# Unified Acceleration Method for Packing and Covering Problems via Diameter Reduction[*]

## Di Wang[1], Satish Rao[1], and Michael W. Mahoney[2]

1   Department of Electrical Engineering and Computer Sciences, UC Berkeley,
    Berkeley, USA
    Email: wangd@eecs.berkeley.edu,satishr@berkeley.edu
2   International Computer Science Institute and Department of Statistics, UC
    Berkeley,
    Berkeley, USA
    Email: mmahoney@stat.berkeley.edu

──── **Abstract** ────

In a series of recent breakthroughs, Allen-Zhu and Orecchia [1, 2] leveraged insights from the linear coupling method [15], which is a first-order optimization scheme, to provide improved algorithms for packing and covering linear programs. The result in [1] is particularly interesting, as the algorithm for packing LP achieves both width-independence and Nesterov-like acceleration, which was not known to be possible before. Somewhat surprisingly, however, while the dependence of the convergence rate on the error parameter $\epsilon$ for packing problems was improved to $O(1/\epsilon)$, which corresponds to what accelerated gradient methods are designed to achieve, the dependence for covering problems was only improved to $O(1/\epsilon^{1.5})$, and even that required a different more complicated algorithm, rather than from Nesterov-like acceleration. Given the primal-dual connection between packing and covering problems and since previous algorithms for these very related problems have led to the same $\epsilon$ dependence, this discrepancy is surprising, and it leaves open the question of the exact role that the linear coupling is playing in coordinating the complementary gradient and mirror descent step of the algorithm. In this paper, we clarify these issues, illustrating that the linear coupling method can lead to improved $O(1/\epsilon)$ dependence for both packing and covering problems in a unified manner, i.e., with the same algorithm and almost identical analysis. Our main technical result is a novel dimension lifting method that reduces the coordinate-wise diameters of the feasible region for covering LPs, which is the key structural property to enable the same Nesterov-like acceleration as in the case of packing LPs. The technique is of independent interest and that may be useful in applying the accelerated linear coupling method to other combinatorial problems.

## 1   Introduction

A fractional covering problem, in its generic form, can be written as the following linear program (LP): $\min_{x \geq 0}\{c^T x : Ax \geq b\}$, where $c \in \mathbb{R}^n_{\geq 0}, b \in \mathbb{R}^m_{\geq 0}$, and $A \in \mathbb{R}^{m \times n}_{\geq 0}$.

─────────────────────

Without loss of generality, one can scale the coefficients, in which case one can write this LP in the standard form:

$$\min_{x \geq 0}\{\vec{1}^T x : Ax \geq \vec{1}\}, \text{ where } A \in \mathbb{R}_{\geq 0}^{m \times n} \tag{1}$$

The fractional packing problem, which is the dual of fractional covering, can be written in the standard form as:

$$\max_{y \geq 0}\{\vec{1}^T y : Ay \leq \vec{1}\}, \text{ where } A \in \mathbb{R}_{\geq 0}^{m' \times n'} \tag{2}$$

We denote by OPT the optimal value of a LP. In this case, we say that $x$ is a $(1 + \epsilon)$-*approximation* for the covering LP if $Ax \geq \vec{1}$ and $\vec{1}^T x \leq (1 + \epsilon) \text{OPT}$, and we say that $y$ is a $(1 - \epsilon)$-*approximation* for the packing LP if $Ay \leq \vec{1}$ and $\vec{1}^T y \geq (1 - \epsilon) \text{OPT}$.

Packing and covering problems are important classes of LPs with wide applications, including most resource allocation problems, and they have long drawn interest in theoretical computer science. Although one can use general LP solvers such as the interior point method to solve packing and covering with convergence rate of $\log(1/\epsilon)$, such algorithms usually have very high per-iteration cost, as methods such as the computation of the Hessian and matrix inversion are involved. In the setting of large-scale problems, low precision iterative solvers are often more popular choices. Such solvers usually run in time with a nearly-linear dependence on the problem size, and they have $\text{poly}(1/\epsilon)$ dependence on the approximation parameter. Most such work falls into one of two categories. The first category follows the approach of transforming LPs to convex optimization problems, then applying efficient first-order optimization algorithms. Examples of work in this category include [1–3, 8, 9, 12], and all except [1, 2] apply to more general classes of LPs. The second category is based on the Lagrangian relaxation framework, and some examples of work in this category include [4–7, 11, 13, 14]. For a more detailed comparison of this prior work, see Table 1 in [1]. Also, based on whether the running time depends on the width $\rho$, a parameter which typically depends on the dimension and the largest entry of $A$, these algorithms can also be divided into width-dependent solvers and width-independent solvers. Width-dependent solvers are usually pseudo-polynomial, as the running time depends at least linearly on $\rho \text{OPT}$, which itself can be large, while width-independent solvers are independent or logarithmically dependent on the width.

In this paper, we describe a solver for covering LPs of the form (1). The solver is width-independent, and it is a first-order method with a linear rate of convergence. That is, if we let $N$ be the number of non-zeros in $A$, then the running time of our algorithm is $O\left(N \frac{\log^2(N/\epsilon) \log(1/\epsilon)}{\epsilon}\right)$. To simplify the following discussion, we will follow the standard practice of using $\tilde{O}$ to hide poly-log factors, in which case the running time of our algorithm for the covering problem is $\tilde{O}(N/\epsilon)$. Among other things, our result is an improvement over the recent bound of $\tilde{O}(N/\epsilon^{1.5})$ provided by Allen-Zhu and Orecchia for the covering problem in [1], and our result corresponds to the linear rate of convergence that accelerated gradient methods are designed to achieve [9].

At least as interesting as the $\tilde{O}(1/\epsilon^{0.5})$ improvement for covering LPs, however, is the context of this problem and the main technical contribution that we developed and exploited to achieve our improvement.

