_k} \sum_{\ell=0}^k \frac{h_\ell^2}{H_\ell} + 2n\epsilon. \] (C.14)

As the sequence \( \{h_\ell\}_{\ell=1}^k \) does not affect the algorithm, we can choose it arbitrarily, as long as \( \frac{h_\ell}{H_\ell} \leq \lambda \) for \( \ell \geq 1. \) Let \( h_0 = H_0 = 1 \) and \( \frac{h_\ell}{H_\ell} = \frac{\beta \epsilon}{8(1 + \beta) \log^2(2pmC)} < \lambda \) for \( \ell \geq 1. \) Then:
\[ G_k \leq \frac{1}{H_k} \frac{2(1 + \beta)}{\beta} n \log^2(2pmC) + \frac{n\epsilon}{4} + 2n\epsilon. \]

As \( \frac{1}{H_\ell} = \frac{H_0}{H_\ell} = \frac{H_0}{H_1} \frac{H_1}{H_2} \cdots \frac{H_ {k-1}}{H_k} = (1 - \frac{h_k}{H_k})^k, \) it follows that \( G_k \leq 3\epsilon. \) By construction, \(-f_\alpha(x_\alpha^{(k+1)}) + f_\alpha(x_\alpha^{(k)}) \leq 3\epsilon, \) and \( x_\alpha^{(k+1)} \) is \((\alpha P)\)-feasible due to Proposition 3.2. \( \square \)
C.3 Convergence Analysis for \( \alpha > 1 \)

Define the vector \( y^{(k)} \) as:
\[
y_i^{(k)} = (A F_\alpha(x^{(k)}))_{i} = \left( A(x^{(k)})^{-\frac{1}{1-\alpha}} \right)_i.
\]
(C.15)

Clearly, \( y^{(k)} \geq 0 \). Observe that:
\[
f_r(x^{(k)}) = -\frac{\langle 1, x^{(k)} \rangle}{1-\alpha} + \frac{\beta}{1+\beta} \sum_{i=1}^{m} (y_i^{(k)})^{1+\beta}
\]
and therefore:
\[
f_r(x^{(k)}) - \langle \nabla f_r(x^{(k)}), x^{(k)} \rangle = \left( \frac{\beta}{1+\beta} + \frac{1}{\alpha-1} \right) \sum_{i=1}^{m} (y_i^{(k)})^{1+\beta}
\]
(C.16)

Recall that the Lagrangian dual of \((\alpha P_c)\) (and, by the change of variables, \((\alpha P)\)) is \( g(y) = -\langle 1, y \rangle + \frac{\alpha}{\alpha-1} \sum_{j=1}^{n} (A^T y)^{\frac{\alpha-1}{\alpha}} \). Interpreting \( y^{(k)} \) as a dual vector, we can bound the duality gap of a solution \( x^{(k)} \) at any iteration \( k \) (using primal feasibility from Proposition 3.2) as:
\[
-f_\alpha(x^{(k)}) + f_\alpha(x^*_\alpha) = -\frac{\langle 1, x^{(k)} \rangle}{1-\alpha} + f_\alpha(x^*_\alpha) \leq -\frac{\langle 1, x^{(k)} \rangle}{1-\alpha} - g(y^{(k)}).
\]
(C.17)

We will assume throughout this section that \( \epsilon \leq \min \{ \frac{1}{2}, \frac{1}{10(\alpha-1)} \} \).

C.3.1 Regularity Conditions for the Duality Gap

The next proposition gives a notion of approximate and aggregate complementary slackness, with \( y^{(k)} \) being interpreted as the vector of dual variables, similar to [29].

**Proposition C.9.** After at most \( O\left(\frac{1}{\epsilon^2}\right) \) initial iterations, in every iteration
\[
\langle 1, y^{(k)} \rangle \leq (1 + \epsilon) \langle A^T y^{(k)}, (x^{(k)})^{-\frac{1}{\alpha}} \rangle.
\]

**Proof.** First, let us argue that after at most \( O\left(\frac{1}{\epsilon^2}\right) \) iterations, there must always exist at least one \( i \) with
\[
(A(x^{(k)})^{-\frac{1}{\alpha}})_i \geq 1 - \epsilon/2.
\]
Suppose that in any given iteration \( \max_i (A(x^{(k)})^{-\frac{1}{\alpha}})_i \leq 1 - \epsilon/4 \). Then, as
\[
x_j^{(k)} \leq 1 \text{ (by feasibility – Proposition 3.2)} \forall j, \nabla_j f_r(x^{(k)}) \geq \frac{1}{1-\alpha} (-1 + Cm \rho (1 - \epsilon/4)^{1/\beta}) \geq \frac{1}{2(\alpha-1)}.
\]
Hence, each \( x_j \) must decrease by a factor at least \( 1 - \frac{\beta(\alpha-1)}{8(1+\alpha \beta)} \), which means that \( (A(x^{(k)})^{-\frac{1}{\alpha}}) \) increases by a factor at least \( (1 - \frac{\beta(\alpha-1)}{8(1+\alpha \beta)})^{\frac{1}{1-\alpha}} \geq 1 + \frac{\beta}{8(1+\alpha \beta)} \). As in any iteration, the most any \( (A(x^{(k)})^{-\frac{1}{\alpha}})_i \) can decrease is by a factor at most \( 1 - \beta \), it follows that after at most initial \( O\left(\frac{1+\alpha \beta}{\beta}\right) \) iterations, it always holds that
\[
\max_i (A(x^{(k)})^{-\frac{1}{\alpha}})_i \geq 1 - \epsilon/2.
\]

Let \( i^* = \arg \max_i (A(x^{(k)})^{-\frac{1}{\alpha}})_i \) and \( S = \{ i : (A(x^{(k)})^{-\frac{1}{\alpha}})_i \geq (1 - \epsilon/4)(A(x^{(k)})^{-\frac{1}{\alpha}})_i \} \). Then, \( \forall \ell \notin S, \ y^{(k)}_\ell \leq (1 - \epsilon/4)^{1/\beta} y^{(k)}_i \leq \frac{4 \epsilon}{4m} y^{(k)}_i \). Hence, \( \sum_{\ell \notin S} y^{(k)}_\ell \leq \epsilon y^{(k)}_i \leq \frac{\epsilon}{4} \sum_{i \in S} y^{(k)}_i \) and we have
\[
\sum_{i \in S} y^{(k)}_i \geq \frac{1}{1+\epsilon/4} \sum_{i=1}^{m} y^{(k)}_i.
\]
It follows that:
\[
\langle y^{(k)}, A(x^{(k)})^{-\frac{1}{\alpha}} \rangle \geq \sum_{i \in S} y^{(k)}_i (A(x^{(k)})^{-\frac{1}{\alpha}})_i \geq (1 - \epsilon/2)(1 - \epsilon/4) \sum_{i \in S} y^{(k)}_i \geq \frac{(1 - \epsilon/2)(1 - \epsilon/4)}{1 + \epsilon/4} \langle 1, y^{(k)} \rangle.
\]

