SIAM J. OPTIM.
Vol. 30, No. 4, pp. 3284–3314

FAIR PACKING AND COVERING ON A RELATIVE SCALE*

JELENA DIAKONIKOLAS†, MARYAM FAZEL‡, AND LORENZO ORECCHIA§

Abstract. Fair resource allocation is a fundamental optimization problem with applications in operations research, networking, and economic and game theory. Research in these areas has led to the general acceptance of a class of α-fair utility functions parameterized by $\alpha \in [0, \infty]$. We consider α-fair packing—the problem of maximizing α-fair utilities under positive linear constraints—and provide a simple first-order method for solving it with relative-error guarantees. The method has a significantly lower convergence time than the state of the art, and to analyze it, we leverage the approximate duality gap technique, which provides an intuitive interpretation of the convergence argument. Finally, we introduce a natural counterpart of α-fairness for minimization problems and motivate its usage in the context of fair task allocation. This generalization yields α-fair covering problems, for which we provide the first near-linear-time solvers with relative-error guarantees.

Key words. resource allocation, fairness, width-independent algorithms, relative error

AMS subject classifications. 90C06, 90C25, 49N15, 65K05

DOI. 10.1137/19M1288516

1. Introduction.

How to split limited resources is a fundamental question studied since antiquity. The study of fairness in economic theory, operations research, and networking led to a single class of utility functions known as α-fair utilities [2, 27]:

$$f_\alpha(x) = \begin{cases} \frac{x^{1-\alpha}}{1-\alpha} & \text{if } \alpha \geq 0,\ \alpha \neq 1, \\ \log(x) & \text{if } \alpha = 1. \end{cases} \tag{1.1}$$

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 α-fair allocation as defined by [27]. This class of problems is well studied in the literature, and axiomatically justified in several works [9, 16, 21]. Notable special cases of α-fair allocations include (i) utilitarian allocations (with linear objectives) for $\alpha = 0$, (ii) proportionally fair allocations that correspond to Nash bargaining solutions [28] for $\alpha = 1$, (iii) TCP-fair objectives that correspond to bandwidth allocations in the internet TCP [19] for $\alpha = 2$, and (iv) max-min fair allocations that correspond to Kalai–Smorodinsky solutions in bargaining theory [18] for $\alpha \to \infty$.

In this paper, we consider the maximization of α-fair utilities subject to positive linear (packing) constraints, to which we refer to as the α-fair packing problems. Given a nonnegative matrix $\mathbf{A} \in {\mathbf{R}_+}^{m \times n}$ and a parameter $\alpha \geq 0$, they are defined

*Received by the editors September 23, 2019; accepted for publication (in revised form) September 28, 2020; published electronically December 10, 2020.
https://doi.org/10.1137/19M1288516
Funding: This work was partially supported by NSF grants CCF-1718342 and CCF-1409836, by NSF TRIPODS Award 1740551, by ONR grant N000141612789, and by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant CCF-1740425.
†Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706 USA (jelena@cs.wisc.edu).
‡Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195 USA (mfazel@uw.edu).
§Department of Computer Science, University of Chicago, Chicago, IL 60637 USA (orecchia @uchicago.edu).

as [25]

$$\text{(P-a)} \qquad \max \left\{ f_\alpha(\mathbf{x}) \stackrel{\text{def}}{=} \sum_{j=1}^{n} f_\alpha(x_j) : \quad \mathbf{A}\mathbf{x} \le \mathbb{1},\ \mathbf{x} \ge \mathbf{0} \right\},$$

where $\mathbf{x} \in \mathbf{R}^n$, $\mathbb{1}$ is an all-ones vector, $\mathbf{0}$ is an all-zeros vector. Packing constraints are natural in many applications, including internet congestion control [22], rate control in software defined networks [26], multiresource allocation in data centers [11, 14, 16], and a variety of applications in operations research, economics, and game theory [10, 15]. When $\alpha = 0$, the problem is equivalent to a packing linear program (LP).

We are interested in solving (P-a) in a *distributed* model of computation, where each coordinate j of the allocation vector $\mathbf{x}$ is updated according to global problem parameters (e.g., m, n), local information for coordinate j (namely, the jth column of $\mathbf{A}$), and local[1] information received in each round. The local per round information for coordinate j is the slack $1-(\mathbf{A}\mathbf{x})_i$ of all the constraints i in which j participates. Such a model is natural for networking applications, where each variable x_j corresponds to the rate assigned to a route j and $1-(\mathbf{A}\mathbf{x})_i$ is the (relative) congestion on each of the links i that belong to the route j (see, e.g., the textbook [19] and references therein). Further, as resource allocation problems frequently arise in large-scale settings in which results must be provided in real time (e.g., in data center resource allocation [11, 14, 16, 17]), the design of distributed solvers that efficiently compute approximate solutions to α-fair allocation problems is of crucial importance.

First-order methods are particularly relevant in this context as they lead to algorithms that can be distributed, have simple updates implementable in large-scale settings, and are efficient in practice. Further, we focus on algorithms that are *width independent*[2] and yield an ϵ-approximate solution in the sense of *relative error*. Width-independent algorithms are of great theoretical interest: algorithms that are not width independent cannot in general be considered polynomial time. Such algorithms have primarily been studied in the context of packing and covering LPs. From an 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*. We also note that methods with relative-error guarantees are much less studied in optimization than their additive-error counterparts, and are typically confined to positive LPs (see, e.g., [30, Chapter 7] and references therein).

Finally, we note that for $\alpha > 0$, α-fair utilities do not possess any of the global regularity properties such as smoothness or Lipschitz continuity that are typically used in convergence analysis of first-order methods. In fact, as any of the coordinates x_j tends to zero, $\nabla_j f_\alpha(x_j) \to \infty$. Notably, it is possible to make the Lipschitz constant of the objective or its gradient finite by enforcing $x_j \ge \delta$ for a sufficiently small δ. However, to ensure that the feasible region contains an ϵ-approximate solution to (P-a), it is necessary that $\delta \le \frac{1}{n \max_{ij} A_{ij}}$, which leads to a prohibitively large Lipschitz constant $\Omega((n \max_{ij} A_{ij})^\alpha)$ of the objective and $\Omega((n \max_{ij} A_{ij})^{\alpha+1})$ of the

[1] There are two reasons why this information should be considered local. The first is that we can, in many situations, view the constraints as distributed agents, in which case the locality of information is clear. The second is that algorithms for fractional packing and covering problems working in the same model as presented here are considered local even in the distributed computing community (see, e.g., [20]). A classical example is fractional matching on graphs, which is widely used as a model for scheduling in wireless networks under interference constraints.

[2] I.e., their iteration complexities scale polylogarithmically with the matrix width, defined as the maximum ratio of $\mathbf{A}$'s nonzero entries.

gradient, and thus algorithms that are not width independent. In particular, if one was to use Nesterov's "smooth minimization of nonsmooth functions" [29] to deal with the constraints, the resulting number of iterations required to construct a solution with the same approximation guarantee as in this paper would be no better than $\frac{\rho((n\rho)^{\alpha+1}+\rho)}{\epsilon}$, where $\rho = \frac{\max_{ij} A_{ij}}{\min_{kl:A_{kl}\neq 0} A_{kl}}$ is the matrix width. Alternatively, because α-fair objectives are α-strongly convex, it is possible to work with the dual problem, which is smooth (gradient-Lipschitz) for $\alpha > 0$. This idea was pursued in [6]. However, the smoothness constant is proportional to the maximum eigenvalue of $\mathbf{A}^T\mathbf{A}$ times $1/\alpha$, which is not width independent and, in the worst case, scales as $mn(\max_{ij} A_{ij})^2/\alpha$, leading to the overall polynomial dependence on m, n, ρ and linear dependence on $1/\epsilon$, similarly to the approach described above. These issues are handled in our analysis by using a more fine-grained smoothness-like property of the objective that only holds locally and with sufficiently small step sizes (see Lemma 3.1).

1.1. Contributions. We obtain improved distributed algorithms for constructing ϵ-approximate[3] solutions to α-fair packing problems. As in [25], our specific convergence results depend on the regime of the parameter α, where each iteration takes linear work in the number of nonzero elements of $\mathbf{A}$.

- For $\alpha \in [0, 1)$, Theorem 4.4 shows that a solution with $(1+\epsilon)$-relative error is reached within $O(\frac{\log(n\rho)\log(mn\rho/\epsilon)}{(1-\alpha)^3\epsilon^2})$ iterations. This bound matches the best known results for parallel packing LP solvers for $\alpha = 0$ [1, 24], and improves the dependence on ϵ compared to [25] from ϵ^{-5} to ϵ^{-2} for $\alpha \in (0, 1)$.
- For $\alpha = 1$, Theorem 4.8 yields ϵ-approximate convergence in $O(\frac{\log^3(mn\rho/\epsilon)}{\epsilon^2})$ iterations. In this case only, the error is additive, as the objective can take both positive and negative values. The dependence on ϵ compared to [25] is improved from ϵ^{-5} to ϵ^{-2}.
- For $\alpha > 1$, Theorem 4.14 shows that a solution with $(1-\epsilon)$-relative error[4] is reached within $O(\max\{\frac{\alpha^3\log(n\rho/\epsilon)\log(mn\rho/\epsilon)}{\epsilon}, \frac{\log(\frac{1}{\epsilon(\alpha-1)})\log(mn\rho/\epsilon)}{\epsilon(\alpha-1)}\})$ iterations. This can be extended to the max-min-fair case for sufficiently large α [25]. The dependence on ϵ compared to [25] is improved from ϵ^{-4} to ϵ^{-1}.

While the analysis for each of these cases is somewhat involved, the algorithms we propose are extremely simple, as described in Algorithm 3.1 of section 3. Moreover, our dependence on ϵ is improved by a factor ϵ^{-3} (the dependence on all the remaining parameters is either the same or improved) and the analysis is simpler than the one from [25], as it leverages the approximate duality gap technique (ADGT) [13].

Our final contribution is to introduce a natural counterpart of α-fairness for minimization problems, which we use to study β-fair covering problems, for $\beta \geq 0$.[5]

$$\text{(C)} \qquad \min\left\{ g_\beta(\mathbf{y}) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \frac{{y_i}^{1+\beta}}{1+\beta} : \quad \mathbf{A}^T\mathbf{y} \geq \mathbb{1},\ \mathbf{y} \geq \mathbf{0} \right\}.$$

As for packing problems, the β-covering formulation can be motivated by the goal of producing an equitable allocation of cost among different agents. For instance, we may want to allocate work hours to different workers so that various production requirements, given as covering constraints, are met. In this case, assigning all work

[3] As in [25], the approximation is multiplicative for $\alpha \neq 1$ and additive for $\alpha = 1$.

[4] The relative error in this case is $1-\epsilon$, because for $\alpha > 1$ the objective functions are negative.

[5] We use β instead of α to distinguish between the different parameters in the convergence analysis.

to one worker may provide a solution that minimizes total work but unfairly singles out this worker. However, a fair solution would allocate work so that no worker gets assigned too much work and every worker performs some portion of the work. This is captured in (C) 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 β-fair covering problems, for which we provide the first width-independent ϵ-approximate solver that converges in $O(\frac{(1+\beta)\log(mn\rho)}{\beta\epsilon})$ iterations, by reducing the analysis to the $\alpha < 1$ case of the α-fair packing problems (section 5).

1.2. Our techniques. Several difficulties arise when considering cases $\alpha > 0$ compared to the linear case ($\alpha = 0$). Unlike linear objectives which are 1-Lipschitz, α-fair objectives for $\alpha > 0$ lack any good global properties typically used in convex optimization, e.g., Lipschitz continuity of the function or its gradient. As mentioned before, it is possible to prune the feasible region to guarantee positivity of the vector $\mathbf{x}$. However, any pruning that retains ϵ-approximate solutions would require the point $\frac{1}{n\rho}\mathbb{1}$ to be in the pruned set, leading to Lipschitz constants of 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 variable, 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, 12], 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.

Similarly to the linear case [1], we use regularization of the constraint set to turn the problem into an unconstrained optimization problem over the nonnegative 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. These regularizers seem more natural than entropy, as local smoothness properties used in algorithm analysis hold regardless of whether the point at which local smoothness is considered satisfies the packing constraints, which is not true for entropic regularization. Furthermore, these regularizers are crucial for reducing fair covering problems to α-fair packing problems with $\alpha < 1$ (see section 5).

While the analysis of the case $\alpha \in [0, 1)$ is similar to the analysis of packing LPs from the unpublished note [12] by a subset of the authors, 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 [13] cannot be applied directly, for a number of reasons: (i) it is hard to argue that the optimality gap of 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 approximate dual in ADGT can be a very crude approximation of the true Lagrangian dual for $\alpha > 1$.

Our main idea is to use the Lagrangian dual of the original, nonregularized problem, with two different arguments. The first argument is *local* and relies on similar ideas as [25]: 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 [25], we require much looser regularity conditions, which eventually leads to a much better dependence of the convergence time on ϵ: 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 α. 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 [13], 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 section 4.3.2).

1.3. Additional related work. A long line of work on packing and covering LPs has resulted in width-independent distributed algorithms (see [4, 5, 20, 23, 31, 32] and references therein). This has culminated in recent results that ensure convergence to an ϵ-approximate optimal solution in $O(\log(n)\log(mn/\epsilon)/\epsilon^2)$ rounds of computation [1, 24]. However, when it comes to the general α-fair resource allocation, only [25] provides a width-independent algorithm. The algorithm of [25] works in a very restrictive setting of stateless distributed computation, which leads to convergence times that are polylogarithmic in the problem parameters, but have high dependence on the error parameter ϵ (namely, the dependence is ϵ^{-5} for $\alpha \leq 1$ and ϵ^{-4} for $\alpha > 1$).

The stateless model of distributed computation requires the algorithm (i) to be able to start from an arbitrary (not necessarily feasible) point, (ii) to not have memory of previous states (previous solution points; except for the last one), and (iii) to work in an asynchronous setting, where individual coordinates can be updated based on stale information. While our algorithm satisfies property (ii) and can be adjusted to satisfy property (i), it does not satisfy the third property, which would force the step sizes to be smaller by a factor ϵ and lead to overall slower convergence. Note that better convergence bounds achieved in this work *cannot* be obtained by simply taking a larger step size and following the same analysis as in [25]. This is because the step sizes from [25] are also crucially used in the analysis to guarantee algorithm progress. What leads to faster convergence in our work is the use of ADGT, which allows us to adapt the step sizes to the gradient (step sizes in [25] are not adaptive) and work with the *aggregate* information over *all* iterations, thus constructing better dual solutions and having a global view of the problem. By contrast, the argument from [25] is local, and only uses information from the last iteration to guarantee the algorithm progress or argue that the algorithm has converged to an ϵ-approximate solution.

1.4. Organization of the paper. Section 2 introduces the necessary notation and background. Section 3 provides the statement of the algorithm for α-fair packing and overviews the main technical ideas. The full technical argument is provided in section 4. Section 5 provides the results for β-fair covering. We conclude in section 6.

2. Preliminaries. We assume that the problems are expressed in their standard scaled form [1, 4, 23, 25], so that the minimum nonzero entry of the constraint matrix

A equals one and the maximum element of **A** is equal to the matrix width ρ. Note that weighted versions of the problems, with objective $\sum_j w_j f_\alpha(x_j)$ for positive weights w_j, can be expressed in this form through rescaling and the change of variable. The only effect on the final bounds is that ρ would also depend on $\frac{\max_j w_j}{\min_k w_k}$, which only appears under a logarithm in our bounds, similarly to [25]. The scaling is necessary only for the analysis; the algorithm can be stated for the nonscaled problem by reverting the change of variable (see [1, 3, 25] for similar arguments). The constraint matrix **A** is $m \times n$; **I** denotes the identity matrix.

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 $\mathbf{x}^a$ denote the vector $[{x_1}^a, {x_2}^a, \ldots, {x_n}^a]^T$, $\exp(\mathbf{x})$ denote the vector $[\exp(x_1), \exp(x_2), \ldots, \exp(x_n)]^T$. The inner product of two vectors is denoted as $\langle \cdot, \cdot \rangle$, while the matrix/vector transpose is denoted by $(\cdot)^T$. $\nabla_j f(\cdot)$ denotes the jth coordinate of ∇f, i.e., $\frac{\partial f}{\partial x_j}$. We use the following notation for the truncated (and scaled) gradient [1] for $\alpha \neq 1$:

$$\overline{\nabla_j f}(\mathbf{x}) = \begin{cases} (1-\alpha)\nabla_j f(\mathbf{x}) & \text{if } (1-\alpha)\nabla_j f(\mathbf{x}) \in [-1,1], \\ 1 & \text{otherwise.} \end{cases} \tag{2.1}$$

As we will see later, the only relevant case for us will be the functions whose gradient coordinates satisfy $(1-\alpha)\nabla_j f(\mathbf{x}) \in [-1, \infty)$. Hence, the gradient truncation is irrelevant for $(1-\alpha)\nabla_j f(\mathbf{x}) < -1$. 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 (differentiable) function $f : \mathbb{R}^n_+ \to \mathbb{R}$ is

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+. \tag{2.2}$$

Convex conjugate of a function $\psi : \mathbb{R}^n_+ \to \mathbb{R}$ is defined as

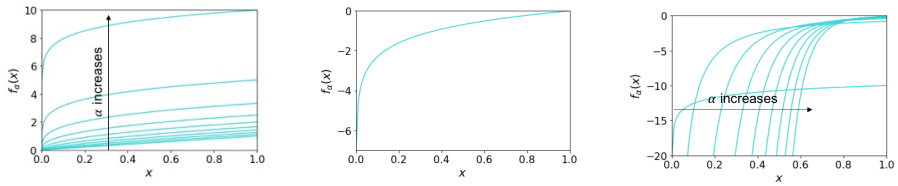
$$\psi^*(\mathbf{z}) \stackrel{\text{def}}{=} \sup_{\mathbf{x} \geq \mathbf{0}} \{\langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{x})\}.$$

In all the examples we consider in this paper, sup can be replaced by max. The following standard fact about convex conjugates can be obtained as a corollary of Danskin's theorem (see, e.g., [7]).