- The context for our results has to do with the linear coupling method that was introduced recently by Allen-Zhu and Orecchia [15]. This is a first order method for solving convex optimization problems, and it provides a conceptually simple way to integrate a gradient descent step and mirror descent step in each iteration. In the setting of standard

smooth convex optimization, the method achieves the same convergence rate as that of the accelerated gradient descent method of Nesterov [9], and indeed the former can be viewed as an insightful reinterpretation of the latter. The high-level view of the method as a coupling of gradient descent steps and mirror descent steps offers more flexibility to the framework, as the combination allows the two steps to complement each other in ways beyond simply Nesterov-like acceleration. Indeed, it has shown initial promise by providing improved algorithms for packing and covering LPs [1, 2]. The packing algorithm of Allen-Zhu and Orecchia in [1] is particularly surprising, as it exploits the linear coupling framework to achieve both width-independence and Nesterov-like acceleration, which is widely believed to be very difficult, and is the first success in a long line of works in this area.

The particular motivation for our work is a striking discrepancy between bounds provided for packing and covering LPs in [1]. In particular, they provide a $(1 - \epsilon)$-approximation solver for the packing problem in $\tilde{O}(N/\epsilon)$, but they are only able to obtain $\tilde{O}(N/\epsilon^{1.5})$ for the covering problem. In the case of covering, they are unable to use the linear coupling method to achieve Nesterov-like acceleration, and even to get width-independence the authors need to integrate some ad-hoc and complicated techniques. This discrepancy between results for packing and covering LPs is rare, due to the duality between them. Filling this gap is of particular interests, as not being able to do so would suggest some fundamental structural differences between the two problems.

Our main technical contribution is a novel diameter reduction method for fractional covering LPs that helps resolve this discrepancy. Recall that the smoothness parameter, e.g., Lipschitz constant, and the diameter of the feasible region are the two most natural limiting factors for most gradient based optimization algorithms. Indeed, many applications of general first-order optimization techniques can be attributed to the existence of norms or proximal setups for the specific problems that gives both good smoothness and diameter properties. In the particular case of coordinate descent algorithms based on the linear coupling idea, we additionally need good coordinate-wise diameter properties to achieve accelerated convergence.

This is easy to accomplish for packing problems, but it is not easy to do for covering problems, and this is the difference that leads to the $\tilde{O}(1/\epsilon^{0.5})$ discrepancy between packing and covering algorithms in previous work [1]. Our diameter reduction method for general covering problems is based on dimension lifting, which transforms the covering problem space to a higher dimensional space, and the feasible region in the lifted space has both good global diameter bounds with respect to the canonical norm for accelerated stochastic coordinate descent (as is needed generally [1, 10]) as well as good coordinate-wise diameter bounds (as is needed for linear coupling [1]). Thus, it is likely of interest more generally for combinatorial optimization problems.

Once the diameter reduction is achieved, covering LP shares all the essential properties necessary to achieve both width-independence and Nesterov-like acceleration as in the case of packing problems, and fits elegantly into the scheme and analysis from [1] that was developed for packing LPs. We obtain improved $\tilde{O}(N/\epsilon)$ results for covering LPs, and this provides a unified acceleration method (unified in the sense that it is with the same algorithm and almost identical analysis) for both packing and covering LPs.

We will start in Section 2 with a discussion of some selected technical ideas and challenges from previous work. Then, in Section 3 we will present our main technical contribution, a novel diameter reduction method for any covering LP of the form given in (1). Finally, in Section 4 we describe how to combine this with previous work to obtain a unified acceleration

method for packing and covering problems.

## 2    High-level Description of Challenges

At a high level, we (as well as Allen-Zhu and Orecchia [1,2]) use the same two-step approach of Nesterov [9]. The first step involves smoothing, which transforms the constrained problem into a *smooth* objective function with trivial or no constraints. By smooth, we mean that the gradient of the objective function has some property in the flavor of Lipschitz continuity. Once smoothing is accomplished, the second step uses one of several first order methods for convex optimization in order to obtain an approximate solution to the smoothed objective. Standard applications of this approach usually lead to width-dependent algorithms, where the width enters the performance analysis as the magnitude of the gradients.

The first width-independent result following the optimization approach in [2] achieves width-independence by truncating the gradient, thus effectively reducing the width to 1. The algorithm uses, in a white-box way, the coupling of mirror descent and gradient descent from [15], where the progress from gradient descent covers the loss incurred by the truncation of the gradient (see Eqn. (7) below for the precise formulation of this loss), thus achieving width-independence. However, the role of gradient descent in the coupling is limited to width-independence, but not acceleration.

To improve the sequential packing solver in [2] with convergence $\tilde{O}(1/\epsilon^3)$ to $\tilde{O}(1/\epsilon)$, the same authors in [1] apply a stochastic coordinate descent method based on the linear coupling idea. Barring the difference between Lipschitz and local Lipschitz continuity, the results in [1] can be viewed as a variant of accelerated coordinate descent method [10]. There are two places where the algorithm achieves an improvement over prior packing-covering results.

- One factor of improvement is due to the better coordinate-wise Lipschitz constant over the full dimensional Lipschitz constant. Intuitively, in the case of packing or covering, the gradient of variable $x_i$ depends on the penalties of constraints involving $x_i$, which further depend on all the variables in those constraints. As a result, if we move all the variables simultaneously, we can only take a small step before changing the gradient of $x_i$ drastically. Sometimes coordinate descent comes with a downside, since if we update one variable each iteration, computing $n$ partial derivatives in $n$ iterations can be much more expensive than computing all the $n$ partial derivatives in the same iteration. However, it can be shown in the case of packing and covering LPs, there is no such computation overhead.

