Computational Guarantees for Restarted PDHG for LP based on "Limiting Error Ratios" and LP Sharpness

Zikai Xiong[*] Robert M. Freund[†]

2025-05-29

Abstract

In recent years, there has been growing interest in solving linear optimization problems – or more simply "LP" – using first-order methods in order to avoid the costly matrix factorizations of traditional methods for huge-scale LP instances. The restarted primal-dual hybrid gradient method (PDHG) – together with some heuristic techniques – has emerged as a powerful tool for solving huge-scale LPs. However, the theoretical understanding of the restarted PDHG and the validation of various heuristic implementation techniques are still very limited. Existing complexity analyses have relied on the Hoffman constant of the LP KKT system, which is known to be overly conservative, difficult to compute (and hence difficult to empirically validate), and fails to offer insight into instance-specific characteristics of the LP problems. These limitations have limited the capability to discern which characteristics of LP instances lead to easy versus difficult LP instances from the perspective of computation. With the goal of overcoming these limitations, in this paper we introduce and develop two purely geometry-based condition measures for LP instances: "limiting error ratio" and LP sharpness. We provide new computational guarantees for the restarted PDHG based on these two condition measures. For limiting error ratio, we provide a computable upper bound and show its relationship with the data instance's proximity to infeasibility under perturbation. For LP sharpness, we prove its equivalence to the stability of the LP optimal solution set under perturbation of the objective function. We validate our computational guarantees in terms of these condition measures via specially constructed instances. Conversely, our computational guarantees validate the practical efficacy of certain heuristic techniques (row preconditioners and step-size tuning) that improve computational performance in practice. Finally, we present computational experiments on LP relaxations from the MIPLIB dataset that demonstrate the promise of various implementation strategies.

1 Introduction, Motivations, and Main Results

The focus of this paper is on solving huge-scale instances of linear optimization problems – or more simply "LP". LP problems abound across a wide variety of applications from manufacturing, transportation, service sciences, to computational science and engineering. Up until very recently, the most successful methods for solving LP problems have been simplex and pivoting methods [13] and interior-point methods [58]; these methods have been extensively studied and implemented in state-of-the-art commercial solvers [21]. In most cases, they are able to obtain a high-accuracy solution, but the success of these methods relies on repeatedly solving a linear system in each

[*] MIT Operations Research Center, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. zikai@mit.edu. Research supported by AFOSR Grant No. FA9550-22-1-0356.

[†] MIT Sloan School of Management, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. rfreund@mit.edu. Research supported by AFOSR Grant No. FA9550-22-1-0356.

iteration. For LP instances of a huge scale, the matrix factorizations required for solving the linear systems can be prohibitively costly. Moreover, the matrix factorizations are often unable to exploit the natural sparsity of a given LP instance and can have prohibitively large memory requirements. In contrast, first-order methods (FOMs) – and in particular the primal-dual hybrid gradient method (PDHG) [11] – are emerging as an alternative for solving huge-scale LP problems because they do not require the repeated solution of linear equations nor do they impose large memory requirements, thus reducing per-iteration costs. The primary task within each iteration of a FOM for LP is the gradient computation, which typically only requires matrix-vector multiplications (and so can fully take advantage of the sparsity of the LP instance). Moreover, FOMs are more suitable for distributed and parallel computation, and can benefit from modern computational architectures that accelerate computation through distributed systems and graphics processing units (GPUs). And indeed this compatibility with modern hardware architectures underscores the growing importance of FOMs for solving larger-scale LP instances.

Perhaps the best-known implementation of an FOM for solving LP is the solver PDLP [1], which is based on the primal-dual hybrid gradient method (PDHG) [11] to solve the saddlepoint formulation of LP. In the experiments reported in [1], PDLP was able to outperform the commercial solver Gurobi when the LP problem was large-scale. A recent GPU implementation of the PDLP further outperforms traditional algorithms implemented in the state-of-art commercial solvers on more LP instances [36]. Furthermore, [44] presents a distributed version of PDLP that is used to solve practical LP problems with 92 billion non-zeros in the constraint matrix – which is way beyond the capability of any simplex or interior-point method. PDLP is based on PDHG [11] – often referred to as the Chambolle-Pock method. PDHG is an operator-splitting method with alternating updates between the primal and dual variables. On top of running the base algorithm PDHG, schemes for restarting PDHG have also been proven in theory to help PDHG achieve faster linear convergence [3, 35] and this theory has yielded impressive speed-ups in practice as well. And in addition to using restarts, PDLP also utilizes various heuristic techniques such as presolving, row preconditioning, and step-size tuning [1].

We work with LP in the standard form:

$$\min_{x\in\mathbb{R}^n} \ c^\top x \quad \text{s.t. } Ax = b,\ x \geq 0 \ , \tag{1.1}$$

where the constraint matrix $A \in \mathbb{R}^{m\times n}$, the right-hand side vector $b \in \mathbb{R}^m$, and the objective vector $c \in \mathbb{R}^n$, whose standard dual problem is:

$$\max_{y\in\mathbb{R}^m} \ b^\top y \quad \text{s.t. } A^\top y \leq c \ . \tag{1.2}$$

The problem (1.1) can also be expressed as the saddlepoint problem:

$$\min_{x\in\mathbb{R}^n_+} \max_{y\in\mathbb{R}^m} L(x,y) := c^\top x + b^\top y - x^\top A^\top y \tag{1.3}$$

where $L(x,y)$ is the Lagrangian function and y is the vector of multipliers on the equation system $Ax = b$, where exchanging the min and max operations in (1.3) leads to (1.2). In this paper we let $\mathcal{X}^\star$ and $\mathcal{Y}^\star$ denote the optimal solution sets of (1.1) and (1.2), let $\mathcal{Z}^\star := \mathcal{X}^\star \times \mathcal{Y}^\star$ denote the solution set of the saddlepoint problem (1.3), and let $z := (x,y) \in \mathbb{R}^{m+n}$ denote the combined primal/dual iterates.