FACT 2.1. *Convex conjugate ψ^* of a function ψ is convex. Moreover, if ψ is strictly convex, ψ^* is differentiable, and the following holds:*

$$\nabla \psi^*(\mathbf{z}) = \operatorname*{argmax}_{\mathbf{x} \geq \mathbf{0}} \{\langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{x})\}.$$

2.2. Fair packing and covering. Recall that α-fair packing problems were defined by (P-a). In the analysis, there are three regimes of α that are handled separately: $\alpha \in [0, 1)$, $\alpha = 1$, and $\alpha > 1$. In these three regimes, the α-fair utilities f_α exhibit very different behaviors, as illustrated in Figure 1. When $\alpha = 0$, f_α is just the linear utility, and (P-a) is a packing LP. As α increases from zero to one, f_α remains nonnegative, but its shape approaches the shape of the natural logarithm. In this regime, any feasible solution $\mathbf{x}$ has optimality gap that is bounded by a constant multiple of $f_\alpha(\mathbf{x}^*)$, where $\mathbf{x}^*$ is the solution to (P-a), as $f_\alpha(\mathbf{x}^*) - f_\alpha(\mathbf{x}) \leq f_\alpha(\mathbf{x}^*)$.



(a) $\alpha \in [0, 1)$ (b) $\alpha = 1$ (c) $\alpha > 1$

FIG. 1. *The three regimes of* α*:* (a) $\alpha \in [0, 1)$*, where* f_α *is nonnegative and equal to zero at* $x = 0$*; for* $\alpha = 0$*,* f_α *is linear, and as* α *approaches* 1 *from below,* f_α *approaches the shifted logarithmic function that equals zero at* $x = 0$*;* (b) $\alpha = 1$*, where* f_α *is the natural logarithm; and* (c) $\alpha > 1$*, where* f_α *is nonpositive; when* α *approaches* 1 *from above,* f_α *approaches the shifted logarithmic function that tends to zero as* x *tends to* ∞*; when* $\alpha \to \infty$*,* f_α *approaches the step function that is* $-\infty$ *for* $x \in (0, 1)$ *and zero for* $x \geq 1$*.*

Thus, in this regime, any algorithm working within the feasible space of (P-a) only needs to reduce the optimality gap by a factor $1/\epsilon$.

When $\alpha = 1$, f_α is the natural logarithm. Even though the range of the possible values of f_α within the feasible set of (P-a) is infinite in this case, it is not hard to choose an initial solution $\mathbf{x}_0$ such that $f_\alpha(\mathbf{x}^*) - f_\alpha(\mathbf{x}_0) \leq n \log(n\rho)$ (see Proposition 2.2). This suffices for our analysis, as we only aim for the final error of the order $n\epsilon$. Notice also that in this case we can only hope for additive error, as the logarithmic function takes both positive and negative values.

When $\alpha > 1$, f_α is nonpositive and its shape approaches the shape of the natural logarithm as $\alpha \to 1$. As α increases and approaches infinity, f_α bends and becomes steeper, approaching the negative indicator of the interval $[0, 1]$. One of the challenges that occurs in the analysis is that it is unclear how to choose an initial feasible point $\mathbf{x}_0$ whose optimality gap would be a constant or polylog factor of $f_\alpha(\mathbf{x}^*)$, as $|f_\alpha(\mathbf{x}_0)| \geq |f_\alpha(\mathbf{x}^*)|$ for any feasible $\mathbf{x}_0$, and any reasonable $\mathbf{x}_0$ could have an optimality gap that is an order $(n\rho)^{\alpha-1}$-factor away from $f_\alpha(\mathbf{x}^*)$. This crucially affects the analysis, as discussed in section 3.3.

Consider the following change of variable:

$$\mathbf{x} = F_\alpha(\hat{\mathbf{x}}) \stackrel{\text{def}}{=} \begin{cases} \hat{\mathbf{x}}^{\frac{1}{1-\alpha}} & \text{if } \alpha \geq 0,\ \alpha \neq 1, \\ \exp(\hat{\mathbf{x}}) & \text{if } \alpha = 1. \end{cases} \tag{2.3}$$

This change of variable does not affect the algorithm; it is only used for the convenience of the analysis. Let $S_\alpha = \mathbb{R}^n_+$, $\hat{f}_\alpha(\mathbf{x}) = \frac{\langle \mathbb{1}, \mathbf{x} \rangle}{1-\alpha}$ for $\alpha \neq 1, \alpha \geq 0$, and $S_\alpha = \mathbb{R}^n$, $\hat{f}_\alpha(\mathbf{x}) = \langle \mathbb{1}, \mathbf{x} \rangle$ for $\alpha = 1$. The problem (P-a) can then equivalently be written (with the abuse of notation) as

$$\max \left\{ \hat{f}_\alpha(\mathbf{x}) : \quad \mathbf{A} F_\alpha(\mathbf{x}) \leq \mathbb{1},\ \mathbf{x} \in S_\alpha \right\}. \tag{P-b}$$

Observe that there is one-to-one correspondence between $\mathbf{x}$ and $F_\alpha(\mathbf{x})$: $\mathbf{x}$ is (P-b)-feasible if and only if $F_\alpha(\mathbf{x})$ is (P-a)-feasible. The objective function value also remains the same: $f_\alpha(F_\alpha(\mathbf{x})) = \hat{f}_\alpha(\mathbf{x})$. Thus, any statements we make about (P-b) can be translated into statements about (P-a); this will be used repeatedly in the analysis.

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

PROPOSITION 2.2. *Let* $\mathbf{x}^*$ *be (any) optimal solution to* (P-b). *Then*

- *if* $\alpha \geq 0$ *and* $\alpha \neq 1$, $\frac{n}{1-\alpha}(n\rho)^{\alpha-1} \leq \hat{f}_\alpha(\mathbf{x}^*) \leq \frac{n}{1-\alpha}$;
- *if* $\alpha = 1$, $-n\log(n\rho) \leq \hat{f}_\alpha(\mathbf{x}^*) \leq 0$.

Proof. The proof is based on the following simple argument. When $F_\alpha(\mathbf{x}) = \frac{1}{n\rho}\mathbb{1}$, $\mathbf{x}$ is feasible and we get a lower bound on the optimal objective value. On the other hand, if $F_\alpha(\mathbf{x}) > \mathbb{1}$, then (as the minimum nonzero entry of $\mathbf{A}$ is at least 1), all constraints are violated, giving an upper bound on the optimal objective value. ∎

Write (P-b) as the following saddle-point problem:

$$\min_{\mathbf{x}\in S_\alpha} -\hat{f}_\alpha(\mathbf{x}) + \max_{\mathbf{y}\geq 0} \langle \mathbf{A}F_\alpha(\mathbf{x}) - \mathbb{1}, \mathbf{y}\rangle . \tag{2.4}$$

The main reason for considering the saddle-point formulation of (P-b) is that after regularization it can be turned into an unconstrained problem over the positive orthant, without losing much in the approximation error, under mild regularity conditions on the steps of the algorithm. In particular, let $(\mathbf{x}^*, \mathbf{y}^*)$ be the optimal primal-dual pair in (2.4). Then, by Fenchel duality (see, e.g., [8, Proposition 5.3.8]), we have that $-\hat{f}_\alpha(\mathbf{x}^*) = \min_{\mathbf{x}\in S_\alpha}\{-\hat{f}_\alpha(\mathbf{x}) + \langle \mathbf{A}F_\alpha(\mathbf{x}) - \mathbb{1}, \mathbf{y}^*\rangle\}$. Hence, $\forall \mathbf{x} \geq \mathbf{0}$,

$$\begin{aligned} -\hat{f}_\alpha(\mathbf{x}^*) &\leq -\hat{f}_\alpha(\mathbf{x}) + \langle \mathbf{A}F_\alpha(\mathbf{x}) - \mathbb{1}, \mathbf{y}^*\rangle - \psi(\mathbf{y}^*) + \psi(\mathbf{y}^*) \\ &\leq -\hat{f}_\alpha(\mathbf{x}) + \max_{\mathbf{y}\geq \mathbf{0}}\{\langle \mathbf{A}F_\alpha(\mathbf{x}) - \mathbb{1}, \mathbf{y}\rangle - \psi(\mathbf{y})\} + \psi(\mathbf{y}^*) \\ &= f_r(\mathbf{x}) + \psi(\mathbf{y}^*), \end{aligned} \tag{2.5}$$

where $f_r(\mathbf{x}) = -\hat{f}_\alpha(\mathbf{x}) + \psi^*(\mathbf{A}F_\alpha(\mathbf{x}) - \mathbb{1})$. Function $\psi^*(\mathbf{A}F_\alpha(\mathbf{x}) - \mathbb{1})$ can also be viewed as an approximate barrier for the packing polytope. The main idea is to show that we can choose the function ψ so that f_r closely approximates $-f_\alpha$ around the optimum $\mathbf{x}^*$ and, further, we can recover a $(1+O(\epsilon))$-approximate solution to (P-a) from a $(1+\epsilon)$-approximate solution to $\min_{\mathbf{x}\geq\mathbf{0}} f_r(\mathbf{x})$. This will allow us to focus on the minimization of f_r, without the need to worry about satisfying the packing constraints from (P-a) in each iteration. The following proposition formalizes this statement and introduces the missing parameters. Its proof is provided in Appendix A. In the choice of $\psi(\cdot)$, the factor $C^{-\beta}$ ensures that the algorithm maintains (strict) feasibility. The case $C = 1$ would allow violations of the constraints by a factor $(1+\epsilon)$.

PROPOSITION 2.3. *Let* $\psi(\mathbf{y}) = \sum_{i=1}^m (\frac{{y_i}^{1+\beta}}{C^\beta(1+\beta)} - y_i)$, *where* $\beta = \frac{\epsilon/4}{(1+\alpha)\log(4mn\rho/\epsilon)}$, $C = (1+\epsilon/2)^{1/\beta}$, *and* $\epsilon \in (0, \min\{\frac{1}{2}, \frac{1}{10|\alpha-1|}\})$ *is the approximation parameter. Then*

1. $f_r(\mathbf{x}) = -\hat{f}_\alpha(\mathbf{x}) + \frac{C\beta}{1+\beta}\sum_{i=1}^m (\mathbf{A}F_\alpha(\mathbf{x}))_i^{\frac{1+\beta}{\beta}}$;
2. *let* $\mathbf{x}_r^* = \operatorname{argmin}_{\mathbf{x}\in S_\alpha} f_r(\mathbf{x})$, $\mathbf{x}_\alpha^*$ *be a solution to* (P-a), *and* $\hat{\mathbf{x}}_r = F_\alpha(\mathbf{x}_r^*)$. *Then* $\hat{\mathbf{x}}_r$ *is* (P-a)*-feasible and*

$$-f_\alpha(\hat{\mathbf{x}}_r) + f_\alpha(\mathbf{x}_\alpha^*) \leq f_r(\mathbf{x}_r^*) + f_\alpha(\mathbf{x}_\alpha^*) \leq 2\epsilon_f \overset{\text{def}}{=} 2\begin{cases} \epsilon n & \text{if } \alpha = 1, \\ \epsilon(1-\alpha) f_\alpha(\mathbf{x}_\alpha^*) & \text{if } \alpha \neq 1. \end{cases}$$

A natural counterpart to α-fair packing problems is the β-fair covering, defined in (C). Similarly as in the case of α-fair packing, when $\beta = 0$, the problem reduces to the covering LP. It is not hard to show (using similar arguments as in [27]) that when $\beta \to \infty$, the optimal solutions to (C) converge to the min-max fair allocation.

For our analysis, it is useful to work with the Lagrangian dual of (C), given by

$$\max_{\mathbf{x}\geq\mathbf{0}} \langle \mathbb{1}, \mathbf{x}\rangle - \frac{\beta}{1+\beta}\sum_{i=1}^{m}(\mathbf{A}\mathbf{x})_i^{(1+\beta)/\beta}.$$

In particular, solving the dual of (C) is the same as minimizing $f_r(\mathbf{x})$ from the packing problem, with $\alpha = 0$ and β from the fair covering formulation (C).

The following two (simple) propositions will be useful in our analysis.

PROPOSITION 2.4. *Let* $\mathbf{y}^*$ *be an optimal solution to* (C). *Then*

$$\left(\frac{1}{m\rho}\right)^{1+\beta}\frac{m}{1+\beta} \leq \sum_{i=1}^{m}\frac{(y_i^*)^{1+\beta}}{1+\beta} \leq \frac{m}{1+\beta}.$$

PROPOSITION 2.5. *Let* $(\mathbf{y}_\beta^*, \mathbf{x}_\beta^*)$ *be the optimal primal-dual pair for* (C). *Then*

$$\left\langle \mathbb{1}, \mathbf{x}_\beta^*\right\rangle = (1+\beta)g_\beta(\mathbf{y}_\beta^*).$$

Proof. By strong duality, $\langle \mathbb{1}, \mathbf{x}_\beta^*\rangle - \frac{\beta}{1+\beta}\sum_{i=1}^{m}(\mathbf{A}\mathbf{x}_\beta^*)_i^{(1+\beta)/\beta} = g_\beta(\mathbf{y}_\beta^*)$ and $\mathbf{y}_\beta^* = (\mathbf{A}\mathbf{x}_\beta^*)^{1/\beta}$. Combining these two identities completes the proof. ∎

3. Fair packing: Algorithm and convergence analysis overview. The algorithm pseudocode is provided in Algorithm 3.1 (FAIRPACKING). All parameter choices will become clear from the analysis.

Observe that Algorithm 3.1 can be implemented in the distributed model described in the introduction, as all that is needed for each distributed agent j are (i) global problem parameters m, n, ρ, α, and ϵ; and (ii) the slack $(\mathbf{A}F_\alpha(\mathbf{x}^{(k-1)}) - \mathbb{1})_i$ for all constraints in which j participates (i.e., for i such that $A_{ij} \neq 0$), in each iteration k, as this information suffices for computing the jth coordinate of the truncated gradient, which in turn suffices for computing the new state $x_j^{(k)}$.

Algorithm 3.1 FAIRPACKING$(\mathbf{A}, \epsilon, \alpha)$.

$\mathbf{x}^{(0)} = \left(\frac{1-\epsilon}{n\rho}\right)^{1-\alpha}\mathbb{1}$ for $\alpha \neq 1$, $\mathbf{x}^{(0)} = \exp\left(\frac{1-\epsilon}{n\rho}\right)\mathbb{1}$ for $\alpha = 1$, $\beta = \frac{\epsilon/4}{(1+\alpha)\log(4mn\rho/\epsilon)}$
if $\alpha < 1$ **then**
  $\mathbf{z}^{(0)} = \exp(\epsilon/4)\mathbb{1}$, $\beta' = \frac{(1-\alpha)\epsilon/4}{\log(n\rho/(1-\epsilon))}$, $h = \frac{(1-\alpha)\beta\beta'}{16\epsilon(1+\alpha\beta)}$
  for $k = 1$ to $K = \lceil 2/((1-\alpha)h\epsilon)\rceil$ **do**
    $\mathbf{x}^{(k)} = (\mathbb{1} + \mathbf{z}^{(k-1)})^{-1/\beta'}$
    $\mathbf{z}^{(k)} = \mathbf{z}^{(k-1)} + \epsilon h\overline{\nabla f}_r(\mathbf{x}^{(k)})$
else if $\alpha = 1$ **then**
  for $k = 1$ to $K = \left\lceil 10\frac{\log^2(8\rho mn/\epsilon)}{\epsilon\beta}\right\rceil$ **do**
    $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \frac{\beta}{4(1+\beta)}\overline{\nabla f}_r(\mathbf{x}^{(k-1)})$
else
  for $k = 1$ to $K = \left\lceil 800\frac{(1+\alpha)^2\log(n\rho/(\epsilon\min\{\alpha-1,1\}))}{\beta\min\{\alpha-1,1\}}\right\rceil$ **do**
    $\mathbf{x}^{(k)} = (\mathbf{I} - \frac{\beta(1-\alpha)\,\mathrm{diag}(\overline{\nabla f}_r(\mathbf{x}^{(k-1)})}{4(1+\alpha\beta))})\mathbf{x}^{(k-1)}$
return $F_\alpha(\mathbf{x}^{(K)})$

We remark here that while the absolute constant in the iteration count for the $\alpha > 1$ may appear large, we expect the actual constant to be much smaller in practice.

This is because we have made no effort to optimize the constants, and have instead focused on reducing the dependence on problem parameters ϵ, m, n, and ρ. We also note that it is possible to improve the empirical performance of the algorithm by using the autoscale idea from [1]. The autoscale idea speeds up the convergence in the initial iterations in which all of the constraints are loose by permitting the coordinates of $\mathbf{x}^{(k)}$ to increase at a faster rate. In particular, if, for some j, it holds $(\mathbf{A}F_\alpha(\mathbf{x}^{(k)}))_i \leq 1-\epsilon$ for all i with $A_{ij} \neq 0$, then $\hat{x}_j = (F_\alpha(\mathbf{x}^{(k)}))_j$ is scaled by $\frac{1-\epsilon}{\max_{i:A_{ij}\neq 0}(\mathbf{A}F_\alpha(\mathbf{x}^{(k)}))_i}$. Implementing autoscale has no effect on the analysis.