The rest of the proof is by \( \frac{1+\epsilon/4}{(1-\epsilon/2)(1-\epsilon/4)} \leq 1 + \epsilon. \)
To construct and use the same argument as before (namely, to guarantee that $H_k G_k \leq H_{k-1} G_{k-1} + O(\epsilon)(1 - \alpha) f_\alpha(x_\alpha^*)$ for some notion of the gap $G_k$), we need to ensure that the argument can be started from a gap $G_0 = O(1)(1 - \alpha) f_\alpha(x_\alpha^*)$. The following lemma gives sufficient conditions for ensuring constant multiplicative gap. When those conditions are not satisfied, we will show (in Lemma C.11) that $f_r(x^{(k)})$ must decrease multiplicatively, which will guarantee that there cannot be many such iterations. Define:

$$S_+ \overset{\text{def}}{=} \left\{ j : (x_j^{(k)})^{\frac{\alpha}{1-\alpha}}(A^T y^{(k)})_j \geq 1 + \frac{1}{10(\alpha-1)} \right\} \text{ and } S_- \overset{\text{def}}{=} \left\{ j : (x_j^{(k)})^{\frac{\alpha}{1-\alpha}}(A^T y^{(k)})_j \leq 1 - \frac{1}{10} \right\}.$$

**Lemma C.10.** After the initial $O\left(\frac{1}{\beta} \right)$ iterations, if all of the following conditions are satisfied:

1. $-\sum_{j \in S_+} x_j^{(k)} \left( 1 - (x_j^{(k)})^{\frac{\alpha}{1-\alpha}}(A^T y^{(k)})_j \right) \leq \frac{1}{10(\alpha-1)} \langle 1, x^{(k)} \rangle$;
2. $\sum_{j \in S_-} x_j^{(k)} \left( 1 - (x_j^{(k)})^{\frac{\alpha}{1-\alpha}}(A^T y^{(k)})_j \right) \leq \frac{1}{10(\alpha-1)} \langle 1, x^{(k)} \rangle$; and
3. $\langle y^{(k)}, A(x^{(k)})^{\frac{1}{\alpha}} \rangle \leq 2 \langle 1, x^{(k)} \rangle$

then $f_r(x^{(k)}) + f_\alpha(x_\alpha^*) \leq -2f_\alpha(x^*)$.

**Proof.** Denote $\Delta_j = (x_j^{(k)})^{\frac{\alpha}{1-\alpha}}(A^T y^{(k)})_j$. Let us start by bounding the true duality gap (using feasibility from Proposition 3.2 and approximate complementary slackness from Proposition C.9):

$$\frac{\langle 1, x^{(k)} \rangle}{\alpha-1} + f_\alpha(x_\alpha^*) \leq \frac{\langle 1, x^{(k)} \rangle}{\alpha-1} - g(y^{(k)})$$

$$\leq \frac{\langle 1, x^{(k)} \rangle}{\alpha-1} + (1 + \epsilon) \left( \langle A^T y^{(k)}, (x^{(k)})^{\frac{1}{\alpha}} \rangle - \frac{\alpha}{\alpha-1} \sum_{j=1}^n (A^T y^{(k)})_j^{\frac{\alpha-1}{\alpha}} \right)$$

$$= \frac{1}{\alpha-1} \sum_{j=1}^n x_j^{(k)} \left( 1 + (\alpha-1)\Delta_j - \alpha \Delta_j^{\frac{\alpha-1}{\alpha}} \right) + \epsilon \left\langle A^T y^{(k)}, (x^{(k)})^{\frac{1}{\alpha}} \right\rangle$$

$$= \sum_{j=1}^n \xi_j + \epsilon \left\langle A^T y^{(k)}, (x^{(k)})^{\frac{1}{\alpha}} \right\rangle,$$  

(C.18)

where $\xi_j = x_j^{(k)} \left( 1 + (\alpha-1)\Delta_j - \alpha \Delta_j^{\frac{\alpha-1}{\alpha}} \right)$. To bound the expression from (C.18), we will split the sum $\sum_{j=1}^n \xi_j$ into two: corresponding to terms with $\Delta_j \geq 1$ and corresponding to terms with $\Delta_j < 1$. For the former, as $\Delta_j^{\frac{\alpha-1}{\alpha}} \geq 1$, we have:

$$\sum_{j: \Delta_j \geq 1} \xi_j \leq \frac{1}{\alpha-1} \sum_{j: \Delta_j \geq 1} x_j^{(k)} (1 + (\alpha-1)\Delta_j - \alpha)$$

$$= \sum_{j: 1 \leq \Delta_j \leq 1 + \frac{1}{10(\alpha-1)}} x_j^{(k)} (\Delta_j - 1) + \sum_{j \in S_+} x_j^{(k)} (\Delta_j - 1)$$

$$\leq \frac{1}{5(\alpha-1)} \left\langle 1, x^{(k)} \right\rangle,$$  

(C.19)

where the last inequality is by $\sum_{j: 1 \leq \Delta_j \leq 1 + \frac{1}{10(\alpha-1)}} x_j^{(k)} (\Delta_j - 1) \leq \frac{\langle 1, x^{(k)} \rangle}{5(\alpha-1)}$ and the first condition from the statement of the lemma.
Consider now the terms with $\Delta_j < 1$. As $\Delta_j^{\alpha-1} \geq \Delta_j$:

$$
\sum_{j: \Delta_j < 1} \xi_j \leq \frac{1}{\alpha - 1} \sum_{j: \Delta_j < 1} x_j^{(k)} \left( 1 + (\alpha - 1)\Delta_j - \alpha \Delta_j \right)
= \frac{1}{\alpha - 1} \sum_{j: \Delta_j < 1} x_j^{(k)} (1 - \Delta_j)
= \frac{1}{\alpha - 1} \sum_{j: 1 - \frac{1}{\alpha} < \Delta_j < 1} x_j^{(k)} (1 - \Delta_j) + \frac{1}{\alpha - 1} \sum_{j \in S_0} x_j^{(k)} (1 - \Delta_j)
\leq \frac{1}{5(\alpha - 1)} \left\langle 1, x^{(k)} \right\rangle.
$$

(C.20)

The third condition from the statement of the lemma guarantees that $\epsilon \left\langle A^T y^{(k)}, (x^{(k)})^{1-\alpha} \right\rangle \leq 2 \epsilon \left\langle 1, x^{(k)} \right\rangle \leq \frac{1}{5(\alpha - 1)} \left\langle 1, x^{(k)} \right\rangle$, as $\epsilon \leq \frac{1}{10(\alpha - 1)}$. Hence, combining (C.18)-(C.20):

$$
\left\langle 1, x^{(k)} \right\rangle + f_\alpha(x^{(k)}_\alpha) \leq \frac{3}{5(\alpha - 1)} \left\langle 1, x^{(k)} \right\rangle,
$$