The family of PDHG algorithms (using various step-sizes, and with/without overlaid restart schemes) is designed to directly tackle the saddlepoint problem (1.3) rather than the original problem

(1.1) or/and its dual (1.2). One step of PDHG for (1.3) at the point $z = (x, y)$ is defined as follows:

$$z^+ = \text{PDHG}\textsc{step}(z) := \begin{cases} x^+ := P_{\mathbb{R}^n_+}\left(x - \tau\left(c - A^\top y\right)\right) \\ y^+ := y + \sigma\left(b - A\left(2x^+ - x\right)\right) \end{cases}, \tag{1.4}$$

in which τ and σ are the primal and dual step-sizes, respectively, and $P_{\mathbb{R}^n_+}$ is the projection operator onto the non-negative orthant $\mathbb{R}^n_+$ (which is the computationally trivial task of taking the nonnegative parts of the components). The vanilla PDHG tackles LP by generating iterates according to: $z^{k+1} \leftarrow \text{PDHG}\textsc{step}(z^k)$ for $k = 0, 1, 2, \ldots$. PDHG with restarts, which is denoted by rPDHG, is a variant of PDHG that regularly restarts PDHG using the average of the previous ℓ iterates where ℓ is chosen according to some rule, see Section 3 for a detailed description of rPDHG.

In this paper we seek to more deeply understand the performance of rPDHG applied to LP problems, both in theory and in practice, and to improve the theory where possible, as well as to explore the extent to which the theory is aligned with computational practice of rPDHG. The starting point of our work is the algorithm and analysis of rPDHG in the paper [3], which contains many new and important ideas both in terms of methods and analysis, and whose main results for rPDHG we now attempt to summarize in a brief and cogent manner. Along with other primal-dual methods, [3] analyzes rPDHG for solving (1.3). Let the iterates of rPDHG be denoted by $z^k = (x^k, y^k)$. Taken together, Theorems 1 and 2 of [3] state that rPDHG requires at most

$$O\left(\frac{\|A\|}{\alpha} \cdot \ln\left(\frac{\|A\|}{\alpha\varepsilon}\right)\right) \tag{1.5}$$

iterations in order to obtain an iterate z^k for which $\text{Dist}(z^k, \mathcal{Z}^\star) \leq \varepsilon$, where the notation $\text{Dist}(z, \mathcal{Z})$ denotes the Euclidean distance from a point z to the set $\mathcal{Z}$. Here the notation $O(\cdot)$ hides only absolute constants, $\|A\|$ is the spectral norm of A, and α is a positive scalar related to the sharpness [51, 9] of a particular functional called the "normalized duality gap" function, that we now describe. The normalized duality gap function is denoted $\rho(r; z)$ and is defined parametrically for a given positive "radius" r as:

$$\rho(r; z) := \left(\frac{1}{r}\right) \max_{\hat{z} \in \widetilde{B}(r;z)} \left[L(x, \hat{y}) - L(\hat{x}, y)\right], \tag{1.6}$$

where $z = (x, y) \in \mathbb{R}^n_+ \times \mathbb{R}^m$, $\widetilde{B}(r; z) := \{\hat{z} := (\hat{x}, \hat{y}) : \hat{x} \geq 0 \text{ and } \|\hat{z} - z\|_M \leq r\}$, and the norm $\|\cdot\|_M$ is a carefully selected matrix norm constructed using A and the step-size parameters of PDHG, where $M = \begin{pmatrix} I & -\eta A^\top \\ -\eta A & I \end{pmatrix}$ in the simple case when the primal and dual step-sizes of PDHG are identically equal to η and $\eta \leq (1/\lambda^+_{\max}(A))$. The scalar α in (1.5) is related to the "sharpness" of $\rho(r; z)$, which we now describe as well. As developed in [51] and extended in [9], the sharpness of a function f essentially measures how fast f grows away from its optimal solution set, and we say that f is β-sharp if $f(v) - f^\star \geq \beta \cdot \text{Dist}(v, \mathcal{V}^\star)$ for any v, where $\mathcal{V}^\star$ is the set of minimizers of f, and $f^\star$ is the minimum value of f. The quantity α in (1.5) is related to these sharpness notions applied to the normalized duality gap function and is defined to be a constant that satisfies

$$\rho(r^k; z^k) \geq \alpha \cdot \text{Dist}(z^k, \mathcal{Z}^\star) \tag{1.7}$$

for all iterates z^k of the algorithm. (Note that this is a bit weaker than the actual sharpness of $\rho(r; \cdot)$ as it only needs to hold for the iterates z^k.) Here the radius parameter r^k depends on the iteration k and is dynamically and adaptively defined by the algorithm's iterates, see [3] for the precise details. In summary, if there exists a positive value α for which (1.7) holds for all iterates z^k of rPDHG, then the algorithm has an overall iteration complexity bound given by (1.5).

Furthermore, Property 3 and Lemma 5 of [3] show that (1.7) always holds for

$$\alpha = \frac{1}{\mathcal{H}(K)\sqrt{1+16\,\mathrm{Dist}(0,\mathcal{Z}^\star)^2}}\ , \tag{1.8}$$

where $\mathcal{H}(K)$ is the Hoffman constant[1] [23] of the matrix K of the Karush-Kuhn-Tucker linear inequality system that defines the optimal solution set, namely

$$K := \begin{pmatrix} I & -A^\top & A^\top & 0 & -c \\ 0 & 0 & 0 & -A & b \end{pmatrix}^\top .$$

Combining (1.5) with (1.8) one also obtains the following iteration bound for rPDHG for LP:

$$O\left(\|A\| \cdot \mathcal{H}(K) \cdot \max\{1, \mathrm{Dist}(0,\mathcal{Z}^\star)\} \cdot \ln\left(\|A\| \cdot \mathcal{H}(K) \cdot \max\{1, \mathrm{Dist}(0,\mathcal{Z}^\star)\} \cdot \frac{1}{\epsilon}\right)\right)\ . \tag{1.9}$$

1.1 Motivating Issues

The rPDHG algorithm and iteration bounds (1.5) and/or (1.9) in [3] are quite significant in at least several ways, including but not limited to the algorithm design (the restart scheme for rPDHG both theoretically and practically), the proof of linear convergence of rPDHG, the use of and the development of properties of the normalized duality gap function $\rho(r;z)$, and the use of the Hoffman constant $\mathcal{H}(K)$ of the KKT system matrix K to bound the constant α in (1.7).