We start by characterizing the "local smoothness" of f_r which will be crucial for the analysis.

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 $\mathbf{x} > 0$, which will hold throughout, due to the particular initialization and the choice of the steps. The proof of Lemma 3.1 is provided in the appendix.

LEMMA 3.1. *Suppose that* $\alpha \neq 1$ *and* $\mathbf{x} > 0$. *If* $\Gamma = \operatorname{diag}(\boldsymbol{\gamma})$, $\gamma_j = -\frac{c_j}{4} \cdot \frac{\beta(1-\alpha)}{1+\alpha\beta}\overline{\nabla_j f}_r(\mathbf{x})$ *for* $c_j \in [0,1]$, *then*

$$f_r(\mathbf{x}+\Gamma\mathbf{x}) - f_r(\mathbf{x}) \leq \sum_{j=1}^{n}\left(1-\frac{c_j}{2}\right)\gamma_j x_j \nabla_j f_r(\mathbf{x}).$$

If $\alpha = 1$ *and* $\Delta\mathbf{x} \geq 0$ *is such that* $\Delta x_j = -\frac{c_j\beta}{4(1+\beta)}\overline{\nabla_j f}_r(\mathbf{x})$ *for* $c_j \in [0,1]$, *then*

$$f_r(\mathbf{x}+\Delta\mathbf{x}) - f_r(\mathbf{x}) \leq \sum_{j=1}^{n}\left(1-\frac{c_j}{2}\right)\Delta x_j \nabla_j f_r(\mathbf{x}).$$

Lemma 3.1 also allows us to guarantee that the algorithm always maintains feasible solutions, as stated in the following proposition.

PROPOSITION 3.2. *Solution* $\mathbf{x}^{(k)}$ *computed by* FAIRPACKING *at any iteration* $k \geq 0$ *is* (P-b)*-feasible. Equivalently,* $F_\alpha(\mathbf{x}^{(k)})$ *is* (P-a)*-feasible.*

Proof. By the initialization and steps of FAIRPACKING, $\mathbf{x}^{(k)} \in S_\alpha\ \forall k$. It remains to show that it must be $\mathbf{A}F_\alpha(\mathbf{x}^{(k)}) \leq \mathbb{1}\ \forall k$. Observe that $\mathbf{A}\mathbf{x}^{(0)} \leq (1-\epsilon)\mathbb{1}$. Suppose that in some iteration k, $\exists i$ such that $(\mathbf{A}F_\alpha(\mathbf{x}^{(k)}))_i \geq 1-\epsilon/8$. Fix one such i and let k be the 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 $A_{ij}(x_j^{(k)})^{\frac{1}{1-\alpha}} \geq \frac{1}{4n}$ (there must exist at least one such j, as $(\mathbf{A}F_\alpha(\mathbf{x}))_i \geq 1-\frac{\epsilon}{8} \geq \frac{7}{8}$), we have $(x_j^{(k)})^{\frac{1}{1-\alpha}} \geq \frac{1}{4n\rho}$ and $\nabla_j f_r(\mathbf{x}^{(k)}) \geq \frac{1}{1-\alpha}(-1+(\frac{1}{4n\rho})^\alpha(1+\frac{\epsilon}{4})^{\frac{1}{\beta}}) > \frac{1}{1-\alpha}$. Hence, using Lemma 3.1, all x_j such that $A_{ij}(x_j^{(k)})^{\frac{1}{1-\alpha}} \geq \frac{1}{4n}$ must decrease, which implies $(\mathbf{A}(\mathbf{x}^{(k+1)})^{\frac{1}{1-\alpha}})_i \leq (\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}})_i$. On the other hand, by Lemma 3.1, the maximum increase of any coordinate in any iteration is by a factor $1+\frac{\beta(1-\alpha)}{4(1+\alpha\beta)}$. Thus, the maximum increase in $(\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}})_i$ in any iteration is by a factor at most $(1+\frac{\beta(1-\alpha)}{4(1+\alpha\beta)})^{1-\alpha} \leq e^{\frac{\beta}{4}} \leq 1+\frac{\epsilon}{8}$, and it follows that it must be $\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \leq \mathbb{1}\ \forall k$. The equivalence of feasibility of $F_\alpha(\mathbf{x}^{(k)})$ was already discussed in section 2.2. ∎

3.2. Main theorem. Our main results are summarized in the following theorem. The theorem is proved through Theorems 4.4, 4.8, and 4.14 in section 4.

THEOREM 3.3. *Given* $\mathbf{A}$, $\alpha \geq 0$, *and* $\epsilon \in (0, \min\{\frac{1}{2}, \frac{1}{10|\alpha-1|}\}]$, *let* $\mathbf{x}_\alpha^{(K)} = F_\alpha(\mathbf{x}^{(K)})$ *be the solution produced by* FAIRPACKING *and let* $\mathbf{x}_\alpha^*$ *be the optimal solution to* (P-a). *Then* $\mathbf{x}_\alpha^{(K)}$ *is* (P-a)*-feasible and* $f_\alpha(\mathbf{x}_\alpha^*) - f_\alpha(\mathbf{x}_\alpha^{(K)}) = O(\epsilon_f)$, *where*

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

The total number of iterations taken by the algorithm is

$$K = \begin{cases} O\big(\frac{\log(n\rho)\log(mn\rho/\epsilon)}{(1-\alpha)^3\epsilon^2}\big) & \text{if } \alpha \in [0,1), \\ O\big(\frac{\log^3(\rho mn/\epsilon)}{\epsilon^2}\big) & \text{if } \alpha = 1, \\ O\big(\max\big\{\frac{\alpha^3\log(1/\epsilon)\log(mn\rho/\epsilon)}{\epsilon}, \frac{\log(\frac{1}{\epsilon(\alpha-1)})\log(mn\rho/\epsilon)}{\epsilon(\alpha-1)}\big\}\big) & \text{if } \alpha > 1. \end{cases}$$

3.3. Approximate duality gap. The proof relies on the construction of an approximate duality gap, in the framework of [13]. 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 $-f_\alpha(\mathbf{x}_\alpha^{(k)}) + f_\alpha(\mathbf{x}_\alpha^*)$ decreases as the function of the iteration count k, where $\mathbf{x}_\alpha^{(k)} = F_\alpha(\mathbf{x}^{(k)})$ (recall that, by Proposition 3.2, $\mathbf{x}_\alpha^{(k)}$ is (P-a)-feasible). By construction of f_r, we have that $f_r(\mathbf{x}^{(k)}) \geq -f_\alpha(\mathbf{x}_\alpha^{(k)})$, hence it is an upper bound on $-f_\alpha(\mathbf{x}_\alpha^{(k)})$. In the analysis, we will use $U_k = f_r(\mathbf{x}^{(k+1)})$ as the upper bound. The lower bound L_k needs to satisfy $L_k \leq -f_\alpha(\mathbf{x}_\alpha^*)$. The approximate optimality (or duality, see [13]) gap at iteration k is defined as $G_k = U_k - L_k$.

The goal is to show that G_k decreases at rate $1/H_k$, 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" error E_k. This argument is equivalent to stating that

$$-f_\alpha(\mathbf{x}_\alpha^{(k+1)}) + f_\alpha(\mathbf{x}_\alpha^*) \leq U_k - L_k = G_k \leq \frac{H_0}{H_k}G_0 + \frac{\sum_{\ell=1}^k E_\ell}{H_k},$$

which gives the standard form of convergence for first-order methods.

For $\alpha > 1$, it is unclear how to initialize the algorithm to guarantee a sufficiently small initial gap (and the right change in the gap in general). Instead, we will only require that the gap argument is valid on a subsequence of the iterates. We will argue that in the remaining iterations, f_r must decrease by a large multiplicative factor, so that either way we approach a $(1+\epsilon)$-approximate solution at the target rate.

Local smoothness and the upper bound. As already mentioned, our upper bound of choice will be $U_k = f_r(\mathbf{x}^{(k+1)})$. The reason that the upper bound "looks one step ahead" is that it will hold a sufficiently lower value than $f_r(\mathbf{x}^{(k)})$ (it will always decrease, due to Lemma 3.1) to compensate for any decrease in the lower bound L_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 2.3):

$$-f_\alpha(\mathbf{x}_\alpha^*) \geq f_r(\mathbf{x}_r^*) - 2\epsilon_f \geq \frac{\sum_{\ell=0}^k h_\ell \left(f_r(\mathbf{x}^{(\ell)}) + \langle \nabla f_r(\mathbf{x}^{(\ell)}), \mathbf{x}_r^* - \mathbf{x}^{(\ell)}\rangle\right)}{H_k}. \tag{3.1}$$

Even though simple, we will show that this lower bound can be used for 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 $\nabla f(\mathbf{x}^{(k)})$ (and global problem parameters), when $\alpha \neq 1$, the decrease is also a function of the current solution $\mathbf{x}^{(k)}$. This means that we would need to be able to relate $\sum_{j=1}^{n} (x_j^{(k)} - x_j^*)^2 / x_j^{(k)}$ to the value of $f_r(\mathbf{x}^*)$, which is not clearly possible (see the convergence argument from section 4.2 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 to our note on packing and covering LP [12]. Denote $\mathbf{x}^* = \mathbf{x}_r^* = \operatorname{argmin}_{\mathbf{u}} f_r(\mathbf{u})$. We have

$$f_r(\mathbf{x}^*) \geq \frac{\sum_{\ell=0}^{k} h_\ell (f_r(\mathbf{x}^{(\ell)}) - \langle \nabla f_r(\mathbf{x}^{(\ell)}), \mathbf{x}^{(\ell)} \rangle) + \sum_{\ell=0}^{k} h_\ell \left\langle \frac{\overline{\nabla f}_r(\mathbf{x}^{(\ell)})}{1-\alpha}, \mathbf{x}^* \right\rangle}{H_k}.$$

Let $\phi : \mathbb{R}_+^n \to \mathbb{R}$ be a convex function (that will be specified later). Adding and subtracting $\frac{\phi(\mathbf{x}^*)}{1-\alpha}$ to the right-hand side of the last inequality, and then replacing $\mathbf{x}^*$ with the minimizer of $\sum_{\ell=0}^{k} h_\ell \langle \overline{\nabla f}_r(\mathbf{x}^{(\ell)}), \mathbf{u} \rangle + \phi(\mathbf{u})$ over $\mathbf{u} \geq 0$, we get

$$\begin{aligned} f_r(\mathbf{x}^*) \geq L_k^{\alpha<1} + 2\epsilon_f \stackrel{\text{def}}{=} & \frac{\sum_{\ell=0}^{k} h_\ell (f_r(\mathbf{x}^{(\ell)}) - \langle \nabla f_r(\mathbf{x}^{(\ell)}), \mathbf{x}^{(\ell)} \rangle) - \frac{1}{1-\alpha}\phi(\mathbf{x}^*)}{H_k} \\ & + \frac{\min_{\mathbf{u} \geq \mathbf{0}} \left\{ \sum_{\ell=0}^{k} h_\ell \left\langle \overline{\nabla f}_r(\mathbf{x}^{(\ell)}), \mathbf{u} \right\rangle + \phi(\mathbf{u}) \right\}}{(1-\alpha) H_k}. \end{aligned}$$

Note that the same lower bound cannot be derived for $\alpha \geq 1$. The reason is that we cannot perform a gradient truncation, as for $\alpha > 1$ (resp., $\alpha = 1$), $\langle \nabla f_r(\mathbf{x}), \mathbf{x}^* \rangle \geq \frac{1}{1-\alpha} \langle \overline{\nabla f}_r(\mathbf{x}), \mathbf{x}^* \rangle$ (resp., $\langle \nabla f_r(\mathbf{x}), \mathbf{x}^* \rangle \geq \langle \overline{\nabla f}_r(\mathbf{x}), \mathbf{x}^* \rangle$) *does not hold*.

For $\alpha > 1$, we make use of the Lagrangian dual of (P-b) $g : \mathbb{R}_+^m \to \mathbb{R}$ given by

$$g(\mathbf{y}) = -\langle \mathbb{1}, \mathbf{y} \rangle - \frac{\alpha}{1-\alpha} \sum_{j=1}^{n} (\mathbf{A}^T \mathbf{y})_j^{-\frac{1-\alpha}{\alpha}}. \tag{3.2}$$

Finally, we note that it is not clear how to use the Lagrangian dual in the case of $\alpha \leq 1$. When $\alpha < 1$, the terms $-\frac{1}{1-\alpha} (\mathbf{A}^T \mathbf{y})_j^{-\frac{1-\alpha}{\alpha}}$ approach $-\infty$ as $(\mathbf{A}^T \mathbf{y})_j$ approaches zero. A similar argument can be made for $\alpha = 1$, in which case the Lagrangian dual is $g(\mathbf{y}) = -\langle \mathbb{1}, \mathbf{y} \rangle + n + \sum_{j=1}^{n} \log(\mathbf{A}^T \mathbf{y})_j$. In [25], this was handled by ensuring that $(\mathbf{A}^T \mathbf{y})_j$ never becomes "too small," which requires step sizes that are smaller by a factor ϵ and generally leads to much slower convergence.

4. Proof of the main theorem. In this section, we provide the complete proof of the main theorem (Theorem 3.3). The proof is provided by proving three separate theorems (Theorems 4.4, 4.8, and 4.14), each dealing separately with the cases $\alpha \in [0, 1)$, $\alpha = 1$, and $\alpha > 1$, respectively.

4.1. Convergence analysis for $\alpha \in [0, 1)$. Recall that in this setting, the algorithm makes updates of the following form:

$$\begin{aligned} \mathbf{x}^{(k)} &= (\mathbb{1} + \mathbf{z}^{(k-1)})^{-1/\beta'}, \\ \mathbf{z}^{(k)} &= \mathbf{z}^{(k-1)} + \epsilon h \overline{\nabla f}_r(\mathbf{x}^{(k)}). \end{aligned}$$

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 ϕ in two steps, as follows:

$$\begin{aligned}\phi(\mathbf{x}) &\stackrel{\text{def}}{=} \psi(\mathbf{x}) - \left\langle \nabla\psi(\mathbf{x}^{(0)}) + h_0\overline{\nabla f}_r(\mathbf{x}^{(0)}), \mathbf{x}\right\rangle,\\ \psi(\mathbf{x}) &\stackrel{\text{def}}{=} \frac{1}{\epsilon}\left(\langle \mathbb{1}, \mathbf{x}\rangle - \frac{\langle \mathbb{1}, \mathbf{x}^{1-\beta'}\rangle}{1-\beta'}\right),\end{aligned} \tag{4.1}$$

where $\beta' = \frac{(1-\alpha)\epsilon/4}{\log(n\rho/(1-\epsilon))}$. This particular choice of ϕ is made for the following reasons. First, $\epsilon\psi(\mathbf{x})$ closely approximates $\langle \mathbb{1}, \mathbf{x}\rangle$ (up to an ϵ multiplicative factor, unless $\langle \mathbb{1}, \mathbf{x}\rangle$ is negligible). This will ensure that $\frac{1}{1-\alpha}\phi(\mathbf{x}^*)$ is within $O(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha)$, which will allow us to bound the initial gap by $O(1-\alpha)f_\alpha(\mathbf{x}^*)$.

To understand the role of $-\langle \nabla\psi(\mathbf{x}^{(0)}) + h_0\overline{\nabla f}_r(\mathbf{x}), \mathbf{x}\rangle$, notice that the steps of FairPacking are defined as $\mathbf{x}^{(k+1)} = \operatorname{argmin}_{\mathbf{u}\geq\mathbf{0}}\{\sum_{\ell=0}^{k} h_\ell\langle\overline{\nabla f}_r(\mathbf{x}^{(\ell)}), \mathbf{u}\rangle + \phi(\mathbf{u})\}$. The role of the term $-\langle \nabla\psi(\mathbf{x}^{(0)}) + h_0\overline{\nabla f}_r(\mathbf{x}^{(0)}), \mathbf{x}\rangle$ is to ensure that $\mathbf{x}^{(1)} = \mathbf{x}^{(0)}$, which will allow us to properly initialize the gap. Finally, the scaling factor $\frac{1}{\epsilon}$ ensures that $\mathbf{z}^{(k)} \leq (1+\epsilon/2)\mathbb{1}$ (see the proof of Lemma 4.13), which will allow us to argue that the steps satisfy the assumptions of Lemma 3.1. We also need to guarantee that

$$\mathbf{z}^{(k)} \stackrel{\text{def}}{=} \epsilon\left(\sum_{\ell=1}^{k} h_\ell\overline{\nabla f}_r(\mathbf{x}^{(\ell)}) - \nabla\psi(\mathbf{x}^{(0)})\right) \tag{4.2}$$

is bounded below element-wise by $-O(\epsilon)$ to ensure that the upper bound can compensate for any decrease in the lower bound. Some of these statements are formalized below.