Rearranging the terms in the last equation, $\left\langle 1, x^{(k)} \right\rangle - \frac{3}{5(\alpha - 1)} \left\langle 1, x^{(k)} \right\rangle \leq -\frac{5}{2} f_\alpha(x^{(k)}_\alpha)$. From the third condition in the statement of the lemma, $f_r(x^{(k)}) \leq \frac{1}{\alpha(\alpha - 1)} \left( 1 + \frac{10\beta(\alpha - 1)}{1+\beta} \right) \leq \frac{1}{\alpha(\alpha - 1)} \left( 1 + \frac{10\beta}{2} \right) \leq \frac{21}{20} \frac{1}{\alpha(\alpha - 1)}$ as $\beta \leq \epsilon/4$ and $\epsilon \leq \frac{1}{10(\alpha - 1)}$. Putting everything together:

$$
f_r(x^{(k)}) + f_\alpha(x^{(k)}_\alpha) \leq \frac{13}{20(\alpha - 1)} \left\langle 1, x^{(k)} \right\rangle \leq -\frac{5}{2} \frac{13}{20} f_\alpha(x^{(k)}_\alpha) \leq -2 f_\alpha(x^{(k)}_\alpha),
$$

as claimed. \qed

**Lemma C.11.** If in iteration $k$ any of the conditions from Lemma C.10 does not hold, then $f_r(x^{(k)})$ must decrease by a factor at most $\max \{ 1 - \theta(\beta(\alpha - 1)), 1 - \theta(\beta) \min \left\{ \frac{1}{10(\alpha - 1)}, 1 \right\} \}$.

**Proof.** If the conditions from Lemma C.10 do not hold, then we must have (at least) one of the following cases.

**Case 1:** $- \sum_{j \in S_+} x_j^{(k)} \left( 1 - x_j^{(k)} \right)^{1-\alpha} (A^T y^{(k)})_{j} \geq \frac{1}{10(\alpha - 1)}$. Observe that, by the definition of $S_+$, for all $j \in S_+$, $\nabla_j f_r(x^{(k)}) \geq \min \left\{ \frac{1}{10(\alpha - 1)}, 1 \right\}$ and $(1 - \alpha) \nabla_j f_r(x^{(k)}) \geq \frac{1}{10(\alpha - 1)} > 0$. From Lemma 3.1:

$$
f(x^{(k+1)}) - f(x^{(k)}) \leq -\frac{\beta(1 - \alpha)}{8(1 + \alpha \beta)} \sum_{j \in S_+} x_j^{(k)} \nabla_j f(x^{(k)}) \nabla_j f_r(x^{(k)})
\leq \min \left\{ \frac{1}{10(\alpha - 1)}, 1 \right\} \frac{\beta}{8(1 + \alpha \beta)} \sum_{j \in S_+} x_j^{(k)} \left( 1 - x_j^{(k)} \right)^{1-\alpha} (A^T y^{(k)})_{j}
\leq -\min \left\{ \frac{1}{10(\alpha - 1)}, 1 \right\} \frac{\beta}{80(\alpha - 1)(1 + \alpha \beta)} \left\langle 1, x^{(k)} \right\rangle.
$$

Assume that $\left\langle y^{(k)}, A(x^{(k)})^{1-\alpha} \right\rangle \leq 2 \left\langle 1, x^{(k)} \right\rangle$ (otherwise we would have Case 3 below). Then $f_r(x^{(k)}) \leq \left( \frac{1}{\alpha - 1} + \frac{2\beta}{1+\beta} \right) \left\langle 1, x^{(k)} \right\rangle$, and, hence $\left\langle 1, x^{(k)} \right\rangle \geq \left( \frac{1}{\alpha - 1} + \frac{2\beta}{1+\beta} \right)^{-1} f_r(x^{(k)}) = \frac{(\alpha - 1)(1+\beta)}{1+\beta+2(\alpha-1)\beta} f_r(x^{(k)}) \geq$
\[
\frac{\alpha - 1}{2} f_r(x^{(k)}) \text{. Therefore, it follows that } f(x^{(k+1)}) - f(x^{(k)}) \leq -\theta \left( \beta \min \left\{ \frac{1}{10(\alpha - 1)}, 1 \right\} \right) f_r(x^{(k)}) \text{.}
\]

**Case 2:** \( \sum_{j \in S_-} x_j^{(k)} \left( 1 - (x_j^{(k)})^{-\alpha} (A_T^T y^{(k)})_j \right) > \frac{1}{10} \langle 1, x^{(k)} \rangle \). Observe that, by the definition of \( S_- \), for all \( j \in S_- \), \( \nabla_j f_r(x^{(k)}) \leq -\frac{1}{10} \) and \( (1 - \alpha) \nabla_j f_r(x^{(k)}) \leq -\frac{1}{10} < 0 \). From Lemma 3.1:

\[
f(x^{(k+1)}) - f(x^{(k)}) \leq -\frac{\beta(1 - \alpha)}{8(1 + \alpha \beta)} \sum_{j \in S_-} x_j^{(k)} \nabla_j f(x^{(k)}) \nabla_j f_r(x^{(k)}) \leq -\frac{\beta}{80(1 + \alpha \beta)} \sum_{j \in S_-} x_j^{(k)} \left( 1 - (x_j^{(k)})^{-\alpha} (A_T^T y^{(k)})_j \right) \leq -\frac{\beta}{800(1 + \alpha \beta)} \langle 1, x^{(k)} \rangle \text{.}
\]

Similar as in the previous case, assume that \( \langle y^{(k)}, A(x^{(k)})^\frac{1}{1 - \alpha} \rangle \geq 2 \langle 1, x^{(k)} \rangle \). Then \( \langle 1, x^{(k)} \rangle \geq \frac{\alpha - 1}{2} f_r(x^{(k)}) \), and we have \( f_r(x^{(k+1)}) - f_r(x^{(k)}) \leq -\theta(\beta(\alpha - 1)) f_r(x^{(k)}) \).