Nevertheless, there are certain issues with the bounds (1.5) and/or (1.9) that are not very satisfactory, and which we seek to overcome. One issue has to do with the reliance on the sharpness constant α of the normalized duality gap function $\rho(r;z)$ in (1.5). The normalized duality gap function $\rho(r;z)$ itself is not such a natural metric of LP behavior, as there is (at best) only a partial equivalence between $\rho(r;z)$ and more typical metrics and stopping criteria that are used in LP solvers such as primal and dual infeasibility and non-optimality measures. Also, the definition of $\rho(r;z)$ depends on the step-sizes of the algorithm through the matrix M. Therefore the sharpness of $\rho(r;z)$ and consequently the value of α depend – at least partially – on the magnitude and ratio of the primal and dual step-sizes of the algorithm and hence depend on more than just the intrinsic properties of the LP. It is thus unclear from the iteration bound (1.5) what natural/intrinsic properties of the LP problem itself (such as geometric properties, data-perturbation metrics, error bounds, etc.) contribute to the performance – theoretically or practically – of rPDHG.

Another issue concerns the iteration bound (1.9). This bound replaces α by two other metrics, namely $\mathrm{Dist}(0,\mathcal{Z}^\star)$ and $\mathcal{H}(K)$. The reliance on $\mathrm{Dist}(0,\mathcal{Z}^\star)$, which measures the norms of primal and dual optimal solutions, seems both natural and appropriate, as in the very least $\mathrm{Dist}(0,\mathcal{Z}^\star)$ bounds changes in optimal solution values under perturbations of b and c and thus is tied to general notions of condition number theory for LP much more broadly, see [52]. However, the reliance in (1.9) on the Hoffman constant $\mathcal{H}(K)$ of the KKT system matrix K is not desirable for a number of reasons. For one, the Hoffman constant of a matrix W is typically extremely large by definition, as it accounts for largest error bound of every linear inequality system $\mathcal{V}_g = \{v : Wv \le g\}$ over all possible right-hand side vectors g for which $\mathcal{V}_g \neq \emptyset$. Furthermore, it is typically an extremely conservative measure since in academic applications one is typically only concerned with the error bound for a single given g. And in the KKT system that single g is given by $(0,-b,b,-c,0)^\top$ and

[1] The Hoffman constant of a matrix W is a global error bound that bounds the distance of any point v to a non-empty set of solutions of a system of linear inequalities $Wv \le g$ in terms of the norm of the residual vector $\|[Wv-g]^+\|$, see [23].

thus has its own special structure by itself and also is tied to the data b and c which appear in the matrix K as well.

A third issue also concerns the non-local nature of the error bounds embedded in the Hoffman constant. If the primal or the dual feasible region has an extreme point whose active constraint system is badly ill-conditioned, then $\mathcal{H}(K)$ will be very large, even if this extreme point is not related at all to the optimal solution in terms of active constraints and/or distance to the optima or objective function value. Indeed, it would be better to have an iteration bound that does not depend so globally on properties of all points in the primal and the dual feasible sets.

The fourth issue with the iteration bounds (1.5) and/or (1.9) has to do with computability in order to test whether or not the bounds align with computational practice. From both the practical and theoretical perspectives, it is important to ask whether an iteration bound aligns with computational practice. In other words, when applied to problems that arise in practice do problems with smaller sharpness constant values α require more iterations of rPDHG than problems with larger such values of α?, and similarly for $\mathcal{H}(K)$? In order to study these questions one has to be able to actually compute (or approximately compute) α and/or $\mathcal{H}(K)$. However, it is not known (at least by us) how to compute α for LP problems in general. And regarding the Hoffman constant, using the results in [50] we know that $\mathcal{H}(K)$ has the following characterization:

$$\mathcal{H}(K) := \max_{\substack{J\subseteq\{1,\dots,2m+2n+1\}\\ K_J \text{ has full row rank}}} \frac{1}{\min_{v\in\mathbb{R}^J_+,\|v\|=1}\left\|K_J^\top v\right\|} \ ,$$

where K_J denotes the submatrix of K formed by selecting the rows indexed by J. Even with this novel characterization, it is still a very difficult task to compute (or even just reliably estimate) $\mathcal{H}(K)$ as it requires enumerating exponentially many submatrices [50]. Thus neither of the bounds (1.5) nor (1.9) are amenable to testing the extent to which they might align with computational practice.

To illustrate the importance of alignment between theory and practice, consider the following extremely simple LP instance with $m = 1$ and $n = 2$, for which computing $\mathcal{H}(K)$ is doable:

$$\min_{(x_1,x_2)\in\mathbb{R}^2_+} \ \cos(\gamma)\cdot x_1 - \sin(\gamma)\cdot x_2 \quad \text{s.t.} \ \sin(\gamma)\cdot x_1 + \cos(\gamma)\cdot x_2 = 1 \tag{LP$_\gamma$}$$

for $\gamma \in (0, \pi/2)$, which is illustrated in the left subfigure of Figure 1. In the subfigure the feasible set is the blue line segment whose distance to $(0, 0)$ is always 1, the optimal solution is the red point $(x_1^\star, x_2^\star) = (0, 1/\cos(\gamma))$, and the direction of the objective vector $c = [\cos(\gamma), -\sin(\gamma)]$ is denoted by the red dashed arrow. The right subfigure shows the values of the Hoffman constant $\mathcal{H}(K)$ of the KKT system, the iteration bound (1.9) based on [3], and the actual iteration count of rPDHG for (LP$_\gamma$) for $\gamma \in (0, \pi/2)$. We see that as $\gamma \searrow 0$, (LP$_\gamma$) becomes more ill-conditioned in terms of $\mathcal{H}(K)$ and this is reflected in the iteration bound (1.9). However, this family of LP instances is very easy for rPDHG to solve for arbitrarily small values of γ. It would be better to have an iteration bound that is more aligned with actual computational practice. (The blue line in the right subfigure is a spoiler: it shows the bound that we develop in Theorem 3.1 of this paper, which for this family of problems is well-aligned with actual iteration counts.) Details of this experiment are presented in Section 6.1.