- The other factor of improvement comes from Nesterov-like acceleration. In addition to giving width-independence as in [2], the gradient descent also covers the regret term incurred by the mirror descent step (see Eqn. (7) below for the precise formulation of this regret), which is the key insight from the original linear coupling result [15] to reproduce Nesterov's accelerated convergence. It turns out that nice diameter properties are necessary for the latter to be possible. On a high level, the regret incurred by mirror descent is proportional to its step size, which has an upper-bound proportional to the coordinate-wise diameter of the feasible region. On the other hand, the progress made by the gradient descent step is also proportional to its step size, which is inversely proportional to the Lipschitz parameter. For both packing and covering problems, the coordinate-wise Lipschitz parameter of $x_i$ is proportional to $1/\|A_{:i}\|_\infty$, as $\|A_{:i}\|_\infty$ captures the impact of $x_i$ on the values of the constraints, which determine the gradient of $x_i$. This works out particularly well for packing problems, since the packing constraints

$Ax \leq \vec{1}$ impose a natural coordinate-wise diameter of $x_i^* \leq 1/\|A_{:i}\|_\infty$ on the feasible region, which aligns the gradient descent step size and mirror descent step size, making the coupling possible to accelerate. The same small coordinate-wise diameter is also crucial to get good global diameter for the proximal setup used in mirror descent, which is necessary for mirror descent to achieve good convergence.

The combination of gradient truncation, stochastic coordinate descent, and acceleration due to the nice diameter properties lead to the $\tilde{O}(N/\epsilon)$ solver for the packing LP [1].

Shifting to solvers for the covering LP, one obvious obstacle to reproducing the packing result is we no longer have the small diameters. Indeed, a naive coordinate-wise upper bound from the covering constraints only gives $x_i^* \leq 1/\min_j\{A_{ji} : A_{ji} > 0\}$, which is far from sufficient to give acceleration as the packing solver in [1]. The authors instead go back to the setup in their earlier work [2], where linear coupling only gives width-independence. The authors use a negative-width technique as in [3] (Theorem 3.3 with $l = \sqrt{\epsilon}$), that leads to the (improved, but still worse than for packing) $\tilde{O}(1/\epsilon^{1.5})$ convergence rate.

To get an $\tilde{O}(1/\epsilon)$ solver for the covering LP, it seems crucial to relate the gradient descent step and mirror descent step the same way as in the packing solver in [1]. Thus, we will work directly to reduce the coordinate-wise diameter. Our main result (presented next in Section 3) is a general diameter reduction method to achieve the same diameter property as in the packing solver, and this enables us (in Section 4) to extend all the crucial ideas of the packing solver in [1], as outlined in this section, to get a covering solver with running time $\tilde{O}(N/\epsilon)$.

## 3 Diameter Reduction Method for General Covering Problems

Given any covering LP of the form in (1), characterized by a matrix $A$, we formulate an equivalent covering LP with good diameter properties. This will involve lifting the instance to higher dimensional space by adding variables and redundant constraints. On a high level, as we discussed in last section, the obstacle for covering problems lies in the discrepency between the large coordinate-wise diameter of the feasible region and the small gradient descent step size. Our answer is essentially for each variable to create multiple copies with different resolutions. Certain copies will be in charge of searching over larger regions, but for them we modify their coefficients in the lifted space to allow larger gradient descent steps. We use $i \in [n]$ to denote the indices of the variables (i.e., columns of $A$) and $j \in [m]$ to denote the indices of constraints (i.e., rows of $A$). For ease of comparison with [1], and since our unified approach for both packing and covering uses their packing solver and a similar analysis, we use the same notation whenever possible.

For any $i \in [n]$, let

$$r_i \stackrel{\text{def}}{=} \frac{\max_j\{A_{ji} : A_{ji} > 0\}}{\min_j\{A_{ji} : A_{ji} > 0\}},$$

be the ratio between the largest non-zero coefficient and the smallest non-zero coefficient of variable $x_i$ in all constraints, and let $n_i \stackrel{\text{def}}{=} \lceil \log r_i \rceil$. We first duplicate each original variable $n_i$ times to obtain $\bar{x}_{(i,l)}, i \in [n], l \in [n_i]$ as the new variables. In terms of the coefficient matrix, we now have a new matrix, call it $\bar{A} \in \mathbb{R}_{\geq 0}^{m \times (\sum_i n_i)}$, which contains $n_i$ copies of the $i$-th column $A_{:i}$. We denote a column of $\bar{A}$ by the tuple $(i,l)$ with $l \in [n_i]$. Obviously, the covering LP given by $\bar{A}$ is equivalent to the original covering LP given by $A$. Adding additional copies of variables, however, will allow us to improve the diameter. To reduce the diameter of this new covering LP, we further decrease some of the coefficients in $\bar{A}$, and we

put upper bounds on the variables. In particular, for $j, i, l$, we have

$$\bar{A}_{j,(i,l)} = \min\{A_{j,i}, 2^l \min_j\{A_{ji} : A_{ji} > 0\}\}, \tag{3}$$

and for variable $\bar{x}_{(i,l)}$, we add the constraint

$$\bar{x}_{(i,l)} \leq \frac{2}{2^l \min_j\{A_{ji} : A_{ji} > 0\}}. \tag{4}$$

The next lemma shows that the covering LP given by $\bar{A}$ and the covering LP given by $A$ are equivalent.

▶ **Lemma 1.** *The covering LP of $A$ and the covering LP of $\bar{A}$ have the same optimal value* OPT. *Furthermore, there exists an optimal solution of the covering LP of $\bar{A}$ inside the region specified by* (4).

**Proof.** Let $\overline{\text{OPT}}$ be the optimal of the LP given by the covering constraints of $\bar{A}$ and the coordinate-wise upper-bounds in (4). We need to show $\text{OPT} = \overline{\text{OPT}}$. Given any feasible solution $\bar{x}$, consider the solution $x$ where $x_i = \sum_{l=1}^{n_i} \bar{x}_{(i,l)}$. It is obvious $\vec{1}^T x = \vec{1}^T \bar{x}$, and $Ax \geq \vec{1}$, as coefficients in $\bar{A}$ are no larger than coefficients in $A$. Thus $\text{OPT} \leq \overline{\text{OPT}}$.