**Case 3:** \( \langle y^{(k)}, A(x^{(k)})^\frac{1}{1 - \alpha} \rangle \geq 2 \langle 1, x^{(k)} \rangle \). Equivalently: \( \frac{1}{2} \langle y^{(k)}, A(x^{(k)})^\frac{1}{1 - \alpha} \rangle \geq \langle 1, x^{(k)} \rangle \). Subtracting \( \langle y^{(k)}, A(x^{(k)})^\frac{1}{1 - \alpha} \rangle \) from both sides and rearranging the terms:

\[
\sum_{j=1}^{n} x_j^{(k)} \left( 1 + (x_j^{(k)})^{-\alpha} (A_T^T y^{(k)})_j \right) \geq 2 \langle 1, x^{(k)} \rangle \text{.}
\]

Let \( \zeta_j = \left| -1 + (x_j^{(k)})^{-\alpha} (A_T^T y^{(k)})_j \right| \). Then:

\[
\sum_{j=1}^{n} x_j^{(k)} \left( 1 + (x_j^{(k)})^{-\alpha} (A_T^T y^{(k)})_j \right) \leq \frac{1}{2} \langle 1, x^{(k)} \rangle + \sum_{j: \zeta_j > 1/2} x_j^{(k)} \zeta_j \leq \frac{1}{4} \langle y^{(k)}, A(x^{(k)})^\frac{1}{1 - \alpha} \rangle + \sum_{j: \zeta_j > 1/2} x_j^{(k)} \zeta_j. \quad (C.22)
\]

As \( f_r(x^{(k)}) \leq \left( \frac{1}{2(\alpha - 1)} + \frac{\beta}{1 + \beta} \right) \langle y^{(k)}, A(x^{(k)})^\frac{1}{1 - \alpha} \rangle \), combining (C.21) and (C.22):

\[
\sum_{j: \zeta_j > 1/2} x_j^{(k)} \zeta_j \geq \frac{1}{4} \langle y^{(k)}, A(x^{(k)})^\frac{1}{1 - \alpha} \rangle \geq \frac{1}{4} \left( \frac{1}{2(\alpha - 1)} + \frac{\beta}{1 + \beta} \right)^{-1} f_r(x^{(k)}). \quad (C.23)
\]

Hence, applying Lemma 3.1, it follows that \( f(x^{(k+1)}) - f(x^{(k)}) \leq -\frac{\beta}{16(1 + \alpha \beta)} \sum_{j: \zeta_j > 1/2} x_j^{(k)} \zeta_j \), which, combined with (C.23), gives: \( f(x^{(k+1)}) \leq (1 - \theta(\beta(\alpha - 1))) f(x^{(k)}) \), as claimed. \( \square \)

### C.3.2 The Decrease in the Duality Gap and the Convergence Bound

Based on the results from the previous subsection, within the first \( O(\frac{1}{\beta + \frac{1}{\beta}} \max \{ \frac{1}{\alpha - 1}, \alpha - 1 \} \log \left( \frac{f_r(x^{(0)})}{f_r(x^{(k)})} \right) \) iterations, there must exist at least one iteration in which the conditions from Proposition C.9 and Lemma C.10 hold. With the (slight) abuse of notation, we will treat first such iteration as our initial \( (k = 0) \) iteration, and
focus on proving the convergence over a subsequence of the subsequent iterations. We will call the iterations over which we will perform the gap analysis the “gap iterations” and we define them as iterations in which:

\[ \langle y^{(k)}, A(x^{(k)}) \rangle \leq 2 \langle 1, x^{(k)} \rangle . \]  

(C.24)

Due to Lemma C.11, in non-gap iterations, \( f_r(x^{(k)}) \) must decrease multiplicatively. Hence, we focus only on the gap iterations, which we index by \( k \) below.

To construct the approximate duality gap, we define the upper bound to be \( U_k = f_r(x^{(k+1)}) \). The lower bound is simply defined through the use of the Lagrangian dual as: \( L_k = \frac{\sum_{i=0}^{k} h_i g(y^{(i)})}{H_k} \).

**Initial gap.** Due to Lemma C.10 and the choice of the initial point \( k = 0 \) described above, we have:

\[ G_0 = U_0 - L_0 \leq -2f_\alpha(x^*_\alpha), \]  

(C.25)

as, using Lemma 3.1, \( U_0 = f_r(x^{(1)}) \leq f_r(x^{(0)}) \).

**The gap decrease.** The next step is to show that, for a suitably chosen sequence \( \{h_k\}_k \), \( H_k G_k - H_{k-1} G_{k-1} \leq O(\epsilon)(1 - \alpha) f_\alpha(x^*_\alpha) \). This would immediately imply \( G_k \leq \frac{H_k G_0}{H_k} + O(\epsilon)(1 - \alpha) f_\alpha(x^*_\alpha) \) which is \( O(\epsilon)(1 - \alpha) f_\alpha(x^*_\alpha) \) when \( H_k / H_{k-1} = O(\epsilon(\alpha - 1)) \), due to the bound on the initial gap (C.25). As \( U_k = f_r(x^{(k+1)}) \geq \frac{\langle f_r(x^{(k+1)}) \rangle}{\alpha} \) and \( L_k \geq -f_\alpha(x^*_\alpha) \), taking \( \hat{x}^{(k)} = (x^{(k+1)})^{\frac{1}{1 - \alpha}} \), it would immediately follow that:

\[ -f_\alpha(\hat{x}^{(k)}) + f_\alpha(x^*_\alpha) \leq O(\epsilon)(1 - \alpha) f_\alpha(x^*_\alpha). \]

Since \( \hat{x}^{(k)} \) is \((\alpha P)\)-feasible (due to Proposition 3.2), \( \hat{x}^{(k)} \) is an \((\epsilon)\)-approximate solution to \((\alpha P)\).

To bound \( H_k G_k - H_{k-1} G_{k-1} \), we will need the following technical proposition that bounds \( H_k L_k - H_{k-1} L_{k-1} \) (the change in the lower bound).

**Proposition C.12.** For any two consecutive gap iterations \( k - 1, k \):

\[ H_k L_k - H_{k-1} L_{k-1} \geq h_k \left[ f_r(x^{(k)}) - \langle \nabla f_r(x^{(k)}), x^{(k)} \rangle - 8\epsilon(\alpha - 1) f_\alpha(x^*_\alpha) + \frac{\alpha}{\alpha - 1} \sum_{j=1}^{n} x_j^{(k)} \left( (1 + \nabla_j f_r(x^{(k)}))^{\frac{\alpha - 1}{\alpha}} - (1 + (1 - \alpha) \nabla_j f_r(x^{(k)})) \right) \right]. \]

**Proof.** By the definition of the lower bound:

\[ H_k L_k - H_{k-1} L_{k-1} = h_k g(y^{(k)}) = h_k \left( -\langle 1, y^{(k)} \rangle \right) + \frac{\alpha}{\alpha - 1} \sum_{j=1}^{n} (A^T y^{(k)})^{\frac{\alpha - 1}{\alpha}}. \]  

(C.26)