Proposition 4.1. *Let* $\mathbf{z}^{(k)}$, $\psi(\cdot)$, *and* $\phi(\cdot)$ *be defined as in* (4.1), (4.2). *Define* $\widehat{\psi}(\mathbf{z}^{(k)}) = -\frac{\beta'/\epsilon}{1-\beta'}\sum_{j=1}^{n}(1+z_j^{(k)})^{-\frac{1-\beta'}{\beta'}}$. *Then*

1. $\widehat{\psi}(\mathbf{z}^{(k)}) = \min_{\mathbf{u}\geq\mathbf{0}}\left\{\sum_{\ell=0}^{k} h_\ell\left\langle\overline{\nabla f}_r(\mathbf{x}^{(\ell)}), \mathbf{u}\right\rangle + \phi(\mathbf{u})\right\}$;
2. $\mathbf{x}^{(1)} = \operatorname{argmin}_{\mathbf{u}\geq\mathbf{0}}\{h_0\left\langle\overline{\nabla f}_r(\mathbf{x}^{(0)}), \mathbf{u}\right\rangle + \phi(\mathbf{u})\} = \mathbf{x}^{(0)}$ *and* $\mathbf{x}^{(k+1)} = \epsilon\nabla\widehat{\psi}(\mathbf{z}^{(k)})$.
3. $(\epsilon/4)\mathbb{1} \leq \mathbf{z}^{(k)} \leq (\epsilon/2)\mathbb{1}$.

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

For the second part, by the definition of ϕ and the first-order optimality condition,

$$\operatorname*{argmin}_{\mathbf{u}\geq\mathbf{0}}\left\{h_0\left\langle\overline{\nabla f}_r(\mathbf{x}^{(0)}), \mathbf{u}\right\rangle + \phi(\mathbf{u})\right\} = \operatorname*{argmin}_{\mathbf{u}\geq\mathbf{0}}\left\{\psi(\mathbf{u}) - \left\langle\nabla\psi(\mathbf{x}^{(0)}), \mathbf{u}\right\rangle\right\} = \mathbf{x}^{(0)}.$$

Similarly, for $\mathbf{x}^{(k+1)}$, we have $\mathbf{x}^{(k+1)} = \operatorname{argmin}_{\mathbf{u}\geq\mathbf{0}}\{\langle\mathbf{z}^{(k)}, \mathbf{u}\rangle + \epsilon\psi(\mathbf{u})\}$. It is not hard to verify that $x_j^{(k+1)} = (1+z_j^{(k)})^{-1/\beta'} = \epsilon\nabla_j\widehat{\psi}(\mathbf{z}^{(k)})$. For the last part, recall that $x_j^{(0)} = (\frac{1-\epsilon}{n\rho})^{\frac{1}{1-\alpha}}$ and observe that $\nabla_j\psi(\mathbf{x}) = \frac{1}{\epsilon}(1 - {x_j}^{-\beta'})$. Hence

$$z_j^{(0)} = \left(\frac{n\rho}{1-\epsilon}\right)^{\frac{\beta'}{1-\alpha}} - 1 = \left(\frac{n\rho}{1-\epsilon}\right)^{\frac{\epsilon/4}{\log(n\rho/(1-\epsilon))}} - 1 = \exp(\epsilon/4) - 1.$$

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

Using Proposition 4.1, we can now bound the initial gap, as follows.

Proposition 4.2. *Let* $h_0 = H_0 = 1$. *Then* $H_0G_0 - 2\epsilon(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha) \leq 2f_\alpha(\mathbf{x}^*_\alpha)$.

Proof. From Proposition 4.1, $U_1 = f_r(\mathbf{x}^{(1)}) = f_r(\mathbf{x}^{(0)})$ and thus: $H_0G_0 = \frac{-\widehat{\psi}(\mathbf{z}^{(0)})+\phi(\mathbf{x}^*)}{1-\alpha} + 2\epsilon(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha)$. The rest of the proof follows by bounding $\widehat{\psi}(\mathbf{z}^{(0)})$ and $\phi(\mathbf{x}^*_r)$. For the former, it is not hard to verify that $\sum_{j=1}^n \frac{(x_j^{(0)})^{1-\beta'}}{1-\beta'} \le (1+\epsilon/2)\langle \mathbb{1}, \mathbf{x}^{(0)}\rangle$. Hence, as $x_j^{(1)} = x_j^{(0)} = (z_j^{(0)})^{-1/\beta'}$, we have

$$-\widehat{\psi}(\mathbf{z}^{(0)}) \le \frac{\beta'(1+\epsilon/2)}{\epsilon}\left\langle \mathbb{1}, \mathbf{x}^{(0)}\right\rangle \le \frac{1}{2}(1-\alpha)\left\langle \mathbb{1}, \mathbf{x}^{(0)}\right\rangle \le \frac{1}{2}(1-\alpha)^2 f_\alpha(\mathbf{x}^*_\alpha).$$

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

$$\begin{aligned}\phi(\mathbf{x}^*) &\le -\left\langle \nabla\psi(\mathbf{x}^{(0)}) + h_0\overline{\nabla f}_r(\mathbf{x}^{(0)}), \mathbf{x}^*\right\rangle \le -\langle(-1/2-1)\mathbb{1}, \mathbf{x}^*\rangle \\ &\le \frac{3}{2}\langle \mathbb{1}, \mathbf{x}^*\rangle \le \frac{3}{2}(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha),\end{aligned}$$

as $\overline{\nabla f}_r(\mathbf{x}) \ge -\mathbb{1}\ \forall \mathbf{x}$ and $\nabla\psi(\mathbf{x}^{(0)}) = \mathbf{z}^{(0)}/\epsilon \ge -(1/2)\mathbb{1}$ (due to Proposition 4.1). ∎

The crucial part of the convergence analysis is to show that for some choice of step sizes h_k, $H_kG_k \le H_{k-1}G_{k-1} + 2h_k\epsilon_f$. Note that to make the algorithm as fast as possible (since its convergence rate is proportional to H_k), we would like to set the h_k's as large as possible. However, enforcing the condition $H_kG_k \le H_{k-1}G_{k-1} + 2h_k\epsilon_f$ will set an upper bound on the choice of h_k. We have the following lemma.

LEMMA 4.3. *If* $G_{k-1} - 2\epsilon_f \le 2f_\alpha(\mathbf{x}^*_\alpha)$ *and* $h_k \le \frac{(1-\alpha)\beta\beta'}{16\epsilon(1+\beta)} = \theta(\frac{(1-\alpha)^2\epsilon}{\log(n\rho)\log(mn\rho/\epsilon)})$, *then* $H_kG_k \le H_{k-1}G_{k-1} + 2h_k\epsilon_f\ \forall k \ge 1$.

Proof. The role of the assumption $G_{k-1} - 2\epsilon_f \le 2(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha)$ is to guarantee that $\mathbf{z}^{(k-1)} \ge -(\epsilon/2)\mathbb{1}$. Namely, if $z_j^{(k-1)} < -\epsilon/2$ for any j, $\psi^*(\mathbf{z}^{(k-1)})$ blows up, making the gap G_{k-1} much larger than $3(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha)$. This is not hard to argue (see also a similar argument in [12]) and hence we omit the details and assume from now on that $\mathbf{z}^{(k-1)} \ge -(\epsilon/2)\mathbb{1}$. Note that this assumption holds initially due to Proposition 4.1. Observe that as $\epsilon < 1/H_k$ and $\overline{\nabla f}_r(\mathbf{x}^{(\ell)}) \le \mathbb{1}\ \forall \ell$, we also have

$$z_j^{(k)} = \epsilon\left(\sum_{\ell=1}^k h_\ell \overline{\nabla_j f}_r(\mathbf{x}^{(k)}) - \nabla_j\psi(\mathbf{x}^{(0)})\right) \le 1 - \epsilon\nabla_j\psi(\mathbf{x}^{(0)}) \le 1 + \epsilon/2,$$

where we have used $\nabla_j\psi(\mathbf{x}^{(0)}) = z_j^{(0)}/\epsilon \ge -\frac{1}{2}$ (due to Proposition 4.1.3). To be able to apply Lemma 3.1, we need to ensure that

$$|x_j^{(k+1)} - x_j^{(k)}| \le c_j\frac{\beta}{4(1+\beta)}|\overline{\nabla_j f}_r(\mathbf{x}^{(k)})|x_j^{(k)}$$

for all j and for $c_j \in (0,1]$. Recalling the definition of $\mathbf{x}^{(k+1)}$, $x_j^{(k+1)} = \nabla_j\widehat{\psi}(\mathbf{z}^{(k)}) = (1+z_j^{(k)})^{-1/\beta'}$. As $z_j^{(k)} = z_j^{(k-1)} + \epsilon h_k\overline{\nabla_j f}_r(\mathbf{x}^{(k)})$, we have

$$x_j^{(k+1)} = (1 + z_j^{(k-1)} + h_k\epsilon\overline{\nabla_j f}_r(\mathbf{x}^{(k)}))^{-1/\beta} = x_j^{(k)}\left(1 + \frac{\epsilon h_k\overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{1+z_j^{(k-1)}}\right)^{-1/\beta'}.$$

Suppose first that $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \le 0$. Then $\frac{\overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{1-\epsilon/2} \le \frac{\overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{1+z_j^{(k-1)}} \le \frac{\overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{2+\epsilon/2}$. As

$\frac{\epsilon h_k}{\beta'} \le (1-\epsilon/2)\frac{(1-\alpha)\beta}{8(1+\beta)}$ and $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \ge -1$ we have

$$1 - \frac{1-\epsilon/2}{2+\epsilon/2} \cdot \frac{(1-\alpha)\beta}{8(1+\beta)} \overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \le \Big(1 + \frac{\epsilon h_k \overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{1+z_j^{(k-1)}}\Big)^{-1/\beta'} \le 1 - \frac{(1-\alpha)\beta}{4(1+\beta)} \overline{\nabla_j f}_r(\mathbf{x}^{(k)}).$$

Similarly, when $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) > 0$, $\frac{\overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{2+\epsilon/2} \le \frac{\overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{1+z_j^{(k-1)}} \le \frac{\overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{1-\epsilon/2}$. As $\frac{\epsilon h_k}{\beta'} \le (1-\epsilon/2)\frac{(1-\alpha)\beta}{8(1+\beta)}$ and $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \le 1$ we have

$$1 - \frac{(1-\alpha)\beta}{4(1+\beta)} \overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \le \Big(1 + \frac{\epsilon h_k \overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{1+z_j^{(k-1)}}\Big)^{-1/\beta'} \le 1 - \frac{1-\epsilon/2}{2+\epsilon/2} \cdot \frac{(1-\alpha)\beta}{8(1+\beta)} \overline{\nabla_j f}_r(\mathbf{x}^{(k)}).$$

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

$$\begin{aligned} H_k U_k - H_{k-1}U_{k-1} &= H_k(U_k - U_{k-1}) + h_k U_{k-1} \\ &= H_k(f_r(\mathbf{x}^{(k+1)}) - f_r(\mathbf{x}^{(k)})) + h_k f_r(\mathbf{x}^{(k)}) \\ &\le h_k f_r(\mathbf{x}^{(k)}) - \frac{H_k \beta}{50(1+\alpha\beta)} \sum_{j=1}^n x_j^{(k)} \nabla_j f_r(\mathbf{x}^{(k)}) \overline{\nabla_j f}_r(\mathbf{x}^{(k)}). \end{aligned} \tag{4.3}$$

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

$$\begin{aligned} &H_k L_k - H_{k-1}L_{k-1} \\ &= h_k \Big(f_r(\mathbf{x}^{(k)}) - \Big\langle \nabla f_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} \Big\rangle\Big) + \frac{1}{1-\alpha}\Big(\widehat{\psi}(\mathbf{z}^{(k)}) - \widehat{\psi}(\mathbf{z}^{(k-1)})\Big) + 2h_k\epsilon_f. \end{aligned} \tag{4.4}$$

Using Taylor's theorem

$$\begin{aligned} &\widehat{\psi}(\mathbf{z}^{(k)}) - \widehat{\psi}(\mathbf{z}^{(k-1)}) \\ &= \Big\langle \nabla\widehat{\psi}(\mathbf{z}^{(k-1)}), \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)} \Big\rangle + \frac{1}{2}\Big\langle \nabla^2\widehat{\psi}(\hat{\mathbf{z}})(\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}), \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)} \Big\rangle, \end{aligned} \tag{4.5}$$

where $\hat{\mathbf{z}} = \mathbf{z}^{(k-1)} + t(\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)})$ for some $t \in [0,1]$. Recall that $\epsilon\nabla_j\widehat{\psi}(\mathbf{z}^{(k-1)}) = (1+z_j^{(k-1)})^{-1/\beta'} = x_j^{(k)}$ and $\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)} = \epsilon h_k \overline{\nabla f}_r(\mathbf{z}^{(k)})$. Observe that $\nabla^2_{jj}\widehat{\psi}(\mathbf{z}) = -\frac{1}{\epsilon\beta'}(1+z_j)^{-(1+\beta')/\beta'}$, $\nabla^2_{jk}\widehat{\psi}(\mathbf{z}) = 0$ for $j \ne k$. As $z_j^{(k-1)} \ge -\epsilon/2$ and $z_j^{(k)} = z_j^{(k-1)} + \epsilon h_k \overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \ge 1 - \epsilon h_k$, we have that

$$(1+z_j^{(k)})^{-\frac{1-\beta'}{\beta'}} \le (1-\epsilon h_k)^{-\frac{1-\beta'}{\beta'}}(1+z_j^{(k-1)})^{-\frac{1-\beta'}{\beta'}} < (1+\epsilon/2)(1+z_j^{(k-1)})^{-\frac{1-\beta'}{\beta'}},$$

as $\frac{\epsilon h_k}{\beta'} \le (1-\epsilon/2)\frac{(1-\alpha)\beta}{8(1+\beta)}$. Further, as $x_j^{(k)} = (1+z_j^{(k-1)})^{-1/\beta'}$ and $z_j^{(k)} \ge 1-\epsilon/2$, we have that $(1+z_j^{(k-1)})^{-\frac{1-\beta'}{\beta'}} \le x_j^{(k)}/(1-\epsilon/2)$. Hence, (4.5) implies

$$\widehat{\psi}(\mathbf{z}^{(k)}) - \widehat{\psi}(\mathbf{z}^{(k-1)}) \ge h_k \Big\langle \overline{\nabla f}_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} \Big\rangle - \frac{3(\epsilon h_k)^2}{2\beta'} \sum_{j=1}^n x_j^{(k)} (\overline{\nabla_j f}_r(\mathbf{x}^{(k)}))^2. \tag{4.6}$$

Using (4.3), (4.4), and (4.6) to complete the proof, it suffices to show that, $\forall j$,

$$\xi_j \stackrel{\text{def}}{=} h_k \left(\nabla_j f_r(\mathbf{x}^{(k)}) - \frac{1}{1-\alpha} \overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \right) + \frac{3(\epsilon h_k)^2}{2\beta'(1-\alpha)} (\overline{\nabla_j f}_r(\mathbf{x}^{(k)}))^2 - \frac{H_k \beta}{50(1+\alpha\beta_k)} \nabla_j f_r(\mathbf{x}^{(k)}) \overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \leq 0.$$

Consider the following two cases for $\overline{\nabla_j f}_r(\mathbf{x}^{(k)})$:

Case 1: $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) < 1$. Then $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) = (1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)})$, and we have

$$\xi_j = \frac{(\overline{\nabla_j f}_r(\mathbf{x}^{(k)}))^2}{1-\alpha} \Big(\frac{3(\epsilon h_k)^2}{2\beta'} - \frac{H_k \beta}{50(1+\alpha\beta)} \Big) < 0,$$

as $\epsilon h_k \leq (1-\epsilon/2)\frac{\beta'(1-\alpha)\beta}{8(1+\beta)}$.

Case 2: $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) = 1$. Then $\nabla_j f_r(\mathbf{x}^{(k)}) \geq \frac{1}{1-\alpha} \geq 1$, and

$$\xi_j = \frac{h_k}{1-\alpha} \Big(\frac{3\epsilon^2 h_k}{\beta'} - 1 \Big) + \nabla_j f_r(\mathbf{x}^{(k)}) \Big(h_k - \frac{H_k \beta}{50(1+\alpha\beta)} \Big) \leq 0,$$

by the choice of h_k. ∎

We are now ready to bound the overall convergence of FairPacking for $\alpha < 1$.

THEOREM 4.4. *Let* $h_0 = 1$, $h_k = h = \frac{(1-\alpha)\beta\beta'}{16\epsilon(1+\beta)}$ *for* $k \geq 1$. *Then, after at most* $K = \lceil \frac{2}{h(1-\alpha)\epsilon} \rceil = \theta(\frac{\log(n\rho)\log(mn\rho/\epsilon)}{(1-\alpha)^3\epsilon^2})$ *iterations of* FairPacking, *we have that* $\mathbf{x}_a^{(K+1)} = F_\alpha(\mathbf{x}^{(K+1)})^{\frac{1}{1-\alpha}} = (\mathbf{x}^{(K+1)})^{1-\alpha}$ *is* (P-a)*-feasibile and*

$$f_\alpha(\mathbf{x}_a^{(K+1)}) - f_\alpha(\mathbf{x}_\alpha^*) \geq -3\epsilon(1-\alpha) f_\alpha(\mathbf{x}_\alpha^*).$$

Proof. Feasibility of $\mathbf{x}_\alpha^{(K+1)}$ follows from Proposition 3.2, as the steps of FairPacking satisfy the conditions of Lemma 3.1.