1.2 New computational guarantees based on "Limiting Error Ratios" and LP Sharpness

We now describe our new computational guarantees. Consider the original LP primal problem (1.1), and let $\mathcal{F}_p$ denote the feasible set, defined as the intersection of the nonnegative orthant $\mathbb{R}^n_+$ and the

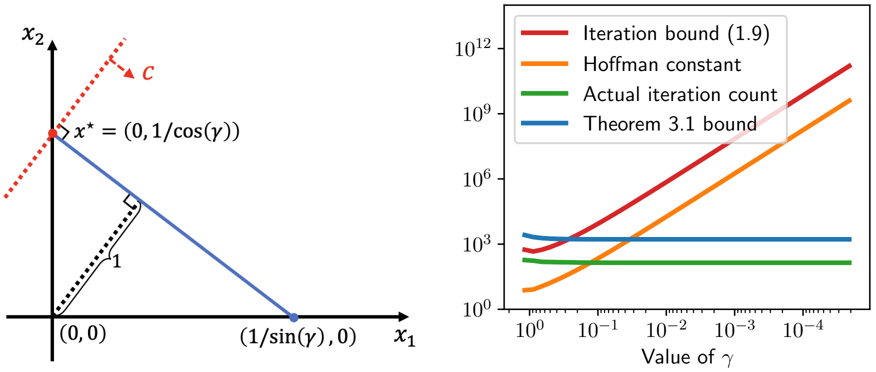


Figure 1: (left) The feasible set (blue line) and the optimal solution (red point) of (LP$_\gamma$). (right) The values of the theoretical iteration bound (1.9) based on [3], the Hoffman constant $\mathcal{H}(K)$ of the KKT system, and the actual iteration count of rPDHG to achieve $z = (x, y)$ for which $\mathrm{Dist}(z, \mathcal{Z}^\star) \le 10^{-10}$. The blue line shows our new bound in Theorem 3.1 of this paper.

affine subspace $V_p := \{x \in \mathbb{R}^n : Ax = b\}$, namely $\mathcal{F}_p = V_p \cap \mathbb{R}^n_+$. The primal optimal solution set is denoted by $\mathcal{X}^\star$. Our new computational guarantees involve two types of condition measures. The first condition measure is denoted by $\theta_p^\star$, and is called the "limiting error ratio," or "LimitingER" for short, which we now define.

Definition 1.1 (Error ratio and limiting error ratio)**.** *For any $x \in V_p \setminus \mathcal{F}_p$ (namely x satisfies the linear equality constraints but lies outside the nonnegative orthant and is thus infeasible), the error ratio (ER) of $\mathcal{F}_p$ at x is defined as:*

$$\theta_p(x) := \frac{\mathrm{Dist}(x, \mathcal{F}_p)}{\mathrm{Dist}(x, \mathbb{R}^n_+)} \ , \tag{1.10}$$

which is the ratio of the distance to the feasible set $\mathcal{F}_p$ to the distance to the nonnegative orthant. Let $\theta_p(x) := 1$ for $x \in \mathcal{F}_p$ for notational completeness. The limiting error ratio (LimitingER) is the quantity $\theta_p^\star$ defined as:

$$\theta_p^\star := \lim_{\varepsilon \to 0} \left(\sup_{x \in V_p,\, \mathrm{Dist}(x, \mathcal{X}^\star) \le \varepsilon} \theta_p(u) \right) \ , \tag{1.11}$$

which is the supremum of $\theta_p(x)$ for all $x \in V_p$ approaching $\mathcal{X}^\star$.

The measure $\theta_p^\star$ (and also $\theta_p(x)$) is similar to error bounds proposed in the literature such as those in [48, 28]. We call $\theta_p(x)$ an *error ratio* because it represents the ratio of two types of errors: the distance to the feasible region and the distance to the nonnegative orthant $\mathbb{R}^n_+$. For any $x \in \mathcal{F}_p \setminus \mathbb{R}^n_+$ it always holds that $\theta_p(x) \ge 1$ because the numerator in (1.10) is always at least as large as the denominator (since $\mathcal{F}_p \subset \mathbb{R}^n_+$). Indeed, many papers have studied different formulations of global upper bounds for the error ratios of linear inequality systems from different perspectives, starting with the celebrated Hoffman bound [23], including [43, 6, 42, 19] among many others, see [48] for a comprehensive survey of relevant results. Note that $\theta_p^\star$ is also bounded above by a Hoffman bound $\mathcal{H}(\bar{A})$ on the linear inequality system $Ax \le b$, $-Ax \le -b$, $-x \ge 0$, where $\bar{A} := [A^\top \ -A^\top \ -I]^\top$, because $\theta_p(x) \le \mathcal{H}(\bar{A})$ for all x. However, unlike $\mathcal{H}(\bar{A})$, $\theta_p^\star$ only depends on local information about the feasible set $\mathcal{F}_p$ near $\mathcal{X}^\star$, and hence $\mathcal{H}(\bar{A})$ is likely to be an excessively conservative bound on $\theta_p^\star$, since (i) $\mathcal{H}(\bar{A})$ is a global bound whereas $\theta_p^\star$ is a local bound, and (ii) x must satisfy $x \in V_p$ in (1.11) of Definition 1.1 and only allows zero error in the system $Ax = b$.

The second condition measure is the LP sharpness and is denoted by μ_p.

Definition 1.2 (LP sharpness)**.** *Let $H_p^\star$ denote the optimal objective hyperplane, i.e., $H_p^\star := \{\hat{x} \in \mathbb{R}^n : c^\top \hat{x} = c^\top x^\star\}$ for an $x^\star \in \mathcal{X}^\star$. The LP sharpness μ_p of the primal problem is defined as:*

$$\mu_p := \inf_{x \in \mathcal{F}_p \setminus \mathcal{X}^\star} \frac{\mathrm{Dist}(x,\ V_p \cap H_p^\star)}{\mathrm{Dist}(x,\ \mathcal{X}^\star)}\ , \tag{1.12}$$

which is the infimum of the ratio between the distance from x to $V_p \cap H_p^\star$ and the distance from x to the optimal set $\mathcal{X}^\star$.

Intuitively speaking, the LP sharpness measures how quickly the objective function grows away from the optimal solution set $\mathcal{X}^\star$ among all feasible points. In the case of LP it is easy to see that $\mu_p > 0$.