From Proposition C.9: \( \langle 1, y^{(k)} \rangle \leq (1 + \epsilon) \left( A^T y^{(k)}, (x^{(k)})^{\frac{1}{1 - \alpha}} \right) \), while from Equation (C.16) \( f_r(x^{(k)}) - \langle \nabla f_r(x^{(k)}), x^{(k)} \rangle = (\frac{\beta}{1 + \beta} + \frac{1}{\alpha - 1}) \left( A^T y^{(k)}, (x^{(k)})^{\frac{1}{1 - \alpha}} \right) \). Hence, we can write:

\[ \langle 1, y^{(k)} \rangle \leq (1 + \epsilon) \left( A^T y^{(k)}, (x^{(k)})^{\frac{1}{1 - \alpha}} \right) = -f_r(x^{(k)}) + \langle \nabla f_r(x^{(k)}), x^{(k)} \rangle + \frac{\alpha}{\alpha - 1} \left( A^T y^{(k)}, (x^{(k)})^{\frac{1}{1 - \alpha}} \right) + \left( \epsilon + \frac{\beta}{1 + \beta} \right) \left( A^T y^{(k)}, (x^{(k)})^{\frac{1}{1 - \alpha}} \right). \]
Since $k$ is a gap iteration, $f_r(x^{(k)}) \geq \left( \frac{1}{2(\alpha - 1)} + \frac{\beta}{1 + \beta} \right) \mathcal{A}^T y^{(k)}, (x^{(k)}) \left( \frac{1}{\alpha} - \alpha \right)$. Hence, it follows that:

\[
\begin{align*}
\langle \mathbf{1}, y^{(k)} \rangle & \leq -f_r(x^{(k)}) + \langle \nabla f_r(x^{(k)}), x^{(k)} \rangle + \frac{\alpha}{\alpha - 1} \langle \mathcal{A}^T y^{(k)}, (x^{(k)}) \mathcal{A}^T y^{(k)} \rangle + 10 \frac{\epsilon f_r(x^{(k)})}{4(\alpha - 1)} - 8(\alpha - 1)\epsilon f_\alpha(x^{\alpha}) \\
& \leq -f_r(x^{(k)}) + \langle \nabla f_r(x^{(k)}), x^{(k)} \rangle + \frac{\alpha}{\alpha - 1} \langle \mathcal{A}^T y^{(k)}, (x^{(k)}) \mathcal{A}^T y^{(k)} \rangle - 8(\alpha - 1)\epsilon f_\alpha(x^{\alpha})
\end{align*}
\]

(C.27)

where the last inequality follows from $f_r(x^{(k)}) \leq f_r(x^{(0)})$ (as $f_r(\cdot)$ decreases in each iteration) and $f_\alpha(x^{(0)}) \leq -\frac{11}{4} f_\alpha(x^{\alpha})$ (by the choice of $x^{(0)}$ and Lemma C.10). Combining (C.26) and (C.27):

\[
\begin{align*}
H_k L_k - H_{k-1} L_{k-1} & \geq h_k \left[ -f_r(x^{(k)}) - \langle \nabla f_r(x^{(k)}), x^{(k)} \rangle + \frac{\alpha}{\alpha - 1} \sum_{j=1}^{\frac{n}{\alpha}} \left( (\mathcal{A}^T y^{(k)}) r_j (x^{(k)}) - (x_j^{(k)}) \right) \right] + h_k 8\epsilon (\alpha - 1)\alpha f_\alpha(x^{\alpha}).
\end{align*}
\]

Finally, as $(\mathcal{A}^T y^{(k)})_j = (x_j^{(k)}) \mathcal{A}^T (1 + (1 - \alpha)\nabla_j f_r(x^{(k)}))$ and $(1 - \alpha)\nabla_j f_r(x^{(k)}) \geq \nabla_j f_r(x^{(k)})$, we have:

\[
\begin{align*}
H_k L_k - H_{k-1} L_{k-1} & \geq h_k \left[ -f_r(x^{(k)}) - \langle \nabla f_r(x^{(k)}), x^{(k)} \rangle + 8\epsilon (\alpha - 1)\alpha f_\alpha(x^{\alpha}) \\
& \quad + \frac{\alpha}{\alpha - 1} \sum_{j=1}^{\frac{n}{\alpha}} x_j^{(k)} \left( (1 + \nabla_j f_r(x^{(k)})) \right) - (1 + (1 - \alpha)\nabla_j f_r(x^{(k)})) \right]
\end{align*}
\]

as claimed.

\[\square\]

**Lemma C.13.** If, for $k \geq 1$, $\frac{h_k}{H_k} \leq \frac{\beta \min\{\alpha - 1, 1\}}{10(1 + \alpha \beta)}$, then $H_k G_k - H_{k-1} G_{k-1} \leq -8h_k \epsilon (\alpha - 1)\alpha f_\alpha(x^{\alpha})$.

**Proof.** Using Lemma 3.1 (and as $f_r(x^{(k)})$ decreases by the Lemma 3.1 guarantees regardless of whether the iteration is a gap iteration or not):

\[
\begin{align*}
H_k U_k - H_{k-1} U_{k-1} & \leq h_k f_r(x^{(k)}) - H_k \frac{\beta(1 - \alpha)}{8(1 + \alpha \beta)} \sum_{j=1}^{\frac{n}{\alpha}} x_j^{(k)} \nabla_j f_r(x^{(k)}) \nabla_j f_r(x^{(k)}).
\end{align*}
\]

Combining with the change in the lower bound from Proposition C.12, it follows that to prove the statement of the lemma it suffices to show that, $\forall j$:

\[
\begin{align*}
\xi_j & \overset{\text{def}}{=} h_k \left[ \nabla_j f_r(x^{(k)}) - \frac{\alpha}{\alpha - 1} \left( (1 + \nabla_j f_r(x^{(k)})) - (1 + (1 - \alpha)\nabla_j f_r(x^{(k)})) \right) \right] \\
& \quad - H_k \frac{\beta(1 - \alpha)}{8(1 + \alpha \beta)} \nabla_j f_r(x^{(k)}) \nabla_j f_r(x^{(k)}) \\
& \leq 0.
\end{align*}
\]

Consider the following three cases:

**Case 1:** $(1 - \alpha)\nabla_j f_r(x^{(k)}) \in [-1/2, 1]$. Then $\nabla_j f_r(x^{(k)}) = (1 - \alpha)\nabla_j f_r(x^{(k)})$. A simple corollary of Taylor’s Theorem is that in this setting:

\[
(1 + \nabla_j f_r(x^{(k)})) \geq 1 + \frac{\alpha - 1}{\alpha} \nabla_j f_r(x^{(k)}) - \frac{\alpha - 1}{\alpha^2} (\nabla_j f_r(x^{(k)}))^2.
\]  

(C.28)
Using Eq. (C.28) from above:

\[
\xi_j \leq h_k \left[ \frac{\nabla_j f_r(x^{(k)})}{1-\alpha} - \frac{\alpha}{\alpha-1} \left( -\frac{1}{\alpha} \nabla_j f_r(x^{(k)}) - \frac{\alpha-1}{\alpha^2} (\nabla_j f_r(x^{(k)}))^2 \right) \right] - \frac{H_k \beta}{8(1+\alpha \beta)} (\nabla_j f_r(x^{(k)}))^2
\]

\[
= (\nabla_j f_r(x^{(k)}))^2 \left( h_k - \frac{H_k \beta}{8(1+\alpha \beta)} \left( \frac{\alpha-1}{\alpha} \right) \right).
\]

As \( \frac{h_k}{H_k} \leq \frac{\beta}{8(1+\alpha \beta)} \), it follows that \( \xi_j \leq 0 \).