For the other direction, consider any feasible $x$. For each $i$, we can assume without loss of generality that

$$x_i \leq \frac{1}{\min_j\{A_{ji} : A_{ji} > 0\}}.$$

Let $l_i$ be the largest index such that

$$x_i \leq \frac{2}{2^{l_i} \min_j\{A_{ji} : A_{ji} > 0\}},$$

and then let

$$\bar{x}_{(i,l)} = \begin{cases} x_i & \text{if } l = l_i \\ 0 & \text{if } l \neq l_i \end{cases}.$$

By construction, $\bar{x}$ satisfies all the upper bounds described in (4). Furthermore, for constraint $j$, we must have $\bar{A}_{j:}\bar{x} \geq 1$. Since for any $i$, $\bar{A}_{j,(i,l_i)}$ differs from $A_{ji}$ only when $A_{ji} > 2^{l_i} \min_j\{A_{ji} : A_{ji} > 0\}$, and we must have $l_i < n_i$ in this case by definition of $n_i$, which gives $\bar{x}_{(i,l_i)} = x_i \geq \frac{1}{2^{l_i} \min_j\{A_{ji}:A_{ji}>0\}}$ by our choice of $l_i$ being the largest possible. Then we know $\bar{A}_{j,(i,l_i)} = 2^{l_i} \min_j\{A_{ji} : A_{ji} > 0\}$, so the $j$-th constraint is satisfied. Thus $\text{OPT} \geq \overline{\text{OPT}}$, and we can conclude $\text{OPT} = \overline{\text{OPT}}$. ◀

Given the equivalence of the covering LP defined by $\bar{A}$ and that defined by $A$, we now point out that the seemingly-redundant constraints of (4) turn out to be crucial. The reason is that we can search over a feasible region with nice diameter properties necessary to tap the full power of the linear coupling method. In particular, we can rewrite the constraints (4) to be

$$\bar{x}_{(i,l)} \leq \frac{2}{\|\bar{A}_{:(i,l)}\|_\infty}.$$

For any $i$, this is the same upper bound on $\bar{x}_{(i,l)}$ for $l < n_i$ (consider the row $j^* = \text{argmax}_j\{A_{ji}, A_{ji} > 0\}$), and it is a relaxation on $\bar{x}_{(i,n_i)}$.

The price we pay for this diameter improvement is that the new LP defined by $\bar{A}$ is larger than that defined by $A$. Two comments on this are in order. First, by Observation 3,

$r_i$ is bounded by $n^2 m/\epsilon^2$, and so the diameter reduction step only increases the problem size by $O(\log(mn/\epsilon))$. Second, we have presented our diameter reduction as an explicit pre-processing step so we can use one unified optimization algorithm (Algorithm 1 below) for both packing and covering, but in practice the diameter reduction would not have to be carried out explicitly. It can equivalently be implemented implicitly within the algorithm (a trivially-modified version of Algorithm 1 below) by randomly choosing a scale after picking the coordinate $i$ and then computing $\bar{A}_{j,(i,l)}$ in (3) by shifting bits on the fly.

Given this reduction, in the rest of the paper, when we refer to the covering LP, we will implicitly be referring to the diameter reduced version, and we have the additional guarantee that there exists an optimal solution $x^*$ to (1) such that

$$0 \leq x_i^* \leq \frac{2}{\|A_{:i}\|_\infty} \quad \forall i \in [n]. \tag{5}$$

## 4 An Accelerated Solver for (Packing and) Covering LPs

In this section, we will show covering LPs fit neatly into the scheme and analysis developed for packing LPs in [1], thus establishing a unified acceleration method for packing and covering problems. To motivate this, recall that for packing problems of the form (2), bounds of the form (5) automatically follow from the packing constraints $Ax \leq \vec{1}$. For readers familiar with the packing LP solver in [1], it should be plausible that—once we have this diameter property—the same stochastic coordinate descent optimization scheme will lead to a $\tilde{O}(N/\epsilon)$ covering LP solver.

In Section 4.1, we'll present some preliminaries and describe how we perform smoothing on the original covering objective function; and then in Section 4.2, we'll present the main algorithm. This algorithm involves a mirror descent step, that will be described in Section 4.3, a gradient descent step, that will be described in Section 4.4, and a careful coupling between the two, that will be described in Section 4.5. Finally, in Section 4.6, we will describe how to ensure we start at a good starting point. Some of the following results are technically-tedious but conceptually-straightforward extensions of analogous results from [1], and some of the results are restated from [1]; we defer most of the proofs to the full version.

### 4.1 Preliminaries and Smoothing the Objective

To start, let's assume that $\min_{j \in [m]} \|A_{j:}\|_\infty = 1$. This assumption is without loss of generality: since we can simply scale $A$ for this to hold without sacrificing approximation quality. With this assumption, the following lemma holds.

▶ **Lemma 2.** $OPT \in [1, m]$

With OPT being at least 1, the error we introduce later in the smoothing step will be small enough that the smoothing function approximates the covering LP well enough with respect to $\epsilon$ around the optimum.

▶ **Observation 3.** It can be shown that to obtain a $(1+O(\epsilon))$-approximation, we can eliminate entries smaller than $\frac{\epsilon}{mn}$ and entries larger than $\frac{n}{\epsilon}$ from matrix $A$.

We will turn the covering LP objective into a smoothed objective function $f_\mu(x)$, as used in [1, 2, 4], and we are going to find a $(1 + \epsilon)$-approximation of the covering LP by approximately minimizing $f_\mu(x)$ over the region

$$\Delta \overset{\text{def}}{=} \{x \in \mathbb{R}^n : 0 \leq x_i \leq \frac{3}{\|A_{:i}\|_\infty}\}.$$

The function $f_\mu(x)$ is

$$f_\mu(x) \stackrel{\text{def}}{=} \vec{1}^T x + \max_{y \geq 0}\{y^T(\vec{1} - Ax) + \mu H(y)\},$$

and it is a smoothed objective in the sense that it turns the covering constraints into soft penalties, with $H(y)$ being a regularization term. Here, we use the generalized entropy $H(y) = -\sum_j y_j \log y_j + y_j$, where $\mu$ is the smoothing parameter balancing the penalty and the regularization. It is straightforward to compute the optimal $y$, and write $f_\mu(x)$ explicitly, as stated in the following lemma.