Due to Proposition 4.2, the assumptions of Lemma 4.3 hold initially and hence they hold for all k (as Lemma 4.3 itself when applied to iteration k implies that its assumptions hold at iteration $k+1$). Thus, we have $G_K \leq \frac{H_0 G_0}{H_K} + \frac{\sum_{\ell=0}^K h_\ell 2\epsilon_f}{H_K} = \frac{H_0 G_0}{H_K} + 2\epsilon_f$. As, from Proposition 4.2, $H_0 G_0 \leq 2 f_\alpha(\mathbf{x}_\alpha^*)$ and $H_K = Kh \geq \frac{2}{(1-\alpha)\epsilon}$, it follows that $G_k \leq 3\epsilon(1-\alpha) f_\alpha(\mathbf{x}_\alpha^*)$. Finally, recalling that by construction, $-f_\alpha(\mathbf{x}_\alpha^{(K+1)}) + f_\alpha(\mathbf{x}_\alpha^*) \leq G_K$, the claimed statement follows. ∎

4.2. Convergence analysis for $\alpha = 1$. In this setting, the algorithm makes updates of the following form:

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \frac{\beta}{4(1+\beta)} \overline{\nabla f}_r(\mathbf{x}^{(k-1)}).$$

Let us start by bounding the coordinates of the running solutions $\mathbf{x}^{(k)}$, for each iteration k. This will allow us to bound the initial-gap-plus-error $H_0 G_0 + \sum_{\ell=1}^k E_i$ in the convergence analysis.

PROPOSITION 4.5. *In each iteration* k, $-\log(2\rho mC)\mathbb{1} \leq \mathbf{x}^{(k)} \leq \mathbf{0}$.

Proof. Using Proposition 3.2, $\mathbf{x}^{(k)} \leq \mathbf{0}$ follows immediately by $\min_{ij:A_{ij}\neq 0} A_{ij} = 1$. Suppose that in some iteration k, $x_j^{(k)} \leq -\log(2\rho mC) + \epsilon/4$. Then, by Proposition 3.2, $AF_\alpha(\mathbf{x}^{(k)}) \leq \mathbb{1}$, and it follows that

$$\begin{aligned}\nabla_j f_r(\mathbf{x}^{(k)}) &= -1 + C\sum_{i=1}^m A_{ij} e^{x_j} (AF_\alpha(\mathbf{x}^{(k)}))_i^{1/\beta} \\ &\leq -1 + C\sum_{i=1}^m \rho e^{-\log(2\rho mC)+\epsilon/4} \\ &\leq -1 + C\cdot \frac{1}{2\rho mC}\exp(\epsilon/4)\rho m \leq -\frac{1-\epsilon}{2},\end{aligned}$$

where in the second line we have used that $A_{ij} \leq \rho\ \forall i, j$.

Hence, $x_j^{(k)}$ must increase in iteration k. Since the maximum decrease in any coordinate and in any iteration is less than $\epsilon/4$, it follows that $\mathbf{x}^{(k)} \geq -\log(2\rho mC)\mathbb{1}$. ∎

Recall that $U_k = f_r(\mathbf{x}^{(k+1)})$ and $L_k = \frac{\sum_{\ell=0}^k h_\ell \left(f(\mathbf{x}^{(\ell)}) + \left\langle \nabla f_r(\mathbf{x}^{(\ell)}), \mathbf{x}^* - \mathbf{x}^{(\ell)}\right\rangle\right)}{H_k} - 2\epsilon n$. Let us start by bounding the initial gap G_0.

PROPOSITION 4.6. *$A_0 G_0 \leq E_0$, where $E_0 = \frac{2(1+\beta)}{\beta}\cdot \frac{{h_0}^2}{H_0}\|\mathbf{x}^* - \mathbf{x}^{(0)}\|^2 + 2h_0\epsilon n$.*

Proof. By the choice of the initial point $\mathbf{x}^{(0)}$, it follows that $\nabla f_r(\mathbf{x}^{(0)}) \leq \mathbf{0}$, and, thus $\overline{\nabla f}_r(\mathbf{x}^{(0)}) = \nabla f_r(\mathbf{x}^{(0)})$. Using the Cauchy–Schwarz inequality

$$H_0 L_0 \geq h_0 f(\mathbf{x}^{(0)}) - h_0\|\nabla f_r(\mathbf{x}^{(0)})\|\cdot\|\mathbf{x}^* - \mathbf{x}^{(0)}\| - 2H_0\epsilon n, \tag{4.7}$$

while, from Lemma 3.1,

$$H_0 U_0 \leq h_0 f(\mathbf{x}^{(0)}) - H_0\frac{\beta}{8(1+\beta)}\|\nabla f_r(\mathbf{x}^{(0)})\|^2. \tag{4.8}$$

Combining (4.7) and (4.8) with $-a^2 + 2ab \leq b^2\ \forall a, b$, and as $H_0 = h_0$, it follows that

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

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 4.6 and the definition of the gap, would imply $f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^*) \leq G_k \leq \frac{\sum_{i=0}^k E_i}{H_k}$, allowing us to bound the approximation error.

LEMMA 4.7. *If, for $k \geq 1$, $\frac{h_k}{H_k} \leq \frac{\beta}{8(1+\beta)\log(2\rho mC)}$, then $H_k G_k - H_{k-1}G_{k-1} \leq E_k$, where $E_k = \frac{2(1+\beta)}{\beta}\cdot\frac{{h_k}^2}{H_k}\|\mathbf{x}^* - \mathbf{x}^{(k)}\|^2$.*

Proof. Applying the Cauchy–Schwarz inequality and Lemma 3.1

$$H_k L_k - H_{k-1}L_{k-1} \geq h_k f_r(\mathbf{x}^{(k)}) - h_k\sum_{j=1}^n |\nabla_j f_r(\mathbf{x}^{(k)})|\cdot|x_j^* - x_j^{(k)}|, \tag{4.9}$$

$$H_k U_k - H_{k-1}U_{k-1} \leq h_k f_r(\mathbf{x}^{(k)}) - \frac{H_k\beta}{8(1+\beta)}\sum_{j=1}^n \overline{\nabla_j f}_r(\mathbf{x}^{(k)})\nabla_j f_r(\mathbf{x}^{(k)}). \tag{4.10}$$

Hence, combining (4.9) and (4.10),

$$H_kG_k - H_{k-1}G_{k-1} \leq \sum_{j=1}^{n}\left(h_k|\nabla_j f_r(\mathbf{x}^{(k)})|\cdot|x_j^* - x_j^{(k)}| - \frac{H_k\beta}{8(1+\beta)}\overline{\nabla_j f}_r(\mathbf{x}^{(k)})\nabla_j f_r(\mathbf{x}^{(k)})\right). \tag{4.11}$$

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

Case 1: $\nabla_j f_r(\mathbf{x}^{(k)}) \leq 1$. Then $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) = \nabla_j f_r(\mathbf{x}^{(k)})$ and using that $-a^2 + 2ab \leq b^2\ \forall a, b$, it follows that

$$e_j \leq \frac{2(1+\beta)}{\beta}\frac{{h_k}^2}{H_k}(x_j^* - x_j^{(k)})^2. \tag{4.12}$$

Case 2: $\nabla_j f_r(\mathbf{x}^{(k)}) > 1$. Then $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) = 1$. By Proposition 4.5, $-\log(2\rho mC) \leq x_j^{(k)} \leq 0$ and similar bounds can be obtained for x_j^* (see [25]). It follows that

$$e_j \leq |\nabla_j f(\mathbf{x}^{(k)})|\left(h_k\log(2\rho mC) - \frac{H_k\beta}{8(1+\beta)}\right) \leq 0, \tag{4.13}$$

as $\frac{h_k}{H_k} \leq \frac{\beta}{8(1+\beta)\log(2\rho mC)}$.

Combining (4.11)–(4.13) completes the proof. ∎

We are now ready to obtain the final convergence bound for $\alpha = 1$.

THEOREM 4.8. *If* $k \geq 10\frac{\log^2(2\rho mC)}{\epsilon\beta} = O(\frac{\log^3(\rho mn/\epsilon)}{\epsilon^2})$, *then* $\mathbf{x}_\alpha^{(k+1)} = \exp(\mathbf{x}^{(k+1)})$ *is* (P-a)*-feasible and* $f_\alpha(\mathbf{x}_\alpha^{(k+1)}) - f_\alpha(\mathbf{x}_\alpha^*) \geq -3\epsilon n$.

Proof. Combining Proposition 4.6 and Lemma 4.7, we have that if for $\ell \geq 1$, $\frac{h_\ell}{H_\ell} \leq \lambda \stackrel{\text{def}}{=} \frac{\beta}{8(1+\beta)\log(2\rho mC)}$, then $G_k \leq \frac{2(1+\beta)}{H_k\beta}\sum_{\ell=0}^{k}\frac{h_\ell^2}{H_\ell}\|\mathbf{x}^* - \mathbf{x}^{(\ell)}\|^2 + 2n\epsilon$. As discussed before, $\|\mathbf{x}^* - \mathbf{x}^{(\ell)}\|^2 \leq n\log^2(2\rho mC)$, and thus

$$G_k \leq \frac{2(1+\beta)}{\beta}n\log^2(2\rho mC)\frac{1}{H_k}\sum_{\ell=0}^{k}\frac{h_\ell^2}{H_\ell} + 2n\epsilon. \tag{4.14}$$

As the sequence $\{h_\ell\}_{\ell=1}^k$ does not affect the analysis, we can choose it arbitrarily, as long as $\frac{h_\ell}{H_\ell} \leq \lambda$ for $\ell \geq 1$. Let $h_0 = 1$ and $\frac{h_\ell}{H_\ell} = \frac{\beta\epsilon}{8(1+\beta)\log^2(2\rho mC)} < \lambda$ for $\ell \geq 1$. Then

$$G_k \leq \frac{1}{H_k}\frac{2(1+\beta)}{\beta}n\log^2(2\rho mC) + \frac{n\epsilon}{4} + 2n\epsilon.$$

As $\frac{1}{H_k} = \frac{H_0}{H_k} = \frac{H_0}{H_1}\frac{H_1}{H_2}\cdots\frac{H_{k-1}}{H_k} = (1-\frac{h_1}{H_1})^k$, it follows that $G_k \leq 3\epsilon n$. By construction, $-f_\alpha(\mathbf{x}_\alpha^{(k+1)}) + f_\alpha(\mathbf{x}_\alpha^*) \leq 3n\epsilon$, and $\mathbf{x}_\alpha^{(k+1)}$ is (P-a)-feasible due to Proposition 3.2. ∎

4.3. Convergence analysis for $\alpha > 1$. Recall that in this setting, the algorithm makes updates of the following form:

$$\mathbf{x}^{(k)} = \Big(\mathbf{I} - \frac{\beta(1-\alpha)\,\mathrm{diag}(\overline{\nabla f}_r(\mathbf{x}^{(k-1)}))}{4(1+\alpha\beta)}\Big)\mathbf{x}^{(k-1)}.$$

Define the vector $\mathbf{y}^{(k)}$ as

$$y_i^{(k)} = (\mathbf{A}F_\alpha(\mathbf{x}^{(k)}))_i^{1/\beta} = \left(\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}}\right)_i^{1/\beta}. \tag{4.15}$$

Clearly, $\mathbf{y}^{(k)} \geq \mathbf{0}$. Observe that

$$f_r(\mathbf{x}^{(k)}) = -\frac{\left\langle \mathbb{1}, \mathbf{x}^{(k)} \right\rangle}{1-\alpha} + \frac{\beta}{1+\beta}\left\langle \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}}, \mathbf{y}^{(k)} \right\rangle. \tag{4.16}$$

Recall that the Lagrangian dual of (P-b) (and, by the change of variables, (P-a)) is $g(\mathbf{y}) = -\langle \mathbb{1}, \mathbf{y} \rangle + \frac{\alpha}{\alpha-1}\sum_{j=1}^n (\mathbf{A}^T\mathbf{y})_j^{\frac{\alpha-1}{\alpha}}$. Interpreting $\mathbf{y}^{(k)}$ as a dual vector, we can bound the duality gap of a solution $\hat{\mathbf{x}}^{(k)} = F_\alpha(\mathbf{x}^{(k)})$ at any iteration k (using primal feasibility from Proposition 3.2) as

$$-f_\alpha(\hat{\mathbf{x}}^{(k)}) + f_\alpha(\mathbf{x}_\alpha^*) = -\frac{\langle \mathbb{1}, \mathbf{x}^{(k)} \rangle}{1-\alpha} + f_\alpha(\mathbf{x}_\alpha^*) \leq -\frac{\langle \mathbb{1}, \mathbf{x}^{(k)} \rangle}{1-\alpha} - g(\mathbf{y}^{(k)}). \tag{4.17}$$

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

4.3.1. Regularity conditions for the duality gap. The next proposition gives a notion of approximate and aggregate complementary slackness, with $\mathbf{y}^{(k)}$ being interpreted as the vector of dual variables, similarly to [25].

PROPOSITION 4.9. *After at most $O(1/\beta)$ initial iterations, in every iteration*

$$\left\langle \mathbb{1}, \mathbf{y}^{(k)} \right\rangle \leq (1+\epsilon)\left\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle.$$

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

Let $i^* = \operatorname{argmax}_i(\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}})_i$ and

$$S = \{i : (\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}})_i \geq (1-\epsilon/4)(\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}})_{i^*}\}.$$

Then, $\forall \ell \notin S$, $y_\ell^{(k)} \leq (1-\epsilon/4)^{1/\beta} y_{i^*}^{(k)} \leq \frac{\epsilon}{4m} y_{i^*}^{(k)}$. Hence, $\sum_{\ell \notin S} y_\ell^{(k)} \leq \frac{\epsilon}{4} y_{i^*}^{(k)} \leq \frac{\epsilon}{4}\sum_{i\in S} y_i^{(k)}$ and we have $\sum_{i\in S} y_i^{(k)} \geq \frac{1}{1+\epsilon/4}\sum_{i'=1}^m y_{i'}^{(k)}$. It follows that

$$\begin{aligned}\left\langle \mathbf{y}^{(k)}, \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle &\geq \sum_{i\in S} y_i^{(k)}(\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}})_i \geq (1-\epsilon/2)(1-\epsilon/4)\sum_{i\in S} y_i^{(k)} \\ &\geq \frac{(1-\epsilon/2)(1-\epsilon/4)}{1+\epsilon/4}\left\langle \mathbb{1}, \mathbf{y}^{(k)} \right\rangle.\end{aligned}$$

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 $H_k G_k \le H_{k-1} G_{k-1} + O(\epsilon)(1-\alpha) f_\alpha(\mathbf{x}^*_\alpha)$ for some gap G_k), we need to ensure that the argument can be started from a gap $G_0 = O(1)(1-\alpha) f_\alpha(\mathbf{x}^*_\alpha)$. The following lemma gives sufficient conditions for ensuring constant multiplicative gap. When those conditions are not met, we show that $f_r(\mathbf{x}^{(k)})$ must decrease multiplicatively (Lemma 4.11), which guarantees that there cannot be many such iterations. Define

$$S_+ \stackrel{\text{def}}{=} \left\{ j : (x_j^{(k)})^{\frac{\alpha}{1-\alpha}} (\mathbf{A}^T \mathbf{y}^{(k)})_j \ge 1 + \frac{1}{10(\alpha-1)} \right\},$$
$$S_- \stackrel{\text{def}}{=} \left\{ j : (x_j^{(k)})^{\frac{\alpha}{1-\alpha}} (\mathbf{A}^T \mathbf{y}^{(k)})_j \le 1 - \frac{1}{10} \right\}.$$

The next lemma gives sufficient conditions for $\mathbf{x}^{(k)}$ to have a constant relative error.

LEMMA 4.10. *After the initial $O(\frac{1}{\beta})$ iterations, if all following conditions hold,*

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

then $f_r(\mathbf{x}^{(k)}) + f_\alpha(\mathbf{x}^*_\alpha) \le -2 f_\alpha(\mathbf{x}^*_\alpha)$.

Proof. Denote $\Delta_j = (x_j^{(k)})^{\frac{\alpha}{1-\alpha}} (\mathbf{A}^T \mathbf{y}^{(k)})_j$. Let us start by bounding the true duality gap (using feasibility from Proposition 3.2 and approximate complementary slackness from Proposition 4.9):

$$\begin{aligned}
\frac{\langle \mathbb{1}, \mathbf{x}^{(k)} \rangle}{\alpha - 1} + f_\alpha(\mathbf{x}^*_\alpha) &\le \frac{\langle \mathbb{1}, \mathbf{x}^{(k)} \rangle}{\alpha - 1} - g(\mathbf{y}^{(k)}) \\
&\le \frac{\langle \mathbb{1}, \mathbf{x}^{(k)} \rangle}{\alpha - 1} + (1+\epsilon) \left\langle \mathbf{A}^T \mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle - \frac{\alpha}{\alpha - 1} \sum_{j=1}^n (\mathbf{A}^T \mathbf{y}^{(k)})_j^{\frac{\alpha-1}{\alpha}} \\
&= \frac{1}{\alpha - 1} \sum_{j=1}^n x_j^{(k)} \left(1 + (\alpha - 1)\Delta_j - \alpha \Delta_j^{\frac{\alpha-1}{\alpha}} \right) + \epsilon \left\langle \mathbf{A}^T \mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle \\
&= \sum_{j=1}^n \xi_j + \epsilon \left\langle \mathbf{A}^T \mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle, \qquad (4.18)
\end{aligned}$$

where $\xi_j = \frac{x_j^{(k)} (1 + (\alpha-1)\Delta_j - \alpha \Delta_j^{(\alpha-1)/\alpha})}{\alpha - 1}$. To bound the expression from (4.18), we will split the sum $\sum_{j=1}^n \xi_j$ into two: corresponding to terms with $\Delta_j \ge 1$ and corresponding to terms with $\Delta_j < 1$. For the former, as $\Delta_j^{\frac{\alpha-1}{\alpha}} \ge 1$, we have

$$\begin{aligned}
\sum_{j : \Delta_j \ge 1} \xi_j &\le \frac{1}{\alpha - 1} \sum_{j : \Delta_j \ge 1} x_j^{(k)} \left(1 + (\alpha - 1)\Delta_j - \alpha \right) \\
&= \sum_{j : 1 \le \Delta_j \le 1 + \frac{1}{10(\alpha-1)}} x_j^{(k)} (\Delta_j - 1) + \sum_{j \in S_+} x_j^{(k)} (\Delta_j - 1) \\
&\le \frac{1}{5(\alpha - 1)} \left\langle \mathbb{1}, \mathbf{x}^{(k)} \right\rangle, \qquad (4.19)
\end{aligned}$$

where the last inequality is by $\sum_{j : 1 \le \Delta_j \le 1 + \frac{1}{10(\alpha-1)}} x_j^{(k)} (\Delta_j - 1) \le \frac{\langle \mathbb{1}, \mathbf{x}^{(k)} \rangle}{10(\alpha-1)}$ and the first condition from the statement of the lemma.