**Case 2:** \((1-\alpha)\nabla_j f_r(x^{(k)}) \in [-1, -1/2] \). Then \( \nabla_j f_r(x^{(k)}) = (1-\alpha)\nabla_j f_r(x^{(k)}) \) and \(|\nabla_j f_r(x^{(k)})| > \frac{1}{2} \).

As in this case \((1+\nabla_j f_r(x^{(k)}))^\frac{\alpha-1}{\alpha} \geq 1 + \nabla_j f_r(x^{(k)}) \), we have:

\[
\xi_j \leq \nabla_j f_r(x^{(k)}) \left( h_k - \frac{H_k \beta}{8(1+\alpha \beta)} (1-\alpha) \right),
\]

which is \( \leq 0 \), as \( \frac{h_k}{H_k} \leq \frac{\beta}{8(1+\alpha \beta)} \) and \( \nabla_j f_r(x^{(k)}) > 0 \) (because \( \alpha > 1 \) and \((1-\alpha)\nabla_j f_r(x^{(k)}) < 0 \)).

**Case 3:** \((1-\alpha)\nabla_j f_r(x^{(k)}) > 1 \). Then \( \nabla_j f_r(x^{(k)}) = 1 \), and we have:

\[
\xi_j \leq h_k \left[ \nabla_j f_r(x^{(k)}) - \frac{\alpha}{\alpha-1} \left( 2^{\frac{\alpha-1}{\alpha}} - 1 - (1-\alpha)\nabla_j f_r(x^{(k)}) \right) \right] - \frac{H_k \beta}{8(1+\alpha \beta)} \nabla_j f_r(x^{(k)})
\]

\[
\leq (1-\alpha)\nabla_j f_r(x^{(k)}) \left( h_k - \frac{H_k \beta}{8(1+\alpha \beta)} \right),
\]

which is non-positive, as \( \frac{h_k}{H_k} \leq \frac{\beta}{8(1+\alpha \beta)} \).

We can now state the final convergence bound.

**Theorem C.14.** Given \( \epsilon \in (0, \min\{1/2, 1/(10(\alpha-1))\}] \), after at most

\[
O\left( \frac{\alpha^3 \log(n \rho) \log(mn \rho/\epsilon)}{\epsilon}, \frac{\log\left( \frac{1}{\epsilon(\alpha-1)} \right) \log(mn \rho/\epsilon)}{\epsilon(\alpha-1)} \right)
\]

iterations of FAIRPACKING, \( f_\alpha(x^{(k+1)}_\alpha) - f_\alpha(x^*_\alpha) \geq 10 \epsilon(\alpha-1) f_\alpha(x^*_\alpha) \), where \( x^{(k+1)}_\alpha = (x^{(k+1)})^\frac{1}{\alpha} \).

**Proof.** At initialization, \( f_r(\cdot) \) takes value less than \( \frac{n(3n \rho)^{\alpha-1}}{\alpha-1} \) and decreases in every subsequent iteration.

From Proposition 2.3, \(-f_\alpha(x^*_\alpha) \geq \frac{n}{\alpha-1} \). As \( f_r(x) \geq \frac{\alpha}{\alpha-1} \) and the algorithm always maintains solutions \( x^{(k)} \) that are feasible in \((\alpha P_c)\), \( \min_k f_r(x^{(k)}) \geq -f_\alpha(x^*_\alpha) \geq \frac{n}{\alpha-1} \). Using Proposition C.9 and Lemma C.11, there can be at most \( O\left( \frac{1}{\beta} \max\{\frac{1}{\alpha-1}, \alpha-1\} \right) \alpha(\alpha-1) \log(n \rho) \right) \) non-gap iterations before \( f_r(\cdot) \) reaches its minimum value. Using the second part of Proposition 2.4, if this happens, it follows that \( f_\alpha(x^{(k+1)}_\alpha) - f_\alpha(x^*_\alpha) \geq -2(1-\alpha) f_\alpha(x^*_\alpha) \), and we are done.

For the gap iterations, choose \( h_0 = H_0 = 1, \frac{h_\ell}{H_\ell} = (1 - \frac{H_{\ell-1}}{H_\ell}) = \frac{\beta \min\{\alpha-1, 1\}}{16(1+\alpha \beta)} \), for \( \ell \geq 1 \). Using Lemma C.13:

\[
G_k \leq \frac{H_0 G_0}{H_k} - 8 \epsilon(\alpha-1) f_\alpha(x^*_\alpha)
\]

\[
= \frac{H_0}{H_1} \cdot \frac{H_1}{H_2} \cdots \frac{H_{k-1}}{H_k} G_0 - 8 \epsilon(\alpha-1) f_\alpha(x^*_\alpha)
\]

\[
= \left( 1 - \frac{\beta \min\{\alpha-1, 1\}}{16(1+\alpha \beta)} \right)^{k} G_0 - 8 \epsilon(\alpha-1) f_\alpha(x^*_\alpha).
\]
As \( G_0 \leq -2f_\alpha(x_0^*) \), after \( k \geq \frac{\log(\frac{1}{\epsilon(\alpha-1)})}{\beta \min(\alpha-1,1)} \), \( 16(1 + \alpha \beta) = O\left( \frac{(1 + \alpha) \log(\frac{1}{\epsilon(\alpha-1)}) \log(mnp/\epsilon)}{\epsilon \min(\alpha-1,1)} \right) \) iterations, it must be \(-f_\alpha(x^{(k+1)}) + f_\alpha(x_0^*) \leq f_r(x^{(k+1)}) + f_\alpha(x_0^*) \leq G_k \leq 10\epsilon(1 - \alpha) f_\alpha(x_0^*)\), as claimed. \( \square \)

## D Convergence Analysis for Fair Covering

Since we do not need to ensure the feasibility of the packing problem, in this section we take \( C = 1 \), so that \( f_r(x) = -\langle 1, x \rangle + \frac{\beta}{1+\beta} \sum_{i=1}^m (Ax)_i \). As before, the upper bound is defined as \( U_k = f_r(x^{(k+1)}) \). The lower bound \( L_k \) is the same as the one from Section C.1, with the choice of \( \beta' \) as in FAIRCOVERING.

**Proposition D.1.** Let \( h_0 = H_0 = 1 \). Then: \( H_k G_0 \leq 2(1 + \beta)g_\beta(y_\beta^*) \).

**Proof.** By the same arguments as in the proof of Proposition C.1, \( x^{(1)} = x^{(0)} \), and hence \( U_0 = f_r(x^{(0)}) \).