▶ **Lemma 4.** $f_\mu(x) = \vec{1}^T x + \mu \sum_{j=1}^m p_j(x)$, where $p_j(x) \stackrel{\text{def}}{=} \exp(\frac{1}{\mu}(1 - (Ax)_j))$.

Optimizing $f_\mu(x)$ over $\Delta$ gives a good approximation to OPT, in the following sense. If we let $x^*$ be an optimal solution satisfying (5), and $u^* \stackrel{\text{def}}{=} (1 + \epsilon/2)x^* \in \Delta$, then we have the properties in the following lemma.

▶ **Lemma 5.** *Setting the smoothing parameter $\mu = \frac{\epsilon}{4\log(nm/\epsilon)}$, we have*
1. $f_\mu(u^*) \leq (1 + \epsilon)\,\text{OPT}$.
2. $f_\mu(x) \geq (1 - \epsilon)\,\text{OPT}$ *for any $x \geq 0$.*
3. *For any $x \geq 0$ satisfying $f_\mu(x) \leq 2\,\text{OPT}$, we must have $Ax \geq (1 - \epsilon)\vec{1}$.*
4. *If $x \geq 0$ satisfies $f_\mu(x) \leq (1 + O(\epsilon))\,\text{OPT}$, then $\frac{1}{1-\epsilon}x$ is a $(1 + O(\epsilon))$-approximation to the covering LP.*
5. *The gradient of $f_\mu(x)$ is*

$$\nabla f_\mu(x) = \vec{1} - A^T \vec{p(x)} \quad \text{where} \quad p_j(x) \stackrel{\text{def}}{=} \exp(\frac{1}{\mu}(1 - (Ax)_j),$$

*and $\nabla_i f_\mu(x) = 1 - \sum_j A_{ji} p_j(x) \in [-\infty, 1]$.*

Although $f_\mu(x)$ gives a good approximation to the covering LP, $f_\mu(x)$ doesn't have the necessary Lipschitz-smoothness property due to the fast changing nature of exponential functions. However, $f_\mu(x)$ is *locally Lipschitz continuous*, in a sense quantified by the following lemma, and so we have a good improvement with a gradient step within certain range.

▶ **Lemma 6.** *Let $L \stackrel{\text{def}}{=} \frac{4}{\mu}$, for any $x \in \Delta$, and $i \in [n]$*
1. *If $\nabla_i f_\mu(x) \in (-1, 1)$, then for all $|\gamma| \leq \frac{1}{L\|A_{:i}\|_\infty}$, we have*

$$|\nabla_i f_\mu(x) - \nabla_i f_\mu(x + \gamma\, \mathbf{e}_i)| \leq L\|A_{:i}\|_\infty |\gamma|.$$

2. *If $\nabla_i f_\mu(x) \leq -1$, then for all $\gamma \leq \frac{1}{L\|A_{:i}\|_\infty}$, we have*

$$\nabla_i f_\mu(x + \gamma\, \mathbf{e}_i) \leq (1 - \frac{L\|A_{:i}\|_\infty}{2}|\gamma|)\nabla_i f_\mu(x).$$

We call $L\|A_{:i}\|_\infty$ the *coordinate-wise local Lipschitz constant*. The significance of Lemma 6 is that for covering LPs the coordinate-wise local Lipschitz constant is inversely proportional to the coordinate-wise diameter. (This fact has been established previously for the case of packing LPs [1].)

---

**Algorithm 1** Accelerated stochastic coordinate descent for both packing and covering

---

**Input**: $A \in \mathbb{R}_{\geq 0}^{m \times n}, x^{\text{start}} \in \Delta, f_\mu, \epsilon$ **Output**: $y_T \in \Delta$

1: $\mu \leftarrow \frac{\epsilon}{4 \log(nm/\epsilon)}, L \leftarrow \frac{4}{\mu}, \tau \leftarrow \frac{1}{8nL}$

2: $T \leftarrow \lceil 8nL \log(1/\epsilon) \rceil = \tilde{O}(\frac{n}{\epsilon})$

3: $x_0, y_0, z_0 \leftarrow x^{\text{start}}, \alpha_0 \leftarrow \frac{1}{nL}$

4: **for** $k = 1$ to $T$ **do**

5:      $\alpha_k \leftarrow \frac{1}{1-\tau}\alpha_{k-1}$

6:      $x_k \leftarrow \tau z_{k-1} + (1-\tau)y_{k-1}$

7:      Select $i \in [n]$ uniformly at random.

     ▷ Gradient truncation:

8:      Let $(\xi_k^{(i)})_i \leftarrow \begin{cases} -\mathbf{e}_i & \nabla_i f_\mu(x_k) < -1 \\ \nabla_i f_\mu(x_k) \cdot \mathbf{e}_i & \nabla_i f_\mu(x_k) \in [-1, 1] \\ \mathbf{e}_i & \nabla_i f_\mu(x_k) > 1 \end{cases}$

     ▷ Mirror descent step:

9:      $z_k \leftarrow z_k^{(i)} \stackrel{\text{def}}{=} \text{argmin}_{z \in \Delta}\{V_{z_{k-1}}(z) + \langle z, n\alpha_k \xi_k^{(i)}\rangle\}$.

     ▷ Gradient descent step:

10:     $y_k \leftarrow y_k^{(i)} \stackrel{\text{def}}{=} x_k + \frac{1}{n\alpha_k L}(z_k^{(i)} - z_{k-1})$

11: **end for**

12: **return** $y_T$.

---

## 4.2 An Accelerated Coordinate Descent Algorithm

We will now show that the accelerated coordinate descent used in packing LP solver in [1] also works as a covering LP solver, with appropriately-chosen starting points and smoothed objective functions. Consider Algorithm 1, which is our main accelerated stochastic coordinate descent for both packing and covering. Note for both packing and covering LPs, we give $\Delta = \{x \in \mathbb{R}^n : 0 \leq x_i \leq \frac{3}{\|A_{:i}\|_\infty}\}$ as the input feasible region. The correctness of this algorithm and its running time guarantees for the packing problem have already been nicely presented in [1], and so here we will focus on the covering problem.