Notions of sharpness were perhaps first introduced by Polyak in [51] as a useful analytical tool in convex minimization. For example, sharpness plus some mild smoothness assumptions can lead to linear convergence of the subgradient descent method via the use of restarts, see [63]. [3] generalizes the sharpness concept from convex optimization to primal-dual saddlepoint problems by defining sharpness for the normalized duality gap functional (1.6). Here we apply the notion of sharpness directly (and naturally) to the LP optimization problem itself.

The two condition measures $\theta_p^\star$ and μ_p are defined for the primal problem in the nonnegative ("cone") variable x. We define analogous condition measures for the dual problem in terms of the slack variables s of the dual constraints, namely the variables $s := c - A^\top y$. Specifically, let $\mathcal{S}^\star$ denote the set of optimal slack variable values in the dual LP. Then it is straightforward to define dual counterparts $\theta_d^\star$ and μ_d of the primal condition measures $\theta_p^\star$ and μ_p; see Section 3 for the formal definitions. Additionally, we let κ denote the standard condition number of A, which is defined to be the ratio the largest to the smallest positive singular value of A.

The main result of this paper is a new computational guarantee for rPDHG that depends on the above condition measures, which we now describe. We suppose that c satisfies $Ac = 0$ (which can be easily enforced by projecting c onto the nullspace of A as part of a presolve scheme), and we suppose that the step-sizes τ and σ are chosen in a certain way that is described in detail in Section 3. Our main result (Theorem 3.1) is as follows:

Main Result 1. *(Less formal restatement of Theorem 3.1) Under the above conditions rPDHG requires at most*

$$O\left(\mathcal{N} \cdot \ln\left[\frac{\mathcal{N}\mathcal{D}\varepsilon_0}{\varepsilon}\right]\right) \tag{1.13}$$

iterations of PDHGstep *in order to compute a pair (x, s) of primal solution and dual slack that satisfies* $\max\{\mathrm{Dist}(x, \mathcal{X}^\star), \mathrm{Dist}(s, \mathcal{S}^\star)\} \le \varepsilon$*, where $\mathcal{N}$ and $\mathcal{D}$ are defined as follows:*

$$\mathcal{N} := \kappa \cdot \left(\frac{1}{\mu_p} + \frac{1}{\mu_d}\right)\left(\theta_p^\star + \theta_d^\star + \frac{\|x^\star\|}{\|b\|_Q} + \frac{\|s^\star\|}{\|c\|}\right) \tag{1.14}$$

and

$$\mathcal{D} := 32e \cdot \kappa \cdot \max\left\{\frac{\|c\|}{\|b\|_Q}, \frac{\|b\|_Q}{\|c\|}\right\}\ , \tag{1.15}$$

and $\varepsilon_0 := \max\{\mathrm{Dist}(x^0, \mathcal{X}^\star), \mathrm{Dist}(s^0, \mathcal{S}^\star)\}$ *is the error of the initial iterate (x^0, s^0) measured using the distance to the set of optimal solution. Here $x^\star$ and $s^\star$ are the least norm optimal primal solution and dual slack solution, respectively. Also $\|b\|_Q$ denotes $\|A^\top(AA^\top)^\dagger b\|$, in which $Q := (AA^\top)^{-1}$ if A has full row rank. The quantity e is the base of the natural logarithm.*

Let us now examine the components of the iteration bound in Main Result 1 a bit closer. First notice that the ratio of the initial error to the target error $\varepsilon_0/\varepsilon$ appears inside the logarithm term and is a consequence of the global linear convergence property of the algorithm. The quantity $\mathcal{N}$ appears both inside and outside of the logarithm term and itself involves the condition number κ of the matrix A, plus six other quantities. These are μ_p and μ_d (the LP sharpness for both the primal and dual problems), $\theta_p^\star$ and $\theta_d^\star$ (the LimitingER for both the primal and dual problems), as well as $\frac{\|x^\star\|}{\|b\|_Q}$ and $\frac{\|s^\star\|}{\|c\|}$. Notice that higher values of μ_p^{-1}, μ_d^{-1}, $\theta_p^\star$, and/or $\theta_d^\star$ result in a higher value of $\mathcal{N}$. Also notice that all four cross-terms between primal and dual LimitingERs and the reciprocals of the LP sharpness are present in $\mathcal{N}$, as well as all four cross-terms between $\frac{\|x^\star\|}{\|b\|_Q}$, $\frac{\|s^\star\|}{\|c\|}$ and the reciprocals of the LP sharpness. The numerator in $\frac{\|x^\star\|}{\|b\|_Q}$ is the norm of the least-norm primal optimal solution, which measures the stability of the dual problem under perturbation of c. In Fact 2.1 of Section 2 we show that the denominator $\|b\|_Q$ is equal to $\operatorname{Dist}(0, V_p)$, which is the distance from 0 to V_p, and so $\|b\|_Q$ can be interpreted as a lower bound on the norm of any (and every) feasible or optimal solution of the primal. Therefore the quotient $\frac{\|x^\star\|}{\|b\|_Q}$ is equal to the geometric measure $\frac{\operatorname{Dist}(0,\mathcal{X}^\star)}{\operatorname{Dist}(0,V_p)}$ and can be interpreted as a relative measure of stability or relative distance to optima. A similar interpretation holds for $\frac{\|s^\star\|}{\|c\|}$.

We note that $\mathcal{N}$ is the only quantity outside of the logarithm term in the iteration bound and hence is the quantity characterizing the rate of linear convergence. We will shortly focus on $\mathcal{N}$ in our discussion and comments.

The quantity $\mathcal{D}$, which appears only in the logarithm term of the iteration bound, is a (positive) scale-invariant measure of the problem data, and is less important since it only appears inside the logarithm term in (1.13). The proofs leading to Main Result 1 actually use a similar strategy as in [3]. Roughly speaking, these proofs proceed by showing that a sharpness condition that is different from (1.7) by the norm holds for $\alpha = \frac{1}{\mathcal{N}}$ for all iterates, see Lemma 3.4 in Section 3.