Let \( x_\beta^* = \arg\min_{x \geq 0} f_r(x) \). Then, the initial gap can be expressed as:

\[
H_0 G_0 = \left\langle \nabla f_r(x^{(0)}), x^{(0)} \right\rangle - \bar{\psi}(x^{(0)}) + \phi(x_\beta^*)
\]

By the choice of \( x^{(0)} \), \( (Ax^{(0)})^{1/\beta} \leq 1 \), and, therefore, \( \nabla f_r(x^{(0)}) \leq 0 \). Thus, \( H_0 G_0 \leq -\bar{\psi}(x^{(0)}) + \phi(x_\beta^*) \).

As \( \beta' \) chosen here is smaller than the one from Section C.1, it follows by the same argument as in the proof of Proposition C.2 that \(-\bar{\psi}(x^{(0)}) = \frac{\beta'}{\epsilon(1 - \beta')}(1, x^{(0)})^{1 - \beta'} \leq \frac{1}{2} \langle 1, x^{(0)} \rangle \), which is at most \( \frac{1}{2}(1 + \beta)g_\beta(y_\beta^*) \), by the choice of the initial point \( x^{(0)} \) and Proposition 2.5. It remains to bound \( \phi(x_\beta^*) = \psi(x_\beta^*) - \left\langle \nabla \psi(x^{(0)}), x_\beta^* \right\rangle \). By the definition of \( \psi \), \( \psi(x_\beta^*) \leq 0 \) and \( \nabla_\beta \psi(x^{(0)}) = \frac{1}{\epsilon}(1 - (x_2^{(0)})^{-\beta'}) \leq -1/2 \). By the definition of \( f_r \), \( \nabla_\beta f_r(x^{(0)}) \geq -1 \). Hence:

\[
-\left\langle \nabla \psi(x^{(0)}), x_\beta^* \right\rangle \leq \frac{3}{2}(1 + \beta)g_\beta(y_\beta^*)
\]

where the last inequality is by Proposition 2.6. \( \square \)

Since the analysis from Section C.1 can be applied in a straightforward way to ensure that after \( \lceil 2/(\epsilon h) \rceil \) iterations we have \( H_k G_k \leq \epsilon(1 + \beta)g_\beta(y_\beta^*) \), what remains to show is that we can recover an approximate solution to \((\beta C)\) from this analysis. Define:

\[
y^{(k)} = (Ax^{(k)})^{1/\beta} \quad \text{and} \quad y_\beta^{(k)} = \frac{\sum_{\ell=1}^k y^{(\ell)}}{k}.
\]

Notice that this is consistent with the definition of \( y_\beta^{(k)} \) from FAIRCOVERING. We are now ready to prove Theorem 4.1.

**Proof of Theorem 4.1.** Observe that, by the definition of \( f_r \) and \( y^{(k)} \), we have \( f_r(x^{(k)}) - \left\langle \nabla f_r(x^{(k)}), x^{(k)} \right\rangle = -\sum_{i=1}^m (y_i^{(k)})^{1+\beta} \). Hence, \( L_k = -\sum_{\ell=0}^k h_\ell g_\beta(y^{(\ell)}) + \bar{\psi}(y^{(k)}) - \phi(x_\beta^*) \).

As, by Proposition D.1 and the analysis from Section C.1 it must be \( G_K \leq 2\epsilon(1 + \beta)g_\beta(y_\beta^*) \), and by Lagrangian duality \( U_K = f_r(x^{(K+1)}) \geq -g_\beta(y_\beta^*) \), we have that \( L_K = U_K - G_K \geq -(1 + 2\epsilon(1 + \beta)g_\beta(y_\beta^*) \).
\( \beta \))g_\beta(y^*_\beta). \) As \( \hat{\psi}(z^{(k)}) \leq 0 \) and \( \phi(x^*_\beta) \geq \psi(x^*_\beta) \geq -\frac{1}{2}\left(1 + \beta\right)g_\beta(y^*_\beta) \) (because \( \nabla \psi(x^{(0)}) + \nabla f_j(x^{(0)}) \geq 0 \) and, by the choice of \( \beta' \), \( \psi(x^*_\beta) \geq -\frac{1}{2}\left(1 + \beta\right)g_\beta(y^*_\beta) \)), we have that:

\[
- \sum_{\ell=0}^K h_\ell g_\beta(y^{(\ell)}) \leq -(1 + 2\epsilon(1 + \beta))g_\beta(y^*_\beta) - \frac{(1 + \beta)g_\beta(y^*_\beta)}{2H_K} \geq -(1 + (9\epsilon/4)(1 + \beta))g_\beta(y^*_\beta). \quad (D.2)
\]

Recall that \( h_0 = 1, h_\ell = h \) for \( \ell \geq 1 \) and \( H_K = \sum_{\ell=0}^K h_\ell = 1 + Kh. \) As \( g_\beta \) is convex, by the definition of \( y^{(K)}_\beta \) and Jensen’s Inequality:

\[
\sum_{\ell=0}^K h_\ell g_\beta(y^{(\ell)}) = \frac{1}{H_0}g_\beta(y^{(0)}) + \frac{h}{1 + hK} \sum_{\ell=1}^K g_\beta(y^{(\ell)}) \geq \frac{hK}{1 + hK}g_\beta(y^{(K)}) \geq \frac{1}{1 + \epsilon/2}g_\beta(y^{(K)}). \quad (D.3)
\]

Hence, combining (D.2) and (D.3), \( g_\beta(y^{(K)}_\beta) \leq (1 + 3\epsilon(1 + \beta))g_\beta(y^*_\beta). \)

It remains to show that \( y^{(K)}_\beta \) is nearly-feasible. By the definition of \( y^{(K)}_\beta \), \( y^{(K)}_\beta \geq 0. \) We claim first that it must be \( z^{(k)} \geq -(\epsilon/2)\mathbb{I}. \) Suppose not. Then \( \hat{\psi}(z^{(k)}) = -\frac{\beta'}{\epsilon(1 - \beta')} \sum_{j=1}^n z_j^{(k)}(1 - 2\epsilon/\beta') \leq -\frac{\beta'}{\epsilon(1 - \beta')} \leq - \frac{\beta'}{\epsilon(1 - \beta')} \leq -H_K(1 + \beta)g_\beta(y^*_\beta) \). As (from the argument above) \( \phi(x^*_\beta) \geq \psi(x^*_\beta) \geq -\frac{1}{2}(1 + \beta)g_\beta(y^*_\beta) \), it follows that \( L_K \leq -(1 + \beta)g_\beta(y^*_\beta) \), which is a contradiction, as we have already shown that \( L_K \geq -(1 + \epsilon(1 + \beta))g_\beta(y^*_\beta) \). Thus, we have, \( \forall j, z_j^{(k)} \geq -\epsilon/2. \) Recall from the definition of \( z^{(k)} \) that:

\[
1 + \sum_{\ell=1}^K h_\ell \nabla_j f_j(x^{(\ell)}) = 1 + \epsilon \left( \sum_{\ell=1}^K h_\ell \nabla_j f_j(x^{(\ell)}) \right)
\]

\[
\leq 1 + \epsilon \sum_{\ell=1}^K h_\ell \nabla_j f_j(x^{(\ell)})
\]

\[
\leq 1 + \epsilon \sum_{\ell=1}^K h_\ell \nabla_j f_j(x^{(\ell)})
\]