Our main result is summarized in the following theorems.

▶ **Theorem 7.** *With $x^{\text{start}}$ computable in time $\tilde{O}(N)$ to be specified later, Algorithm 1 outputs $y_T$ satisfying $\mathbb{E}[f_\mu(y_T)] \leq (1 + 6\epsilon)\,\text{OPT}$, and the running time is $\tilde{O}(N/\epsilon)$.*

A standard application of Markov bound gives the following corollary.

▶ **Corollary 8.** *There is a algorithm that, with probability at least $9/10$, computes a $(1+O(\epsilon))$-approximation to the fractional covering problem and has $\tilde{O}(N/\epsilon)$ expected running time.*

Before proceeding with our proof of these theorems, we discuss briefly the optimization scheme from [1] we will use. First, the $A$-norm is used as the proximal setup for mirror descent, where

$$\|x\|_A = \sqrt{\sum_i \|A_{:i}\|_\infty x_i^2}, \tag{6}$$

The corresponding distance generating function is $w(x) = \frac{1}{2}\|x\|_A^2$, and the Bregman divergence is $V_x(y) = \frac{1}{2}\|x - y\|_A^2$.[1]

---

[1] In particular, $w$ is a 1-strongly convex function with respect to $\|\cdot\|_A$, and $V_x(y) \stackrel{\text{def}}{=} w(y) - \langle \nabla w(x), y - x\rangle - w(x)$. See [15] for a detailed discussion of mirror descent as well as and several interpretations.

Next, observe that Algorithm 1 works as follows. Each iteration integrates a mirror descent step and a gradient descent step. The standard analysis of mirror descent gives a convergence of $\frac{1}{\epsilon^2}$, and it depends on the width of the problem. Here is how the coupling of gradient descent and mirror descent achieves both width-independence and linear-rate acceleration.

- To eliminate the width from the convergence rate, the gradient $\nabla_i f_\mu(x_k)$ is split into the small component, $\xi_k^{(i)} = \max\{-1, \nabla_i f_\mu(x_k)\}\mathbf{e}_i$, and the large component, $\eta_k^{(i)} = \nabla_i f_\mu(x_k)\mathbf{e}_i - \xi_k^{(i)}$. Only the small component $\xi^{(i)}$ is given to the mirror descent step, and thus the width is effectively 1. However, the truncation incurs loss from the large component, as the mirror descent only acts on the small component. The progress from the gradient descent step is used to cover that loss.
- In order to get to $1/\epsilon$ convergence, recall that the $1/\epsilon^2$ in the convergence of mirror descent is largely due to the regret term accumulated along all iterations of mirror descent. The progress from the gradient step also covers the regret from the mirror descent step (see Eqn. (7) below for the precise formulation of this loss and regret). This enables the coupling to get Nesterov-like acceleration using the same approach in [15].

Before we moving to formalize the above discussion, here are some lemmas about the algorithm. The first lemma says that the gradient step we take is always valid (i.e., in $\Delta$), which is crucial in the sense that we need the step length to be at least $\frac{1}{n\alpha_k L}$ of the mirror descent step length for the coupling to work.

▶ **Lemma 9.** *We have $x_k, y_k, z_k \in \Delta$ for all $k = 0, 1, \ldots, T$.*

The second lemma is clearly crucial to achieve the nearly linear time $\tilde{O}(N/\epsilon)$ algorithm.

▶ **Lemma 10.** *Each iteration can be implemented in expected $O(N/n)$ time.*

## 4.3   Mirror Descent Step

We now analyze the mirror descent step of Algorithm 1:

$$z_k \leftarrow z_k^{(i)} \stackrel{\text{def}}{=} \operatorname*{argmin}_{z \in \Delta}\{V_{z_{k-1}}(z) + \langle z, n\alpha_k \xi_k^{(i)}\rangle\}.$$

▶ **Lemma 11.** $\langle n\alpha_k \xi_k^{(i)}, z_{k-1} - u^*\rangle \leq n^2\alpha_k^2 L\langle \xi_k^{(i)}, x_k - y_k^{(i)}\rangle + V_{z_{k-1}}(u^*) - V_{z_k}(u^*)$

Also, we note that the mirror descent step, defined above in a variational way, can be explicitly written as

1. $z_k^{(i)} \leftarrow z_{k-1}$
2. $z_k^{(i)} \leftarrow z_k^{(i)} - n\alpha_k \xi_k^{(i)}/\|A_{:i}\|_\infty$
3. If $z_{k,i}^{(i)} < 0, z_{k,i}^{(i)} \leftarrow 0$; if $z_{k,i}^{(i)} > 3/\|A_{:i}\|_\infty, z_{k,i}^{(i)} \leftarrow 3/\|A_{:i}\|_\infty$.

## 4.4   Gradient Descent Step

We now analyze the gradient descent step of Algorithm 1. In particular, from the explicit formulation of the mirror descent step, we have that $|z_{k,i}^{(i)} - z_{k-1,i}| \leq \frac{n\alpha_k |\xi_k^{(i)}|}{\|A_{:i}\|_\infty}$, which gives

$$|y_{k,i}^{(i)} - x_{k,i}| = \frac{1}{n\alpha_k L}|z_{k,i}^{(i)} - z_{k-1,i}| \leq \frac{|\xi_k^{(i)}|}{L\|A_{:i}\|_\infty}.$$

The gradient step we take is within the local region, and so Lemma 6 applies. We bound the progress from the gradient descent step in the following lemma.

▶ **Lemma 12.** $f_\mu(x_k) - f_\mu(y_k^{(i)}) \geq \frac{1}{2}\langle \nabla f_\mu(x_k), x_k - y_k^{(i)}\rangle$

## 4.5   Coupling of Gradient and Mirror Descent

Here, we will analyze the coupling between the gradient descent and mirror descent steps. This and the next section will give a proof of Theorem 7.