1.3 Remarks on Main Result 1

Geometric nature of the new iteration bound. The iteration bound (1.13) in Main Result 1, as well as the specific quantity $\mathcal{N}$ that is outside the logarithm term, essentially relies on the primal and dual LimitingER values $\theta_p^\star$ and $\theta_d^\star$, the primal and dual LP sharpness values μ_p and μ_d, and the minimum norms of solutions $x^\star$ and $s^\star$, as well as on data metrics involving the condition number κ of A and other norms of the data (A, b, c). The primal and dual LimitingER values and the LP sharpness values are all geometric in nature and (in our view) arise naturally in the study of the behavior or "conditioning" of an LP instance at its optima. Recall that $\theta_p^\star$ and $\theta_d^\star$ are local supremums of the error ratios $\theta_p(x)$ and $\theta_d(s)$ near $\mathcal{X}^\star$ and $\mathcal{S}^\star$, and μ_p and μ_d quantify the sharpness of the objective function near the optimal solution set. In this regard $\theta_p^\star$, $\theta_d^\star$, μ_p, and μ_d depend only on the optimal solution set and an arbitrarily small neighborhood around the optimal solution set. Especially because the bound in Main Result 1 is described only by local information around the optima, it is likely to be significantly tighter than (1.9) which depends on the Hoffman constant $\mathcal{H}(K)$. (An ill-conditioned basis at a non-optimal extreme point may significantly increase $\mathcal{H}(K)$, yet it will have no impact on $\mathcal{N}$.)

Dependence only on local behavior near the optima. Because the iteration bound in Main Result 1 only depends on local behavior near the optimal solution set, it therefore shows that the performance of PDHG – at least in theory – is only tied to local properties of the LP instance local to the optimal solution set.

Computability of the bound in Main Result 1. From both a practical and theoretical perspective, it is important to ask whether the bound in Main Result 1 aligns with computational practice. In other words, when applied to problems that arise in practice do problems with smaller values of (1.13) require fewer iterations of rPDHG than problems with larger such values? (More generally it is important to ask this question for any algorithm for any optimization problem; however it is all the more important for LP given its pervasive use in practice, and it is important for rPDHG in order to better understand whether/where one might discover improvements in the theory or in practice in this nascent stage.) In order to answer this and related questions it is necessary to efficiently compute the component quantities involved in (1.13). The quantities $\|b\|_Q$, $\|c\|$, and κ are not difficult to compute (or estimate with high accuracy). Also $\|x^\star\|$ and $\|s^\star\|$ are readily computable once an optimal solution has been computed. Hence the challenge in computing the bound in Main Result 1 lies in computing $\theta_p^\star$ and μ_p and their analogous dual quantities. Let us first consider $\theta_p^\star$. While we have not uncovered a direct way to compute $\theta_p^\star$ exactly, we have developed an efficient way to compute an upper bound $\theta_p^\star$ via the following result:

Property 1. (essentially Proposition 4.2.) *Suppose $x_a \in \mathcal{X}^\star$ and there exists R_a for which $\mathcal{X}^\star \subset \{x : \|x - x_a\| \le R_a\}$, then it holds that $\theta_p^\star \le G^\star$ for $G^\star$ defined as follows:*

$$G^\star := \inf_{r>0,\ x\in\mathbb{R}^n} \frac{R_a + \|x - x_a\|}{r} \quad \text{s.t. } \ x \in V_p,\ \ x \ge r \cdot e \ . \tag{1.16}$$

In order to compute $G^\star$ above one needs to know a ball containing the optimal solution set; if $\mathcal{X}^\star$ is a singleton then it is sufficient to know an optimal solution, and if there are multiple optima then the analytic center of $\mathcal{X}^\star$ can furnish such information, see Sonnevend [55], also [45]. Property 1 essentially states that $\theta_p^\star$ cannot be too large if (i) the radius of the optimal solution set of (1.1) is not too large, and (ii) there is a feasible solution x that is not too close to the boundary of $\mathbb{R}^n_+$ and not too far from the optimal solution set. Such a solution is related to the concept of a "reliable solution" in [15]. Similar results also hold for $\theta_d^\star$. This result is formally stated as the first assertion of Proposition 4.2 in Section 4, where we will also show that the optimization problem in (1.16) can be reformulated as a convex conic optimization problem with m linear equalities, $n + 1$ linear inequalities, and one second-order cone.

Let us now consider the computation of LP sharpness μ_p. Not surprisingly, there is a nice polyhedral characterization of LP sharpness μ_p that enables its computation, which we develop in Section 5. If the optimal solution set is a singleton, then the LP sharpness μ_p is the smallest objective function growth rate along all of the edges of $\mathcal{F}_p$ emanating from $x^\star$, which can be computed easily if the number of such edges is not excessive. For LP problems with multiple optimal solutions, computing μ_p requires computing the smallest sharpness along all edges of $\mathcal{F}$ that intersect $\mathcal{X}^\star$, which might be more challenging. A similar approach applies to computing μ_d. For details see Section 5.

Relation to other condition numbers. Especially since the condition measures LimitingER and LP sharpness play the central role in Main Result 1, it is useful to understand how they may be related to other more traditional condition measures for LP, such as Renegar's data-perturbation condition numbers [52]. It turns out that the LimitingER is upper-bounded by (and hence is tighter than) a simple quantity involving the data-perturbation condition number of Renegar [52], see Corollary 4.4. We also show that LP sharpness μ_p is related to the stability of $\mathcal{X}^\star$ under perturbation of the objective function vector c; in fact μ_p is equal to the least-norm relative perturbation Δc of c for which the new optimal solution set is not a subset of existing optimal solution set, see Theorem 5.1. Similar arguments also hold for μ_d.