Recall that \( \nabla_j f_j(x^{(\ell)}) = -1 + (A^T y^{(\ell)})_j. \) Hence

\[
A^T y^{(K)}_\beta = \frac{\sum_{\ell=1}^K A^T y^{(\ell)}}{K} \geq z^{(k)} + \epsilon hK \geq (1 - \epsilon/2)\mathbb{I},
\]

as claimed. \( \square \)

References

[1] Z. Allen-Zhu and L. Orecchia. Using optimization to break the epsilon barrier: A faster and simpler width-independent algorithm for solving positive linear programs in parallel. In Proc. ACM-SIAM SODA’15, 2015.

[2] A. B. Atkinson. On the measurement of inequality. J. Econ. Theory, 2(3):244–263, 1970.

[3] B. Awerbuch and R. Khandekar. Stateless distributed gradient descent for positive linear programs. SIAM J. Comput., 38(6):2468–2486, 2009.
[4] Y. Azar, N. Buchbinder, T. H. Chan, S. Chen, I. R. Cohen, A. Gupta, Z. Huang, N. Kang, V. Nagarajan, J. Naor, et al. Online algorithms for covering and packing problems with convex objectives. In Proc. IEEE FOCS, 2016.

[5] Y. Bartal, J. Byers, and D. Raz. Global optimization using local information with applications to flow control. In Proc. IEEE FOCS’97, 1997.

[6] J. Bentham. The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation. Clarendon Press, 1996.

[7] D. P. Bertsekas. Convex optimization theory. Athena Scientific Belmont, 2009.

[8] D. Bertsimas, V. F. Farias, and N. Trichakis. The price of fairness. Operations research, 59(1):17–31, 2011.

[9] D. Bertsimas, V. F. Farias, and N. Trichakis. On the efficiency-fairness trade-off. Manag. Sci, 58(12):2234–2250, 2012.

[10] R. Binns. Fairness in machine learning: Lessons from political philosophy. Proceedings of Machine Learning Research, pages 1–11. Forthcoming. Available at SSRN: https://ssrn.com/abstract=3086546.

[11] T. Bonald and J. Roberts. Multi-resource fairness: Objectives, algorithms and performance. In Proc. ACM SIGMETRICS’15, 2015.

[12] S. Borst, N. Walton, and B. Zwart. Network iso-elasticity and weighted α-fairness. Performance Evaluation, 70(11):995–1000, 2013.

[13] N. Buchbinder and S. Naor. Online primal-dual algorithms for covering and packing. Math. of OR, 34(2):270286, 2009.

[14] N. Devanur, C. Papadimitriou, A. Saberi, and V. Vazirani. Market equilibrium via a primal–dual algorithm for a convex program. Journal of the ACM (JACM), 55(5):22, 2008.

[15] J. Diakonikolas and L. Orecchia. The approximate duality gap technique: A unified theory of first-order methods, 2017. arXiv preprint, arXiv:1712.02485.

[16] J. Diakonikolas and L. Orecchia. Solving packing and covering linear programs in $\tilde{O}(\epsilon^{-2})$ distributed iterations with a single algorithm and simpler analysis, 2017. arXiv preprint, arXiv:1710.09002.

[17] R. Eghbali and M. Fazel. Designing smoothing functions for improved worst-case competitive ratio in online optimization. In Proc. NIPS, 2016.

[18] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica. Dominant resource fairness: Fair allocation of multiple resource types. In Proc. USENIX NSDI’11, 2011.

[19] K. Jain and V. Vazirani. Eisenberg-gale markets: Algorithms and structural properties. In Proc. ACM STOC’07, 2007.

[20] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang. Multiresource allocation: Fairness-efficiency tradeoffs in a unifying framework. IEEE/ACM Trans. Netw., 21(6):1785–1798, 2013.

[21] E. Kalai and M. Smorodinsky. Other solutions to nash’s bargaining problem. Econometrica: Journal of the Econometric Society, pages 513–518, 1975.
[22] F. Kelly and E. Yudovina. *Stochastic networks*, volume 2. Cambridge University Press, 2014.

[23] J. Kleinberg, Y. Rabani, and É. Tardos. Fairness in routing and load balancing. In *Proc. IEEE FOCS’99*, 1999.

[24] F. Kuhn, T. Moscibroda, and R. Wattenhofer. The price of being near-sighted. In *Proc. ACM-SIAM SODA’06*, 2006.

[25] T. Lan, D. Kao, M. Chiang, and A. Sabharwal. An axiomatic theory of fairness in network resource allocation. In *Proc. IEEE INFOCOM’10*, 2010.

[26] S. Low, F. Paganini, and J. Doyle. Internet congestion control. *IEEE Control Systems*, 22(1):28–43, 2002.

[27] M. Luby and N. Nisan. A parallel approximation algorithm for positive linear programming. In *Proc. ACM STOC’93*, 1993.

[28] M. W. Mahoney, S. Rao, D. Wang, and P. Zhang. Approximating the solution to mixed packing and covering LPs in parallel $O(\epsilon^{-3})$ time. In *Proc. ICALP’16*.

[29] J. Marasevic, C. Stein, and G. Zussman. A fast distributed stateless algorithm for alpha-fair packing problems. In *Proc. ICALP’16*, 2016.

[30] B. McCormick, F. Kelly, P. Plante, P. Gunning, and P. Ashwood-Smith. Real time alpha-fairness based traffic engineering. In *Proc. ACM HotSDN’14*, 2014.

[31] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.*, 8(5):556–567, Oct. 2000.

[32] J. F. Nash Jr. The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162, 1950.

[33] L. Orecchia, S. Sachdeva, and N. K. Vishnoi. Approximating the exponential, the Lanczos method and an $\tilde{O}(m)$-time spectral algorithm for balanced separator. In *Proc. ACM STOC’12*, 2012.

[34] C. Papadimitriou and M. Yannakakis. Linear programming without the matrix. In *Proc. ACM STOC’93*, 1993.

[35] J. Rawls. A theory of justice. *Mass.: Harvard University*, 1971.

[36] Y. Yi and M. Chiang. Stochastic network utility maximisation a tribute to Kelly’s paper published in this journal a decade ago. *Eur. Trans. Telecommun.*, 19(4):421–442, 2008.

[37] N. Young. Sequential and parallel algorithms for mixed packing and covering. In *Proc. IEEE FOCS’01*, 2001.