As we take steps on random coordinates, we will write the full gradient as

$$\nabla f_\mu(x_k) = \mathbb{E}_i[n \nabla_i f_\mu(x_k)] = \mathbb{E}_i[n\eta_k^{(i)} + n\xi_k^{(i)}].$$

As discussed earlier, we have the small component $\xi_k^{(i)} \in (-1,1)\,\mathbf{e}_i$ and the large component $\eta_k^{(i)} = \nabla_i f_\mu(x_k) - \xi_k^{(i)} \in (-\infty, 0]\,\mathbf{e}_i$. We put the gradient and mirror descent steps together, and we bound the gap to optimality at iteration $k$:

$$
\begin{aligned}
\alpha_k(f_\mu(x_k) - f_\mu(u^*)) \leq & \langle \alpha_k \nabla f_\mu(x_k), x_k - u^* \rangle \\
= & \langle \alpha_k \nabla f_\mu(x_k), x_k - z_{k-1} \rangle + \langle \alpha_k \nabla f_\mu(x_k), z_{k-1} - u^* \rangle \\
= & \langle \alpha_k \nabla f_\mu(x_k), x_k - z_{k-1} \rangle + \mathbb{E}_i[\langle n\alpha_k \eta_k^{(i)}, z_{k-1} - u^* \rangle] \\
& + \mathbb{E}_i[\langle n\alpha_k \xi_k^{(i)}, z_{k-1} - u^* \rangle] \\
= & \frac{1-\tau}{\tau} \alpha_k \langle \nabla f_\mu(x_k), y_{k-1} - x_k \rangle + \mathbb{E}_i[\langle n\alpha_k \eta_k^{(i)}, z_{k-1} - u^* \rangle] \\
& + \mathbb{E}_i[\langle n\alpha_k \xi_k^{(i)}, z_{k-1} - u^* \rangle] \\
\leq & \frac{1-\tau}{\tau} \alpha_k (f_\mu(y_{k-1}) - f_\mu(x_k)) + \mathbb{E}_i[\langle n\alpha_k \eta_k^{(i)}, z_{k-1} - u^* \rangle] \\
& + \mathbb{E}_i[n^2 \alpha_k^2 L \langle \xi_k^{(i)}, x_k - y_k^{(i)} \rangle] + V_{z_{k-1}}(u^*) - V_{z_k^{(i)}}(u^*).
\end{aligned}
$$

The first line is due to convexity. The fourth line is due to $x_k = \tau z_{k-1} + (1-\tau)y_{k-1}$, so $\tau(x_k - z_{k-1}) = (1-\tau)(y_{k-1} - x_k)$. The last line is by Lemma 11.

We need to use the progress from the gradient step given in Lemma 12 to cover the loss from $\eta_k^{(i)}$, and the regret from the mirror descent step:

$$\underbrace{\mathbb{E}_i[\langle n\alpha_k \eta_k^{(i)}, z_{k-1} - u^* \rangle]}_{\text{loss from } \eta_k^{(i)}} + \underbrace{\mathbb{E}_i[n^2 \alpha_k^2 L \langle \xi_k^{(i)}, x_k - y_k^{(i)} \rangle]}_{\text{regret from mirror descent}}, \tag{7}$$

The following lemma crucially relies on the nice coordinate-wise diameters of the feasible region $\Delta$.

▶ **Lemma 13.** *The (scaled) progress from the gradient step covers both the loss from gradient truncation and the regret incurred by the mirror descent step*

$$\mathbb{E}_i[\langle n\alpha_k \eta_k^{(i)}, z_{k-1} - u^* \rangle] + \mathbb{E}_i[n^2 \alpha_k^2 L \langle \xi_k^{(i)}, x_k - y_k^{(i)} \rangle] \leq \mathbb{E}_i[8n\alpha_k L(f_\mu(x_k) - f_\mu(y_k^{(i)}))].$$

Now we can show this gives Nesterov-like acceleration. We have

$$
\begin{aligned}
\alpha_k(f_\mu(x_k) - f_\mu(u^*)) \leq & \frac{1-\tau}{\tau} \alpha_k(f_\mu(y_{k-1}) - f_\mu(x_k)) + \mathbb{E}_i[8n\alpha_k L(f_\mu(x_k) - f_\mu(y_k^{(i)})] \\
& + \mathbb{E}_i[V_{z_{k-1}}(u^*) - V_{z_k^{(i)}}(u^*)].
\end{aligned}
$$

With our choice of $\tau = \frac{1}{8nL}$, $\alpha_k = \frac{1}{1-\tau}\alpha_{k-1}$, we get

$$-\alpha_k f_\mu(u^*) \leq 8nL\alpha_{k-1} f_\mu(y_{k-1}) - \mathbb{E}_i[8nL\alpha_k f_\mu(y_k^{(i)})] + \mathbb{E}_i[V_{z_{k-1}}(u^*) - V_{z_k^{(i)}}(u^*)].$$

Telescoping the above inequality $^2$ along $k = 1, \ldots, T$, we get

$$\mathbb{E}[8nL\alpha_T f_\mu(y_T)] \leq \sum_{k=1}^{T} \alpha_k f_\mu(u^*) + 8nL\alpha_0 f_\mu(y_0) + V_{z_0}(u^*),$$

and thus

$$\mathbb{E}[f_\mu(y_T)] \leq \frac{\sum_{k=1}^{T} \alpha_k}{8nL\alpha_T} f_\mu(u^*) + \frac{\alpha_0}{\alpha_T} f_\mu(y_0) + \frac{1}{8nL\alpha_T} V_{z_0}(u^*).$$

We have $\sum_{k=1}^{T} \alpha_k = \alpha_T \sum_{k=0}^{T-1} (1 - \frac{1}{8nL})^k = 8nL\alpha_T(1 - (1 - \frac{1}{8nL})^T) \leq 8nL\alpha_T$, and by our choice of $T = \lceil 8nL \log(1/\epsilon) \rceil$, we also have

$$\frac{\alpha_0}{\alpha_T} = (1 - \frac{1}{8nL})^T \leq \epsilon, \frac{1}{8nL\alpha_T} \leq \frac{\epsilon}{8nL\alpha_0} = \frac{\epsilon}{8},$$

and thus

$$\mathbb{E}_i[f_\mu(y_T)] \leq f_\mu(u^*) + \epsilon f_\mu(y_0) + \frac{\epsilon}{8} V_{z_0}(u^*). \tag{8}$$

## 4.6    Finding a Good Starting Point

From (8), we see a good starting point $y_0 = x^{\text{start}}$ for Algorithm 1 is a point that is not too far away from the optimal in terms of the function value (i.e small $f_\mu(y_0)$), and not too far away from $u^*$ in $A$-norm (i.e. small $V_{z_0}(u^*)$). For packing problems, starting with the all-0's vector will work, but this will not work for covering problems. Instead, for covering problems, we will show now a good enough $x^{\text{start}}$ can be obtained in $\tilde{O}(N)$.