Consider now the terms with $\Delta_j < 1$, As $\Delta_j^{\frac{\alpha-1}{\alpha}} \geq \Delta_j$,

$$\begin{aligned}\sum_{j:\Delta_j<1} \xi_j &\leq \frac{1}{\alpha-1} \sum_{j:\Delta_j<1} x_j^{(k)}\big(1+(\alpha-1)\Delta_j - \alpha\Delta_j\big) \\ &= \frac{1}{\alpha-1} \sum_{j:1-\frac{1}{10}<\Delta_j<1} x_j^{(k)}(1-\Delta_j) + \frac{1}{\alpha-1}\sum_{j\in S_-} x_j^{(k)}(1-\Delta_j) \\ &\leq \frac{1}{5(\alpha-1)}\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle. \end{aligned} \tag{4.20}$$

The third condition from the statement of the lemma guarantees that

$$\epsilon\left\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}}\right\rangle \leq 2\epsilon\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle \leq \frac{1}{5(\alpha-1)}\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle,$$

as $\epsilon \leq \frac{1}{10(\alpha-1)}$. Hence, combining (4.18)–(4.20), $\frac{\langle \mathbb{1}, \mathbf{x}^{(k)}\rangle}{\alpha-1} + f_\alpha(\mathbf{x}_\alpha^*) \leq \frac{3}{5(\alpha-1)}\langle \mathbb{1}, \mathbf{x}^{(k)}\rangle$. Equivalently, $\frac{\langle \mathbb{1}, \mathbf{x}^{(k)}\rangle}{\alpha-1} \leq -\frac{5}{2} f_\alpha(\mathbf{x}_\alpha^*)$. Using (4.16), $\mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \leq 1$ (by feasibility—Proposition 3.2), and the third condition in the lemma statement,

$$\begin{aligned} f_r(\mathbf{x}^{(k)}) &= \frac{\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle}{\alpha-1} + \frac{\beta}{1-\beta}\left\langle \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}}, \mathbf{y}^{(k)}\right\rangle \\ &\leq \frac{\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle}{\alpha-1}\Big(1+\frac{2\beta(\alpha-1)}{1+\beta}\Big) \leq \frac{\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle}{\alpha-1}\Big(1+\frac{\epsilon(\alpha-1)}{2}\Big) \leq \frac{21}{20}\frac{\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle}{\alpha-1}, \end{aligned}$$

as $\beta \leq \epsilon/4$ and $\epsilon \leq \frac{1}{10(\alpha-1)}$. Putting everything together,

$$f_r(\mathbf{x}^{(k)}) + f_\alpha(\mathbf{x}_\alpha^*) \leq \frac{13}{20(\alpha-1)}\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle \leq -\frac{5}{2}\cdot\frac{13}{20} f_\alpha(\mathbf{x}_\alpha^*) \leq -2 f_\alpha(\mathbf{x}_\alpha^*). \qquad \square$$

LEMMA 4.11. *If in iteration k any of the conditions from Lemma 4.10 does not hold, then $f_r(\mathbf{x}^{(k)})$ must decrease by a factor at most*

$$\max\Big\{1-\theta(\beta(\alpha-1)),\ 1-\theta(\beta)\min\Big\{\frac{1}{10(\alpha-1)}, 1\Big\}\Big\}.$$

Proof. If the conditions from Lemma 4.10 do not hold, then we must have (at least) one of the following cases.

Case 1: $-\sum_{j\in S_+} x_j^{(k)}(1-(x_j^{(k)})^{\frac{\alpha}{1-\alpha}}(\mathbf{A}^T\mathbf{y}^{(k)})_j) > \frac{\langle \mathbb{1}, \mathbf{x}^{(k)}\rangle}{10(\alpha-1)}$. Observe that, by the definition of S_+, for all $j \in S_+$, $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \geq \min\{\frac{1}{10(\alpha-1)}, 1\}$ and $(1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)}) \geq \frac{1}{10(\alpha-1)} > 0$. From Lemma 3.1, as $x_j^{(k)}\nabla_j f(\mathbf{x}^{(k)})\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \geq 0\ \forall j$,

$$\begin{aligned} &f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^{(k)}) \\ &\quad\leq -\frac{\beta(1-\alpha)}{8(1+\alpha\beta)}\sum_{j\in S_+} x_j^{(k)}\nabla_j f(\mathbf{x}^{(k)})\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \\ &\quad\leq \min\Big\{\frac{1}{10(\alpha-1)}, 1\Big\}\frac{\beta}{8(1+\alpha\beta)}\sum_{j\in S_+} x_j^{(k)}\Big(1-(x_j^{(k)})^{\frac{\alpha}{1-\alpha}}(\mathbf{A}^T\mathbf{y}^{(k)})_j\Big) \\ &\quad\leq -\min\Big\{\frac{1}{10(\alpha-1)}, 1\Big\}\frac{\beta}{80(\alpha-1)(1+\alpha\beta)}\left\langle \mathbb{1}, \mathbf{x}^{(k)}\right\rangle. \end{aligned}$$

Assume that $\langle \mathbf{y}^{(k)}, \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle \leq 2\langle \mathbb{1}, \mathbf{x}^{(k)} \rangle$ (otherwise we would have Case 3 below). Then $f_r(\mathbf{x}^{(k)}) \leq (\frac{1}{\alpha-1} + \frac{2\beta}{1+\beta})\langle \mathbb{1}, \mathbf{x}^{(k)} \rangle$ and, hence,

$$\left\langle \mathbb{1}, \mathbf{x}^{(k)} \right\rangle \geq \left(\frac{1}{\alpha-1} + \frac{2\beta}{1+\beta} \right)^{-1} f_r(\mathbf{x}^{(k)}) \geq \frac{\alpha-1}{2} f_r(\mathbf{x}^{(k)}).$$

Therefore, it follows that $f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^{(k)}) \leq -\theta(\beta \min\{\frac{1}{10(\alpha-1)}, 1\}) f_r(\mathbf{x}^{(k)})$.

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

$$\begin{aligned} f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^{(k)}) &\leq -\frac{\beta(1-\alpha)}{8(1+\alpha\beta)} \sum_{j \in S_-} x_j^{(k)} \nabla_j f(\mathbf{x}^{(k)}) \overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \\ &\leq -\frac{\beta}{80(1+\alpha\beta)} \sum_{j \in S_-} x_j^{(k)} \left(1 - (x_j^{(k)})^{\frac{\alpha}{1-\alpha}} (\mathbf{A}^T \mathbf{y}^{(k)})_j \right) \\ &< -\frac{\beta}{800(1+\alpha\beta)} \left\langle \mathbb{1}, \mathbf{x}^{(k)} \right\rangle. \end{aligned}$$

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

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

$$\sum_{j=1}^{n} x_j^{(k)} \big(-1 + (x_j^{(k)})^{\frac{\alpha}{1-\alpha}} (\mathbf{A}^T \mathbf{y}^{(k)})_j \big) \geq \frac{1}{2} \left\langle \mathbf{y}^{(k)}, \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle. \tag{4.21}$$

Let $\zeta_j = -1 + (x_j^{(k)})^{\frac{\alpha}{1-\alpha}} (\mathbf{A}^T \mathbf{y}^{(k)})_j$. Then

$$\begin{aligned} \sum_{j=1}^{n} x_j^{(k)} \big(-1 + (x_j^{(k)})^{\frac{\alpha}{1-\alpha}} (\mathbf{A}^T \mathbf{y}^{(k)})_j \big) &\leq \frac{1}{2} \left\langle \mathbb{1}, \mathbf{x}^{(k)} \right\rangle + \sum_{j : \zeta_j > 1/2} x_j^{(k)} \zeta_j \\ &\leq \frac{1}{4} \left\langle \mathbf{y}^{(k)}, \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle + \sum_{j : \zeta_j > 1/2} x_j^{(k)} \zeta_j. \end{aligned} \tag{4.22}$$

As $f_r(\mathbf{x}^{(k)}) \leq (\frac{1}{2(\alpha-1)} + \frac{\beta}{1+\beta}) \langle \mathbf{y}^{(k)}, \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle$, combining (4.21) and (4.22),

$$\sum_{j : \zeta_j > 1/2} x_j^{(k)} \zeta_j \geq \frac{1}{4} \left\langle \mathbf{y}^{(k)}, \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle \geq \frac{1}{4} \left(\frac{1}{2(\alpha-1)} + \frac{\beta}{1+\beta} \right)^{-1} f_r(\mathbf{x}^{(k)}). \tag{4.23}$$

Using Lemma 3.1, it follows that, $f_r(\mathbf{x}^{(k+1)}) - f_r(\mathbf{x}^{(k)}) \leq -\frac{\beta}{16(1+\alpha\beta)} \sum_{j : \zeta_j > 1/2} x_j^{(k)} \zeta_j$, which, combined with (4.23), gives $f_r(\mathbf{x}^{(k+1)}) \leq (1 - \theta(\beta(\alpha-1))) f_r(\mathbf{x}^{(k)})$. ∎

4.3.2. The decrease in the duality gap and the convergence bound. Using Lemma 4.11, within the first $O(\frac{1}{\beta} + \frac{1}{\beta} \max\{\frac{1}{\alpha-1}, \alpha-1\} \log(\frac{f_r(\mathbf{x}^{(0)})}{f_r(\mathbf{x}_r^*)}))$ iterations, there must exist at least one iteration in which the conditions from Proposition 4.9 and Lemma 4.10 hold. With the (slight) abuse of notation, we treat the first such

an iteration as our initial ($k = 0$) iteration, and focus on proving the convergence over a subsequence of iterations that come after it. We call the iterations over which we perform the gap analysis the "gap iterations" and we define them as iterations in which

$$\left\langle \mathbf{y}^{(k)}, \mathbf{A}(\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle \leq 2\left\langle \mathbb{1}, \mathbf{x}^{(k)} \right\rangle. \tag{4.24}$$

Due to Lemma 4.11, in nongap iterations, $f_r(\mathbf{x}^{(k)})$ must decrease multiplicatively. Hence, we focus only on the gap iterations, which we index by k below.

To construct G_k, we define the upper bound to be $U_k = f_r(\mathbf{x}^{(k+1)})$. The lower bound is simply defined through the use of the Lagrangian dual as $L_k = \frac{\sum_{\ell=0}^{k} h_\ell g(\mathbf{y}^{(\ell)})}{H_k}$.

Initial gap. Using Lemma 3.1, $U_0 = f_r(\mathbf{x}^{(1)}) \leq f_r(\mathbf{x}^{(0)})$. Thus, by Lemma 4.10 and the choice of the initial point $k = 0$ described above, we have

$$G_0 = U_0 - L_0 \leq -2f_\alpha(\mathbf{x}^*_\alpha). \tag{4.25}$$

The gap decrease. The next step is to show that, for a suitably chosen sequence $\{h_k\}_k$, $H_kG_k - H_{k-1}G_{k-1} \leq O(\epsilon)(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha)$. This would immediately imply $G_k \leq \frac{H_0G_0}{H_k} + O(\epsilon)(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha)$ which is $= O(\epsilon)(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha)$ when $H_0/H_k = O(\epsilon(\alpha-1))$, due to the bound on the initial gap (4.25). As $U_k = f_r(\mathbf{x}^{(k+1)}) \geq \frac{\langle \mathbb{1}, \mathbf{x}^{(k+1)} \rangle}{\alpha - 1}$ and $L_k \geq -f_\alpha(\mathbf{x}^*_\alpha)$, taking $\hat{\mathbf{x}}^{(k)} = (\mathbf{x}^{(k+1)})^{\frac{1}{1-\alpha}}$, it would immediately follow that

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

Since $\hat{\mathbf{x}}^{(k)}$ is (P-a)-feasible (Proposition 3.2), $\hat{\mathbf{x}}^{(k)}$ is an $O(\epsilon)$-approximate solution to (P-a).

To bound $H_kG_k - H_{k-1}G_{k-1}$, we will need the following technical proposition that bounds $H_kL_k - H_{k-1}L_{k-1}$ (the change in the lower bound).

PROPOSITION 4.12. *For any two consecutive gap iterations $k-1, k$,*

$$\begin{aligned} &H_kL_k - H_{k-1}L_{k-1} \\ &\quad \geq h_k \Bigg[f_r(\mathbf{x}^{(k)}) - \left\langle \nabla f_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} \right\rangle - 8\epsilon(\alpha-1)f_\alpha(\mathbf{x}^*_\alpha) \\ &\qquad + \frac{\alpha}{\alpha-1} \sum_{j=1}^{n} x_j^{(k)} \left(\left(1 + \overline{\nabla_j f}_r(\mathbf{x}^{(k)})\right)^{\frac{\alpha-1}{\alpha}} - \left(1 + (1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)})\right) \right) \Bigg]. \end{aligned}$$

Proof. By the definition of the lower bound,

$$H_kL_k - H_{k-1}L_{k-1} = h_k g(\mathbf{y}^{(k)}) = h_k \left(-\left\langle \mathbb{1}, \mathbf{y}^{(k)} \right\rangle + \frac{\alpha}{\alpha-1} \sum_{j=1}^{n} (\mathbf{A}^T\mathbf{y}^{(k)})_j^{\frac{\alpha-1}{\alpha}} \right). \tag{4.26}$$

From Proposition 4.9, $\langle \mathbb{1}, \mathbf{y}^{(k)} \rangle \leq (1+\epsilon)\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle$, while from (4.16),

$$f_r(\mathbf{x}^{(k)}) - \left\langle \nabla f_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} \right\rangle = \left(\frac{\beta}{1+\beta} + \frac{1}{\alpha-1} \right) \left\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \right\rangle.$$

Hence,

$$\begin{aligned}\langle \mathbb{1}, \mathbf{y}^{(k)} \rangle &\le (1+\epsilon)\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle \\ &= -f_r(\mathbf{x}^{(k)}) + \langle \nabla f_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} \rangle + \frac{\alpha}{\alpha-1}\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle \\ &\quad + \left(\epsilon + \frac{\beta}{1+\beta}\right)\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle.\end{aligned}$$

Since k is a gap iteration, $f_r(\mathbf{x}^{(k)}) \ge (\frac{1}{2(\alpha-1)} + \frac{\beta}{1+\beta})\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle$. Hence,

$$\begin{aligned}\langle \mathbb{1}, \mathbf{y}^{(k)} \rangle &\le - f_r(\mathbf{x}^{(k)}) + \langle \nabla f_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} \rangle \\ &\quad + \frac{\alpha}{\alpha-1}\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle + \frac{10}{4}(\alpha-1)\epsilon f_r(\mathbf{x}^{(k)}) \\ &\le - f_r(\mathbf{x}^{(k)}) + \langle \nabla f_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} \rangle \\ &\quad + \frac{\alpha}{\alpha-1}\langle \mathbf{A}^T\mathbf{y}^{(k)}, (\mathbf{x}^{(k)})^{\frac{1}{1-\alpha}} \rangle - 8(\alpha-1)\epsilon f_\alpha(\mathbf{x}^*_\alpha), \qquad (4.27)\end{aligned}$$

where the last inequality follows from $f_r(\mathbf{x}^{(k)}) \le f_r(\mathbf{x}^{(0)})$ (as $f_r(\cdot)$ decreases in each iteration) and $f(\mathbf{x}^{(0)}) \le -\frac{11}{4} f_\alpha(\mathbf{x}^*_\alpha)$ (by the choice of $\mathbf{x}^{(0)}$ and Lemma 4.10). Combining (4.26) and (4.27),

$$\begin{aligned}H_k L_k - H_{k-1}L_{k-1} \ge h_k \Bigg(& f_r(\mathbf{x}^{(k)}) - \langle \nabla f_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} \rangle + 8\epsilon(\alpha-1) f_\alpha(\mathbf{x}^*_\alpha) \\ & + \frac{\alpha}{\alpha-1}\sum_{j=1}^{n}\left((\mathbf{A}^T\mathbf{y}^{(k)})_j^{\frac{\alpha-1}{\alpha}} - (x_j^{(k)})^{\frac{1}{1-\alpha}}(\mathbf{A}^T\mathbf{y}^{(k)})_j \right)\Bigg).\end{aligned}$$

Finally, as $(\mathbf{A}^T\mathbf{y}^{(k)})_j = (x_j^{(k)})^{\frac{-\alpha}{1-\alpha}}(1 + (1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)}))$ and $(1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)}) \ge \overline{\nabla_j f}_r(\mathbf{x}^{(k)})$, the statement of the proposition follows. ∎