Invariance under simple scalar rescaling. It has been observed in practice that, with proper choice of step-sizes, rPDHG's performance is invariant under the scalar rescaling $(\alpha A, \beta b, \gamma c)$ of the LP instance data (A, b, c) for $\alpha, \beta, \gamma > 0$, see [1]. Notice that this observation is in synch with the iteration bound in Main Result 1. To see this, notice that the quantity $\mathcal{N}$ in (1.14) is invariant under scalar rescaling, since in particular each of the condition measures – μ_p, μ_d, $\theta_p^\star$, and $\theta_d^\star$ is invariant under the rescaling, as is the matrix condition number κ. Likewise, because $\frac{\|x^\star\|}{\|b\|_Q}$ and $\frac{\|s^\star\|}{\|c\|}$ can be interpreted as the relative distances to optima, they are also invariant under the rescaling. Curiously, the quantity $\mathcal{D}$ (which only appears inside the logarithm term) is not invariant under rescaling; while the first term of $\mathcal{D}$ (which is κ) is scale invariant, the second term is always at least 1 and setting $\alpha = \beta = \gamma = \|b\|_Q / \|c\|$ results in $\mathcal{D} = \kappa$.

In addition to the above desirable features of the iteration bound in Main Result 1, the bound also suggests ways to think about practical enhancements of rPDHG to improve performance, in particular row-preconditioning of A as well as tuning the step-sizes τ and σ, which we now discuss.

Row-preconditioning of A. It has been observed in practice that heuristic row- and column-preconditioning of A can improve the practical performance of rPDHG, see [1]. The iteration bound in Main Result 1 provides a theoretical justification of the value of row-preconditioning as follows. Observe from Main Result 1 that the iteration bound is at least linear in the matrix condition number κ which appears outside the logarithm term (and inside the logarithm term as well). Now consider the row-preconditioned system $HAx = Hb$ for some rank-m matrix H. Replacing (A, b) with (HA, Hb) does not change the geometry of the primal x or dual s variables, and so leaves μ_p, μ_d, $\theta_p^\star$, and $\theta_d^\star$ invariant. However, it does change the value of κ, and thus heuristics to compute H that will reduce the value of κ will have the effect of reducing the theoretical iteration bound in Main Result 1. Therefore, row-preconditioning of A is a natural way to improve the performance of the algorithm – at least in theory. Furthermore, in Section 6 we explore and confirm the practical effect of row-preconditioning.

Tuning the ratio of primal and dual step-sizes. The theory for PDHG is premised on the primal and dual step-sizes τ and σ satisfying $\tau \cdot \sigma \le (\sigma_{\max}^+(A))^{-2}$ where $\sigma_{\max}^+(A)$ is the largest positive singular value of A, see [10, 3]. However, there is leeway in the ratio of the stepsizes; notice that for $\gamma > 0$ if we replace $(\tau, \sigma) \to (\gamma\tau, \sigma/\gamma)$ then the product $\tau \cdot \sigma$ is unchanged but the stepsize ratio τ/σ changes by γ^2. Furthermore, it has been observed in practice that tuning the ratio τ/σ can significantly improve the performance of rPDHG, see [1, 3]. Our analysis points to theoretical benefits in the iteration bound of rPDHG if the stepsize ratio is tuned in a special way. In Theorem 3.2 in Section 3 we show that a specially chosen step-size ratio leads to an iteration bound with a similar structure as in (1.13) but with $\mathcal{N}$ and $\mathcal{D}$ replaced by:

$$\widehat{\mathcal{N}} := \kappa \cdot \left(\frac{\theta_p^\star}{\mu_p} + \frac{\theta_d^\star}{\mu_d} + \frac{\|x^\star\|}{\mu_d \|b\|_Q} + \frac{\|s^\star\|}{\mu_p \|c\|} \right) \tag{1.17}$$

and $\widehat{\mathcal{D}} := 32e \cdot \kappa \cdot \max\left\{ \frac{\mu_p \|c\|}{\mu_d \|b\|_Q}, \frac{\mu_d \|b\|_Q}{\mu_p \|c\|} \right\}$. Since the quantity $\widehat{\mathcal{D}}$ only appears in the logarithm term, let us focus on the quantity $\widehat{\mathcal{N}}$ as compared to $\mathcal{N}$. Notice that $\widehat{\mathcal{N}}$ contains fewer cross-terms involving μ_p, μ_d, $\theta_p^\star$, $\theta_d^\star$, $\frac{\|x^\star\|}{\|b\|_Q}$ and $\frac{\|s^\star\|}{\|c\|}$, and so has the potential to be significantly smaller than $\mathcal{N}$. This lends theoretical credence to the value of tuning the step-size ratio in practical implementations of rPDHG. Indeed, we confirm the benefit of this strategy in our experiments in Section 6. The special stepsize formula that leads to this theoretical improvement is presented in equation (3.6) in Theorem 3.2. Notice that the stepsize formula in (3.6) involves the LP sharpness quantities μ_p and μ_d. Unfortunately, these two quantities are typically not known *a priori* nor are they easy to estimate, and for this reason the improved iteration bound involving $\widehat{\mathcal{N}}$ and $\widehat{\mathcal{D}}$ is essentially just theoretical in nature. Nevertheless, the improved bound points to the usefulness of heuristically tuning the step-size ratio in practical implementations of PDHG.

We end this section with a review of related works for large-scale LP.

1.4 Related works for large-scale LP

In addition to [1, 3] discussed earlier, there are several other investigations and analyses of the performance of PDHG and its variants for solving LP problems. [38] studies PDHG (without restarts) applied to LP instances, and uncovers a two-phase behavior of PDHG: the initial phase is characterized by sublinear convergence, followed by a second phase with linear convergence. The linear convergence of the latter phase is upper bounded using the Hoffman constant of a reduced linear system defined by the limiting point of the algorithm's trajectory. The duration of the initial phase inversely depends on the smallest nonzero of the limiting point. [34] introduces a stochastic variant of PDHG for solving LPs. [2] studies how to use PDHG for detecting infeasible LP instances, and [35] shows that the PDHG without restarts also achieves linear convergence on LPs, though at a slower rate compared to rPDHG. [22] shows that the rPHDG has polynomial-time complexity for totally unimodular LPs, and [37] proposes to solve convex QP using PDHG-based methods.

Concurrent with our own efforts in revising the present work, several more papers on PDHG have been posted on arXiv. From a computational perspective, recent efforts include improved implementations of PDHG-based LP solvers [41], other extensions to convex quadratic and conic optimization [25, 33], as well as [40, 62, 29]. From a theoretical perspective, recent works have provided refined analyses for special families of LPs [59, 39], average-case complexity guarantees [60], and convex conic optimization extensions [62].