To do so, recall that we can get a 2-approximation $x^{\#}$ to the original covering LP in time $\tilde{O}(N)$ using various nearly linear time covering solvers, e.g., those of [4, 6, 7, 14]. Without loss of generality, we can assume $x_i^{\#} \in [0, \frac{2}{\|A_{:i}\|_\infty}]$, since we can use the diameter reduction process as specified in Lemma 1 to get an equivalent solution satisfying the conditions. Then, we have the following lemma.

▶ **Lemma 14.** *Let* $x^{\text{start}} = (1 + \epsilon/2)x^{\#}$, *we have* $x^{\text{start}} \in \Delta$, $f_\mu(x^{\text{start}}) \leq 4\,\text{OPT}$, *and* $V_{x^{\text{start}}}(u^*) \leq 6\,\text{OPT}$

It is now clear from (8) that we have

$$\mathbb{E}_i[f_\mu(y_T)] \leq f_\mu(u^*) + \epsilon f_\mu(y_0) + \frac{\epsilon}{8} V_{z_0}(u^*) \leq (1+\epsilon)\,\text{OPT} + 4\epsilon\,\text{OPT} + \epsilon\,\text{OPT} = (1+6\epsilon)\,\text{OPT}.$$

Thus, we have the approximation guarantee in Theorem 7. The running time follows directly from Lemma 10 and $T = \tilde{O}(n/\epsilon)$.

---

$^2$    More accurately, the telescoping works on

$$-\alpha_k f_\mu(u^*) \leq 8nL\alpha_{k-1} \mathbb{E}_{I_{k-1}}[f_\mu(y_{k-1})] - \mathbb{E}_{I_k}[8nL\alpha_k f_\mu(y_k^{(i)})] + \mathbb{E}_{I_{k-1}}[V_{z_{k-1}}(u^*)] - \mathbb{E}_{I_k}[V_{z_k^{(i)}}(u^*)].$$

where $I_k$ is all the random coordinate choices made through the first iteration till $k$-th iteration. The final expectation on $f_\mu(y_T)$ is over all the $T$ random choices.

## References

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

**2** Zeyuan Allen-Zhu and Lorenzo Orecchia. Using optimization to break the epsilon barrier: A faster and simpler width-independent algorithm for solving positive linear programs in parallel. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1439–1456, 2015.

**3** Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012. URL: `http://www.theoryofcomputing.org/articles/v008a006`, `doi:10.4086/toc.2012.v008a006`.

**4** Baruch Awerbuch and Rohit Khandekar. Stateless distributed gradient descent for positive linear programs. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 691–700, 2008. URL: `http://doi.acm.org/10.1145/1374376.1374476`, `doi:10.1145/1374376.1374476`.

**5** Lisa Fleischer. A fast approximation scheme for fractional covering problems with variable upper bounds. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004, New Orleans, Louisiana, USA, January 11-14, 2004*, pages 1001–1010, 2004. URL: `http://dl.acm.org/citation.cfm?id=982792.982942`.

**6** Christos Koufogiannakis and Neal E. Young. A nearly linear-time PTAS for explicit fractional packing and covering linear programs. *Algorithmica*, 70(4):648–674, 2014. URL: `http://dx.doi.org/10.1007/s00453-013-9771-6`, `doi:10.1007/s00453-013-9771-6`.

**7** Michael Luby and Noam Nisan. A parallel approximation algorithm for positive linear programming. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, May 16-18, 1993, San Diego, CA, USA*, pages 448–457, 1993. URL: `http://doi.acm.org/10.1145/167088.167211`, `doi:10.1145/167088.167211`.

**8** Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. URL: `http://dx.doi.org/10.1137/S1052623403425629`, `doi:10.1137/S1052623403425629`.

**9** Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005. URL: `http://dx.doi.org/10.1007/s10107-004-0552-5`, `doi:10.1007/s10107-004-0552-5`.

**10** Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. URL: `http://dx.doi.org/10.1137/100802001`, `doi:10.1137/100802001`.

**11** Serge A. Plotkin, David B. Shmoys, and Éva Tardos. Fast approximation algorithms for fractional packing and covering problems. In *32nd Annual Symposium on Foundations of Computer Science, San Juan, Puerto Rico, 1-4 October 1991*, pages 495–504, 1991. URL: `http://dx.doi.org/10.1109/SFCS.1991.185411`, `doi:10.1109/SFCS.1991.185411`.

**12** James Renegar. Efficient first-order methods for linear programming and semidefinite programming. *CoRR*, abs/1409.5832, 2014. URL: `http://arxiv.org/abs/1409.5832`.

**13** Neal E. Young. Sequential and parallel algorithms for mixed packing and covering. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 538–546, 2001. URL: `http://dx.doi.org/10.1109/SFCS.2001.959930`, `doi:10.1109/SFCS.2001.959930`.

**14** Neal E. Young. Nearly linear-time approximation schemes for mixed packing/covering and facility-location linear programs. *CoRR*, abs/1407.3015, 2014. URL: `http://arxiv.org/abs/1407.3015`.

**15**    Zeyuan Allen Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *CoRR*, abs/1407.1537, 2014. URL: `http://arxiv.org/abs/1407.1537`.