LEMMA 4.13. *If, for $k \ge 1$, $\frac{h_k}{H_k} \le \frac{\beta \min\{\alpha-1,1\}}{16(1+\alpha\beta)}$, then*

$$H_k G_k - H_{k-1}G_{k-1} \le -8h_k\epsilon(\alpha-1)f_\alpha(\mathbf{x}^*_\alpha).$$

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

$$H_k U_k - H_{k-1}U_{k-1} \le h_k f_r(\mathbf{x}^{(k)}) - H_k \frac{\beta(1-\alpha)}{8(1+\alpha\beta)}\sum_{j=1}^{n} x_j^{(k)}\nabla_j f_r(\mathbf{x}^{(k)})\overline{\nabla_j f}_r(\mathbf{x}^{(k)}).$$

Combining this with the change in the lower bound from Proposition 4.12, it follows that to prove the statement of the lemma it suffices to show that, $\forall j$,

$$\begin{aligned}\xi_j \stackrel{\text{def}}{=} h_k & \left[\nabla_j f_r(\mathbf{x}^{(k)}) - \frac{\alpha}{\alpha-1}\left(\left(1+\overline{\nabla_j f}_r(\mathbf{x}^{(k)})\right)^{\frac{\alpha-1}{\alpha}} - \left(1+(1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)})\right)\right)\right] \\ & - H_k\frac{\beta(1-\alpha)}{8(1+\alpha\beta)}\nabla_j f_r(\mathbf{x}^{(k)})\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) \le 0.\end{aligned}$$

Consider the following three cases:

Case 1: $(1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)}) \in [-1/2, 1]$. Then $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) = (1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)})$. A simple corollary of Taylor's theorem is that, in this setting,

$$\left(1+\overline{\nabla_j f}_r(\mathbf{x}^{(k)})\right)^{\frac{\alpha-1}{\alpha}} \geq 1 + \frac{\alpha-1}{\alpha}\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) - \frac{\alpha-1}{\alpha^2}(\overline{\nabla_j f}_r(\mathbf{x}^{(k)}))^2. \tag{4.28}$$

Using (4.28) from above

$$\begin{aligned}\xi_j &\leq h_k\left[\frac{\overline{\nabla_j f}_r(\mathbf{x}^{(k)})}{1-\alpha} - \frac{\alpha}{\alpha-1}\left(-\frac{1}{\alpha}\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) - \frac{\alpha-1}{\alpha^2}\left(\overline{\nabla_j f}_r(\mathbf{x}^{(k)})\right)^2\right)\right] \\ &\quad - \frac{H_k\beta}{8(1+\alpha\beta)}(\overline{\nabla_j f}_r(\mathbf{x}^{(k)}))^2 \\ &= (\overline{\nabla_j f}_r(\mathbf{x}^{(k)}))^2\left(\frac{h_k}{\alpha} - \frac{H_k\beta}{8(1+\alpha\beta)}\right).\end{aligned}$$

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

Case 2: $(1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)}) \in [-1, -1/2)$. Then $\overline{\nabla_j f}_r(\mathbf{x}^{(k)}) = (1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)})$ and $|\overline{\nabla_j f}_r(\mathbf{x}^{(k)})| > \frac{1}{2}$. As in this case $(1+\overline{\nabla_j f}_r(\mathbf{x}^{(k)}))^{\frac{\alpha-1}{\alpha}} \geq 1 + \overline{\nabla_j f}_r(\mathbf{x}^{(k)})$, we have

$$\xi_j \leq \nabla_j f_r(\mathbf{x}^{(k)})\left(h_k - H_k\frac{\beta(\alpha-1)}{16(1+\alpha\beta)}\right),$$

which is ≤ 0, as $\frac{h_k}{H_k} \leq \frac{\beta\min\{\alpha-1,1\}}{16(1+\alpha\beta)}$ and $\nabla_j f_r(\mathbf{x}^{(k)}) > 0$.

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

$$\begin{aligned}\xi_j &\leq h_k\left[\nabla_j f_r(\mathbf{x}^{(k)}) - \frac{\alpha}{\alpha-1}\left(2^{\frac{\alpha-1}{a}} - 1 - (1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)})\right)\right] \\ &\quad - H_k\frac{\beta(1-\alpha)}{8(1+\alpha\beta)}\nabla_j f_r(\mathbf{x}^{(k)}) \\ &\leq (1-\alpha)\nabla_j f_r(\mathbf{x}^{(k)})\left(h_k - \frac{H_k\beta}{8(1+\alpha\beta)}\right),\end{aligned}$$

which is nonpositive, as $\frac{h_k}{H_k} \leq \frac{\beta}{8(1+\alpha\beta)}$. ∎

We can now state the final convergence bound.

THEOREM 4.14. *Given* $\epsilon \in (0, \min\{1/2, 1/(10(\alpha-1))\}]$, *after at most*

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

iterations of FAIRPACKING,

$$f_\alpha(\mathbf{x}_\alpha^{(k+1)}) - f_\alpha(\mathbf{x}_\alpha^*) \geq 10\epsilon(\alpha-1)f_\alpha(\mathbf{x}_\alpha^*),$$

where $\mathbf{x}_\alpha^{(k+1)} = (\mathbf{x}^{(k+1)})^{\frac{1}{1-\alpha}}$.

Proof. At initialization, $f_r(\cdot)$ takes a value less than $\frac{n(3n\rho)^{\alpha-1}}{\alpha-1}$ and decreases in every subsequent iteration. From Proposition 2.2, $-f_\alpha(\mathbf{x}_\alpha^*) \geq \frac{n}{\alpha-1}$. As $f_r(\mathbf{x}) \geq \frac{\langle \mathbb{1}, \mathbf{x}\rangle}{\alpha-1}$ and the algorithm always maintains solutions $\mathbf{x}^{(k)}$ that are feasible in (P-b), $\min_k f_r(\mathbf{x}^{(k)}) \geq -f_\alpha(\mathbf{x}_\alpha^*) \geq \frac{n}{\alpha-1}$. Using Proposition 4.9 and Lemma 4.11, there are at most $O(\frac{1}{\beta}\max\{\frac{1}{\alpha-1}, \alpha-1\}(\alpha-1)\log(n\rho)) = O(\frac{(1+\alpha)\max\{(\alpha-1)^2,1\}\log(n\rho)\log(mn\rho/\epsilon)}{\epsilon})$

nongap iterations before $f_r(\cdot)$ reaches its minimum value. Using the second part of Proposition 2.3, if this happens, it follows that $f_\alpha(\hat{\mathbf{x}}^{(k+1)}) - f_\alpha(\mathbf{x}^*_\alpha) \geq -2\epsilon(1-\alpha)f_\alpha(\mathbf{x}^*)$, 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 4.13,

$$G_k \leq \frac{H_0 G_0}{H_k} - 8\epsilon(\alpha-1)f_\alpha(\mathbf{x}^*_\alpha) = \left(1 - \frac{\beta\min\{\alpha-1,1\}}{16(1+\alpha\beta)}\right)^k G_0 - 8\epsilon(\alpha-1)f_\alpha(\mathbf{x}^*_\alpha).$$

As $G_0 \leq -2f_\alpha(\mathbf{x}^*_\alpha)$, after $k \geq \frac{\log(\frac{1}{\epsilon(\alpha-1)})}{\beta\min\{\alpha-1,1\}}16(1+\alpha\beta) = O(\frac{(1+\alpha)\log(\frac{1}{\epsilon(\alpha-1)})\log(mn\rho/\epsilon)}{\epsilon\min\{\alpha-1,1\}})$ iterations, it must be $-f_\alpha(\hat{\mathbf{x}}^{(k+1)}) + f_\alpha(\mathbf{x}^*_\alpha) \leq G_k \leq 10\epsilon(1-\alpha)f_\alpha(\mathbf{x}^*_\alpha)$, as claimed. ∎

5. Fair covering. In this section, we show how to reduce the fair covering problem to the $\alpha < 1$ case from section 4.1. We will be assuming throughout that $\beta \geq \frac{\epsilon/4}{\log(mn\rho/\epsilon)}$, as otherwise the problem can be reduced to the linear covering (see, e.g., [12]). Note that the only aspect of the analysis that relies on β being sufficiently small in the $\alpha \in [0,1)$ case is to ensure that f_r closely approximates $-f_\alpha$ around the optimum of (P-a), (P-b). Here, we will need to choose β' to be sufficiently small to ensure that the lower bound from the $\alpha < 1$ case closely approximates $-g_\beta$ around the optimum $\mathbf{y}^*$. Since we do not need to ensure the feasibility of the packing problem, in this section we take $C = 1$, so that $f_r(\mathbf{x}) = -\langle \mathbb{1}, \mathbf{x}\rangle + \frac{\beta}{1+\beta}\sum_{i=1}^m (\mathbf{A}\mathbf{x})_i^{(1+\beta)/\beta}$. As before, the upper bound is defined as $U_k = f_r(\mathbf{x}^{(k+1)})$. The lower bound L_k is the same as the one from section 4.1, with the choice of β' as in Algorithm 5.1 (FairCovering).

Algorithm 5.1 FairCovering$(\mathbf{A}, \epsilon, \beta)$.

If $\beta \leq 0$, set $\beta = \frac{\epsilon/4}{\log(mn\rho/\epsilon)}$. Initialize: $\mathbf{x}_j^{(0)} = \frac{1}{n\rho}\left(\frac{1}{m\rho}\right)^\beta \mathbb{1}$, $\mathbf{y}_\beta^{(0)} = \mathbf{0}$.
$\mathbf{z}^{(0)} = \exp(\epsilon/4)\mathbb{1}$, $\beta' = \frac{\epsilon/4}{(1+\beta)\log(mn\rho/\epsilon)}$, $h = \frac{\beta\beta'}{16\epsilon}$
for $k = 1$ to $K = 1 + \lceil 2/(h\epsilon)\rceil$ **do**
  $\mathbf{x}^{(k)} = \left(\mathbb{1} + \mathbf{z}^{(k-1)}\right)^{-1/\beta'}$
  $\mathbf{z}^{(k)} = \mathbf{z}^{(k-1)} + \epsilon h \overline{\nabla f}_r(\mathbf{x}^{(k)})$
  $\mathbf{y}_\beta^{(k)} = \frac{k-1}{k}\mathbf{y}_\beta^{(k-1)} + (\mathbf{A}\mathbf{x}^{(k)})^{1/\beta}/k$
return $(1+\epsilon)\mathbf{y}_\beta^{(K)}$

We start by bounding the initial gap.

Proposition 5.1. *Let $h_0 = H_0 = 1$. Then: $H_0 G_0 \leq 2(1+\beta)g_\beta(\mathbf{y}^*_\beta)$.*

Proof. By the same arguments as in the proof of Proposition 4.1, $\mathbf{x}^{(1)} = \mathbf{x}^{(0)}$ and, hence, $U_0 = f_r(\mathbf{x}^{(0)})$. Let $\mathbf{x}^*_\beta = \operatorname{argmin}_{\mathbf{x}\geq\mathbf{0}} f_r(\mathbf{x})$. Then, the initial gap can be expressed as $H_0 G_0 = \left\langle \nabla f_r(\mathbf{x}^{(0)}), \mathbf{x}^{(0)}\right\rangle - \widehat{\psi}(\mathbf{z}^{(0)}) + \phi(\mathbf{x}^*_\beta)$.

By the choice of $\mathbf{x}^{(0)}$, $(\mathbf{A}\mathbf{x}^{(0)})^{1/\beta} \leq \mathbb{1}$ and, therefore, $\nabla f_r(\mathbf{x}^{(0)}) \leq \mathbf{0}$. Thus, $H_0 G_0 \leq -\widehat{\psi}(\mathbf{z}^{(0)}) + \phi(\mathbf{x}^*_\beta)$. As β' chosen here is smaller than the one from section 4.1, it follows by the same argument as in the proof of Proposition 4.2 that $-\widehat{\psi}(\mathbf{z}^{(0)}) = \frac{\beta'}{\epsilon(1-\beta')}\langle \mathbb{1}, (\mathbf{x}^{(0)})\rangle^{1-\beta'} \leq \frac{1}{2}\langle \mathbb{1}, \mathbf{x}^{(0)}\rangle$, which is at most $\frac{1}{2}(1+\beta)g_\beta(\mathbf{y}^*_\beta)$ by the choice of the initial point $\mathbf{x}^{(0)}$ and Proposition 2.4. It remains to bound $\phi(\mathbf{x}^*_\beta) = \psi(\mathbf{x}^*_\beta) - \langle \nabla\psi(\mathbf{x}^{(0)}) + \overline{\nabla f}_r(\mathbf{x}^{(0)}), \mathbf{x}^*_\beta\rangle$. By the definition of ψ, $\psi(\mathbf{x}^*_\beta) \leq 0$ and $\nabla_j\psi(\mathbf{x}^{(0)}) =$

$\frac{1}{\epsilon}(1-(x_j^{(0)})^{-\beta'}) \geq -1/2$. By the definition of f_r, $\overline{\nabla f}_r(\mathbf{x}^{(0)}) \geq -\mathbb{1}$. Hence,

$$-\left\langle \nabla\psi(\mathbf{x}^{(0)}) + \overline{\nabla f}_r(\mathbf{x}^{(0)}), \mathbf{x}_\beta^* \right\rangle \leq \frac{3}{2}\left\langle \mathbb{1}, \mathbf{x}_\beta^* \right\rangle \leq \frac{3}{2}(1+\beta)g_\beta(\mathbf{y}_\beta^*),$$

where the last inequality is by Proposition 2.5. ∎

Since the analysis from section 4.1 can be applied in a straightforward way to ensure that after $\lceil 2/(h\epsilon)\rceil$ iterations we have $H_k G_k \leq \epsilon(1+\beta)g_\beta(\mathbf{y}_\beta^*)$, what remains to show is that we can recover an approximate solution to (C) from from this analysis. Define

$$\mathbf{y}^{(k)} = (\mathbf{A}\mathbf{x}^{(k)})^{1/\beta} \quad \text{and} \quad \mathbf{y}_\beta^{(k)} = \frac{\sum_{\ell=1}^k \mathbf{y}^{(\ell)}}{k}. \tag{5.1}$$

Notice that this is consistent with the definition of $\mathbf{y}_\beta^{(k)}$ from FairCovering. We are now ready to state and prove the main result from this section.

Theorem 5.2. *The solution* $\mathbf{y}_\beta^{(k)}$ *produced by* FairCovering *after* $K = 1 + \lceil 2/(h\epsilon)\rceil = O(\frac{(1+\beta)\log(mn\rho)}{\beta\epsilon})$ *iterations satisfies*

$$\mathbf{A}^T\mathbf{y}_\beta^{(K)} \geq (1-\epsilon/2)\mathbb{1} \quad \textit{and} \quad g_\beta(\mathbf{y}_\beta^{(k)}) - g_\beta(\mathbf{y}_\beta^*) \leq 3\epsilon(1+\beta)g_\beta(\mathbf{y}_\beta^*).$$

Proof. By the definition of f_r and $\mathbf{y}^{(k)}$, we have $f_r(\mathbf{x}^{(k)}) - \langle \nabla f_r(\mathbf{x}^{(k)}), \mathbf{x}^{(k)}\rangle = -\sum_{i=1}^m \frac{(y_i^{(k)})^{1+\beta}}{1+\beta} = -g_\beta(\mathbf{y}^{(k)})$. Hence,

$$L_k = \frac{-\sum_{\ell=0}^k h_\ell g_\beta(\mathbf{y}^{(\ell)}) + \widehat{\psi}(\mathbf{z}^{(k)}) - \phi(\mathbf{x}_\beta^*)}{H_k}.$$

As, by Proposition 5.1 and the analysis from section 4.1, it must be $G_K \leq 2\epsilon(1+\beta)g_\beta(\mathbf{y}_\beta^*)$, and by Lagrangian duality $U_K = f_r(\mathbf{x}^{(K+1)}) \geq -g_\beta(\mathbf{y}_\beta^*)$, we have that $L_K = U_K - G_K \geq -(1+2\epsilon(1+\beta))g_\beta(\mathbf{y}_\beta^*)$. As $\widehat{\psi}(\mathbf{z}^{(k)}) \leq 0$ and $\phi(\mathbf{x}_\beta^*) \geq \psi(\mathbf{x}_\beta^*) \geq -\frac{1}{2}\langle \mathbb{1}, \mathbf{x}_\beta^*\rangle = -\frac{1}{2}(1+\beta)g_\beta(\mathbf{y}_\beta^*)$ (because $\nabla\psi(\mathbf{x}^{(0)}) + \overline{\nabla f}_r(\mathbf{x}^{(0)}) \geq \mathbf{0}$ and, by the choice of β', $\psi(\mathbf{x}_\beta^*) \geq -\frac{1}{2}\langle \mathbb{1}, \mathbf{x}_\beta^*\rangle$), we have that

$$\begin{aligned} -\frac{\sum_{\ell=0}^K h_\ell g_\beta(\mathbf{y}^{(\ell)})}{H_K} &\geq -(1+2\epsilon(1+\beta))g_\beta(\mathbf{y}_\beta^*) - \frac{(1+\beta)g_\beta(\mathbf{y}_\beta^*)}{2H_K} \\ &\geq -(1+(9\epsilon/4)(1+\beta))g_\beta(\mathbf{y}_\beta^*). \end{aligned} \tag{5.2}$$

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_β is convex, by the definition of $\mathbf{y}_\beta^{(K)}$ and Jensen's inequality,

$$\begin{aligned} \frac{\sum_{\ell=0}^K h_\ell g_\beta(\mathbf{y}^{(\ell)})}{H_K} &= \frac{1}{H_0}g_\beta(\mathbf{y}^{(0)}) + \frac{h\sum_{\ell=1}^K g_\beta(\mathbf{y}^{(\ell)})}{1+hK} \geq \frac{hK}{1+hK}g_\beta(\mathbf{y}_\beta^{(K)}) \\ &\geq \frac{1}{1+\epsilon/2}g_\beta(\mathbf{y}_\beta^{(K)}). \end{aligned} \tag{5.3}$$

Hence, combining (5.2) and (5.3), $g_\beta(\mathbf{y}_\beta^K) \leq (1+3\epsilon(1+\beta))g_\beta(\mathbf{y}_\beta^*)$.

It remains to show that $\mathbf{y}_\beta^{(K)}$ is nearly feasible. By its definition, $\mathbf{y}_\beta^{(K)} \geq \mathbf{0}$. We claim first that it must be $\mathbf{z}^{(k)} \geq -(\epsilon/2)\mathbb{1}$. Suppose not. Then $\widehat{\psi}(\mathbf{z}^{(k)}) =$

$-\frac{\beta'}{\epsilon(1-\beta')}\sum_{j=1}^{n}(1+z_j^{(k)})^{-\frac{1-\beta'}{\beta'}} \le -\frac{\beta'}{\epsilon(1-\beta')}(1-\epsilon/2)^{-\frac{1-\beta'}{\beta'}} << -H_K(1+\beta)g_\beta(\mathbf{y}_\beta^*)$. As (from the argument above) $\phi_\beta(\mathbf{x}_\beta^*) \ge -\frac{1}{2}(1+\beta)g_\beta(\mathbf{y}_\beta^*)$, it follows that $L_K \ll -(1+\beta)g_\beta(\mathbf{y}_\beta^*)$, which is a contradiction, as we have already shown that $L_K \ge -(1+\epsilon(1+\beta))g_\beta(\mathbf{y}_\beta^*)$. Thus, we have, $\forall j$, $z_j^{(k)} \ge -\epsilon/2$. By the definition of $\mathbf{z}^{(k)}$,

$$1+z_j^{(K)} \le 1+\epsilon\sum_{\ell=1}^{K} h_\ell \overline{\nabla_j f}_r(\mathbf{x}^{(\ell)}) \le 1+\epsilon\sum_{\ell=1}^{K} h_\ell \nabla_j f_r(\mathbf{x}^{(\ell)}).$$

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

$$\mathbf{A}^T\mathbf{y}_\beta^{(K)} = \frac{\sum_{\ell=1}^{K}\mathbf{A}^T\mathbf{y}^{(\ell)}}{K} \ge \mathbf{z}^{(K)} + \epsilon h K \mathbb{1} \ge (1-\epsilon/2)\mathbb{1}. \qquad \square$$

Observe that Algorithm 5.1 returns the point $(1+\epsilon)\mathbf{y}^{(K)}$. This is to ensure that all of the covering constraints are satisfied. The approximation error in the statement of Theorem 5.2 is then affected only by a factor $(1+\epsilon)^{1+\beta}$.

6. Conclusion. We presented efficient distributed algorithms for solving the class of α-fair packing and covering problems on a relative scale. This class of problems 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, 12, 24]. Our results greatly improve upon the only known width-independent solver for the general α-fair packing [25], both in terms of simplicity of the convergence analysis and in terms of the resulting convergence time.

Appendix A. Omitted proofs.

Proof of Proposition 2.3. The proof of the first part follows by solving

$$\psi^*(\mathbf{A}F_\alpha(\mathbf{x}) - \mathbb{1}) = \max_{\mathbf{y}\ge\mathbf{0}}\left\{\langle \mathbf{A}F_\alpha(\mathbf{x}), \mathbf{y}\rangle - \frac{1}{C^\beta}\sum_{i=1}^{m}\frac{y_i^{1+\beta}}{1+\beta}\right\},$$

which is solved for $y_i = C(\mathbf{A}F_\alpha(\mathbf{x}))_i^{1/\beta}$.

Let $\mathbf{x} = F_\alpha^{-1}((1-\epsilon)\mathbf{x}_\alpha^*)$. Then, $\forall i$, $(\mathbf{A}F_\alpha(\mathbf{x}))_i \le 1-\epsilon$ and thus

$$C(\mathbf{A}F_\alpha(\mathbf{x}))_i^{\frac{1+\beta}{\beta}} \le (1-\epsilon/4)^{1/\beta} \le \Big(\frac{\epsilon}{4mn\rho}\Big)^{\alpha+1}.$$

Hence,

$$C\sum_{i=1}^{m}(\mathbf{A}F_\alpha(\mathbf{x}))_i^{\frac{1+\beta}{\beta}} \le m\Big(\frac{\epsilon}{4mn\rho}\Big)^{\alpha+1} \le \Big(\frac{\epsilon}{4n\rho}\Big)^{\alpha+1}. \tag{A.1}$$

From Proposition 2.2, we have for $\alpha \ne 1$ that $(1-\alpha)f_\alpha(\mathbf{x}_\alpha^*) \ge n(n\rho)^{\alpha-1} \ge \frac{1}{\rho}$. Thus, (A.1) implies that in this case $C\sum_{i=1}^{m}(\mathbf{A}F_\alpha(\mathbf{x}))_i^{\frac{1+\beta}{\beta}} \le \frac{\epsilon}{4}(1-\alpha)f_\alpha(\mathbf{x}_\alpha^*)$. For $\alpha = 1$, we can simply use that $(\frac{\epsilon}{4n\rho})^{\alpha+1} \le \frac{\epsilon n}{16}$.

As $f_\alpha((1-\epsilon)\mathbf{x}_\alpha^*) = (1-\epsilon)^{1-\alpha}f_\alpha(\mathbf{x}_\alpha^*) \ge (1-\frac{3\epsilon(1-\alpha)}{2})f_\alpha(\mathbf{x}_\alpha^*)$ for $\alpha \ne 1$ and $f_\alpha((1-\epsilon)\mathbf{x}_\alpha^*) = n\log(1-\epsilon) + f_\alpha(\mathbf{x}_\alpha^*) \ge -\frac{3}{2}\epsilon n + f_\alpha(\mathbf{x}_\alpha^*)$, it follows that

$$f_r(\mathbf{x}_r^*) \le f_r(\mathbf{x}) = -f_\alpha((1-\epsilon)\mathbf{x}_\alpha^*) + \frac{\beta C}{1+\beta}\sum_{i=1}^{m}(\mathbf{A}F_\alpha(\mathbf{x}))_i^{\frac{1+\beta}{\beta}} \le -f_\alpha(\mathbf{x}_\alpha^*) + 2\epsilon_f.$$

Finally, the (P-b)-feasibility of $\mathbf{x}_r^*$ (and, by the change of variables, (P-a)-feasiblity of $\hat{\mathbf{x}}_r$) follows from Proposition 3.2 and Lemma 3.1. ∎

Proof of Lemma 3.1. We will only prove the first part of the lemma, as he second part uses the same ideas. Writing a Taylor approximation of $f_r(\mathbf{x}+\Gamma\mathbf{x})$, we have

$$f_r(\mathbf{x}+\Gamma\mathbf{x}) = f_r(\mathbf{x}) + \langle \nabla f_r(\mathbf{x}), \Gamma\mathbf{x}\rangle + \frac{1}{2}\left\langle \nabla^2 f_r(\mathbf{x}+t\Gamma\mathbf{x})\Gamma\mathbf{x}, \Gamma\mathbf{x}\right\rangle \tag{A.2}$$

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

$$\nabla_j f_r(\mathbf{x}) = \frac{1}{1-\alpha}\left(-1 + \sum_i A_{ij} x_j{}^{\frac{1}{1-\alpha}-1} C(\mathbf{A}F_\alpha(\mathbf{x}))_i^{1/\beta}\right), \tag{A.3}$$

$$\begin{aligned}\nabla^2_{jk} f_r(\mathbf{x}) &= \mathbb{1}_{\{j=k \text{ and } \alpha\neq 0\}} \frac{\alpha}{(1-\alpha)^2}\sum_i A_{ij} x_j{}^{\frac{1}{1-\alpha}-2} C(\mathbf{A}F_\alpha(\mathbf{x}))_i^{1/\beta} \\ &\quad + \frac{1/\beta}{(1-\alpha)^2}\sum_{i'} A_{i'j}A_{i'k}(x_j x_k)^{\frac{1}{1-\alpha}-1} C(\mathbf{A}F_\alpha(\mathbf{x}))_{i'}^{1/\beta-1}.\end{aligned} \tag{A.4}$$

To have 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 γ_m be the maximum (absolute) multiplicative update. Then, to have $\nabla^2_{jk} f_r(\mathbf{x}+\Gamma\mathbf{x}) \le 2\nabla^2_{jk} f_r(\mathbf{x})$, it is sufficient to enforce: (i) $(1\pm\gamma_m)^{\frac{1}{1-\alpha}-2\pm\frac{1}{\beta(1-\alpha)}} \le 2$ (from the first term in (A.4)) and (ii) $(1\pm\gamma_m)^{\frac{2}{1-\alpha}-2+\frac{1-\beta}{\beta(1-\alpha)}} \le 2$ (from the second term in (A.4)). Combining (i) and (ii), it is not hard to verify that it suffices to have: $\gamma_m \le \frac{\beta|1-\alpha|}{2(1+\alpha\beta)}$.

Assume from now on that $|\gamma_j| \le \gamma_m \le \frac{\beta|1-\alpha|}{2(1+\alpha\beta)}\ \forall j$. Then, we have

$$\begin{aligned}\frac{1}{2}\left\langle \nabla^2 f_r(\mathbf{x}+t\Gamma\mathbf{x})\Gamma\mathbf{x}, \Gamma\mathbf{x}\right\rangle &\le \sum_j \frac{\alpha}{(1-\alpha)^2}\sum_i {\gamma_j}^2 A_{ij} x_j{}^{\frac{1}{1-\alpha}} C(\mathbf{A}F_\alpha(\mathbf{x}))_i^{\frac{1}{\beta}} \\ &\quad + \frac{1/\beta}{(1-\alpha)^2}\sum_{i'} C(\mathbf{A}F_\alpha(\mathbf{x}))_{i'}^{\frac{1}{\beta}-1}(\mathbf{A}\Gamma\mathbf{x}^{\frac{1}{1-\alpha}})_{i'}^2.\end{aligned}$$

Observe that, by the Cauchy–Schwarz inequality,

$$(\mathbf{A}\Gamma\mathbf{x}^{\frac{1}{1-\alpha}})_{i'}^2 = \left(\sum_j A_{i'j} x_j{}^{\frac{1}{1-\alpha}}\gamma_j\right)^2 \le (\mathbf{A}F_\alpha(\mathbf{x}))_{i'}\sum_j A_{i'j} x_j{}^{\frac{1}{1-\alpha}}{\gamma_j}^2.$$

Therefore, applying the last inequality and changing the order of summation,

$$\frac{1}{2}\left\langle \nabla^2 f_r(\mathbf{x}+t\Gamma\mathbf{x})\Gamma\mathbf{x}, \Gamma\mathbf{x}\right\rangle \le \frac{1+\alpha\beta}{\beta(1-\alpha)^2}\sum_j {\gamma_j}^2 x_j\left((1-\alpha)\nabla_j f_r(\mathbf{x})+1\right). \tag{A.5}$$

Since $\langle \nabla f_r(\mathbf{x}), \Gamma\mathbf{x}\rangle = \sum_j \gamma_j x_j \nabla_j f_r(\mathbf{x})$ and $|\overline{\nabla_j f}_r(\mathbf{x})| \le 2|\frac{(1-\alpha)\nabla_j f_r(\mathbf{x})}{1+(1-\alpha)\nabla_j f_r(\mathbf{x})}|$, choosing $\gamma_j = -\frac{c_j}{4}\cdot\frac{\beta(1-\alpha)}{1+\alpha\beta}\overline{\nabla_j f}_r(\mathbf{x})$ and combining (A.5) and (A.2),

$$\begin{aligned}f_r(\mathbf{x}+\Gamma\mathbf{x}) - f_r(\mathbf{x}) &\le -\frac{\beta(1-\alpha)}{1+\alpha\beta}\sum_j \frac{c_j}{4}\left(1-\frac{c_j}{2}\right) x_j \nabla_j f_r(\mathbf{x})\overline{\nabla_j f}_r(\mathbf{x}) \\ &= \sum_{j=1}^n \left(1-\frac{c_j}{2}\right)\gamma_j x_j \nabla f_r(\mathbf{x}).\end{aligned}$$

∎

Acknowledgment. We thank Ken Clarkson for his useful comments and suggestions regarding the presentation of the paper.

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 Proceedings ACM-SIAM SODA'15, SIAM, Philadelphia, 2015, pp. 1439–1456.

[2] A. B. Atkinson, *On the measurement of inequality*, J. Econom. Theory, 2 (1970), pp. 244–263.

[3] B. Awerbuch, Y. Azar, and R. Khandekar, *Fast load balancing via bounded best response*, in Proceedings ACM-SIAM SODA'08, SIAM, Philadelphia, 2008, pp. 314–322.

[4] B. Awerbuch and R. Khandekar, *Stateless distributed gradient descent for positive linear programs*, SIAM J. Comput., 38 (2009), pp. 2468–2486.

[5] Y. Bartal, J. Byers, and D. Raz, *Global optimization using local information with applications to flow control*, in Proceedings IEEE FOCS'97, IEEE Computer Society, Los Alamitos, CA, 1997, pp. 303–312.

[6] A. Beck, A. Nedić, A. Ozdaglar, and M. Teboulle, *An $O(1/k)$ gradient method for network resource allocation problems*, IEEE Trans. Control Netw. Syst., 1 (2014), pp. 64–73.

[7] D. P. Bertsekas, *Control of Uncertain Systems with a Set-Membership Description of the Uncertainty*, Ph.D. thesis, MIT, Cambridge, MA, 1971.

[8] D. P. Bertsekas, *Convex Optimization Theory*, Athena Scientific, Belmont, MA, 2009.

[9] D. Bertsimas, V. F. Farias, and N. Trichakis, *The price of fairness*, Oper. Res., 59 (2011), pp. 17–31.

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

[11] T. Bonald and J. Roberts, *Multi-resource fairness: Objectives, algorithms and performance*, in Proceedings ACM SIGMETRICS'15, ACM, New York, 2015.

[12] 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*, preprint, arXiv:1710.09002, 2017.

[13] J. Diakonikolas and L. Orecchia, *The approximate duality gap technique: A unified theory of first-order methods*, SIAM J. Optim., 29 (2019), pp. 660–689.

[14] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, *Dominant resource fairness: Fair allocation of multiple resource types*, in Proceedings USENIX NSDI'11, USENIX Association, Berkeley, CA, 2011, pp. 323–336.

[15] K. Jain and V. Vazirani, *Eisenberg-Gale markets: Algorithms and structural properties*, in Proceedings ACM STOC'07, ACM, New York, 2007.

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

[17] L. Jose, S. Ibanez, M. Alizadeh, and N. McKeown, *A distributed algorithm to calculate max-min fair rates without per-flow state*, Proc. ACM Meas. Analy. Comput. Systems, 3 (2019), 21.

[18] E. Kalai and M. Smorodinsky, *Other solutions to Nash's bargaining problem*, Econometrica, 43 (1975), pp. 513–518.

[19] F. Kelly and E. Yudovina, *Stochastic Networks*, Vol. 2, Cambridge University Press, Cambridge, 2014.

[20] F. Kuhn, T. Moscibroda, and R. Wattenhofer, *The price of being near-sighted*, in Proceedings ACM-SIAM SODA'06, SIAM, Philadelphia, 2006, pp. 980–989.

[21] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, *An axiomatic theory of fairness in network resource allocation*, in Proceedings IEEE INFOCOM'10, IEEE, Piscataway, NJ, 2010.

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

[23] M. Luby and N. Nisan, *A parallel approximation algorithm for positive linear programming*, in Proceedings ACM STOC'93, ACM, New York, 1993.

[24] M. W. Mahoney, S. Rao, D. Wang, and P. Zhang, *Approximating the solution to mixed packing and covering LPs in parallel $\tilde{O}(\epsilon^{-3})$ time*, in Proceedings ICALP'16, Schloss Dagstuhl, Wadern, Germany, 2016, 52.

[25] J. Marasevic, C. Stein, and G. Zussman, *A fast distributed stateless algorithm for alpha-fair packing problems*, in Proceedings ICALP'16, Schloss Dagstuhl, Wadern, Germany, 2016, 54.

[26] B. McCormick, F. Kelly, P. Plante, P. Gunning, and P. Ashwood-Smith, *Real time alpha-fairness based traffic engineering*, in Proceedings ACM HotSDN'14, ACM, New York, 2014.
[27] J. Mo and J. Walrand, *Fair end-to-end window-based congestion control*, IEEE/ACM Trans. Netw., 8 (2000), pp. 556–567.
[28] J. F. Nash Jr, *The bargaining problem*, Econometrica, 18 (1950), pp. 155–162.
[29] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
[30] Y. Nesterov, *Lectures on Convex Optimization*, Springer Optim. Appl. 137, Springer, Cham, Switzerland, 2018.
[31] C. Papadimitriou and M. Yannakakis, *Linear programming without the matrix*, in Proceedings ACM STOC'93, ACM, New York, 1993.
[32] N. Young, *Sequential and parallel algorithms for mixed packing and covering*, in Proceedings IEEE FOCS'01, IEEE Computer Society, Los Alamitos, CA, 2001.