In addition to PDHG, a number of other FOMs have also been studied for solving huge-scale LP instances. Early efforts included the steepest ascent method [8], feasible direction methods [64], the projected gradient algorithm [27], and others. More recently, several more practical FOM-based solvers have been proposed. ABIP [32, 14] solves conic linear programs (including linear programs) using an ADMM-based interior-point method applied to the homogeneous self-dual embedding. SCS [47, 46] employs a similar ADMM-based approach to solve the homogeneous self-dual embedding. OSQP [56, 54] uses an ADMM-based method to solve convex quadratic programs, which include LPs. HPR-LP [12] is a recently developed GPU-based solver for LP that uses the Halpern Peaceman-Rachford method with semi-proximal terms. [30] proposes a semismooth Newton augmented Lagrangian method for LP problems and proves its superlinear convergence. ECLIPSE [5] is a distributed LP solver designed specifically for addressing large-scale LPs encountered in web applications.

1.5 Notation

For a matrix $A \in \mathbb{R}^{m\times n}$, let $\mathrm{Null}(A) := \{x \in \mathbb{R}^n : Ax = 0\}$ denote the null space of A and $\mathrm{Im}(A) := \{Ax : x \in \mathbb{R}^n\}$ denote the image of A. For any set $\mathcal{X} \subset \mathbb{R}^n$, let $P_{\mathcal{X}} : \mathbb{R}^n \to \mathbb{R}^n$ denote the Euclidean projection onto $\mathcal{X}$, namely, $P_{\mathcal{X}}(x) := \arg\min_{\hat{x}\in\mathcal{X}} \|x - \hat{x}\|$. Unless otherwise specified, $\|\cdot\|$ denotes the Euclidean norm. For $M \in \mathbb{S}^n_+$, the set of symmetric positive-semi-definite matrices in $\mathbb{R}^{n\times n}$, we use $\|\cdot\|_M$ to denote the semi-norm $\|z\|_M := \sqrt{z^\top Mz}$. For any $x \in \mathbb{R}^n$ and $\mathcal{X} \subset \mathbb{R}^n$, the Euclidean distance between x and $\mathcal{X}$ is denoted by $\mathrm{Dist}(x, \mathcal{X}) := \min_{\hat{x}\in\mathcal{X}} \|x - \hat{x}\|$ and the M-norm distance between x and $\mathcal{X}$ is denoted by $\mathrm{Dist}_M(x, \mathcal{X}) := \min_{\hat{x}\in\mathcal{X}} \|x-\hat{x}\|_M$. For simplicity of notation, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. For $A \in \mathbb{R}^{n\times n}$, $A^\dagger$ denotes the Moore-Penrose inverse of A. For any matrix A, $\sigma^+_{\max}(A)$ and $\sigma^+_{\min}(A)$ denote the largest and smallest non-zero singular values of A. For an affine subset V, let $\vec{V}$ denote the associated linear subspace of V, namely $V = \vec{V} + v$ for every $v \in V$. Let $\mathbb{R}^n_+$ and $\mathbb{R}^n_{++}$ denote the nonnegative and strictly positive orthant in $\mathbb{R}^n$, respectively. Let e denote the vector of ones, namely $e = (1, \ldots, 1)^\top$ whose dimension is dictated by context. For a vector $v \in \mathbb{R}^n$, v^+ and v^- respectively denote the vector of positive parts and negative parts of v, i.e., the components of v^+ and v^- are $(v^+)_i = \max\{v_i, 0\}$ and $(v^-)_i = \max\{-v_i, 0\}$ for $i \in [n]$. The operator norm $\|A\|$ of a matrix A is defined as $\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}$. For a symmetric matrix A, $A \succeq 0$ means $A \in \mathbb{S}^n_+$. For a linear subspace $\vec{V} \subset \mathbb{R}^n$, $\vec{V}^\perp$ denotes the orthogonal complement of $\vec{V}$.

1.6 Organization

The other sections of this paper are organized as follows. Section 2 contains preliminaries LP and presents a detailed review of PDHG. In Section 3 we present our new computational guarantees for rPDHG for LP based on LimitingER and LP sharpness. In Sections 4 and 5 we discuss and derive computable upper bounds for LimitingER and a computable representation of LP sharpness, and we relate both of these condition measures to other condition numbers. Finally, in Section 6 we present computational experiments that give credence to both our theoretical iteration bounds and the effectiveness of heuristic enhancements inspired by these iteration bounds.

2 Preliminaries for LP and PDHG

As mentioned in Section 1.2, we use the lens of focusing on the nonnegative variables x of the primal and the nonnegative dual variables s (the slack variables) of the dual. We first review the "symmetric" formulation of primal and dual LP problems from this perspective.

2.1 Symmetric primal and dual formulations of LP

The dual problem (1.2) can be formulated with explicit slack variables s as follows:

$$\max_{y\in\mathbb{R}^m,\ s\in\mathbb{R}^n} \ b^\top y \quad \text{s.t. } A^\top y + s = c,\ s \geq 0\ . \tag{2.1}$$

Define $q := A^\top(AA^\top)^\dagger b$; then (2.1) is equivalent to the following (dual) problem on s:

$$\max_{s\in\mathbb{R}^n} \ q^\top(c - s) \qquad \text{s.t. } s \in c + \operatorname{Im}(A^\top),\ s \geq 0\ , \tag{2.2}$$

and the corresponding dual solution in the variable y is any such y for which $A^\top y = c - s$. (This is because for any dual feasible solutions y satisfying $A^\top y = c - s$, the corresponding objective value of y is equal to $q^\top(c - s)$: $b^\top y = q^\top A^\top y = q^\top(c - s)$, where the first equality is due to $Aq = b$.)

Let $V_p := \{x \in \mathbb{R}^n : Ax = b\} = q + \operatorname{Null}(A)$ and $V_d := c + \operatorname{Im}(A^\top)$. To summarize, we can rewrite the primal problem (1.1) and the dual problem (2.2) in the following symmetric formats:

$$\begin{array}{llll} \text{(P)} & \mathcal{X}^\star := \arg\min_{x\in\mathbb{R}^n} \ c^\top x & \text{(D)} & \mathcal{S}^\star := \arg\max_{s\in\mathbb{R}^n} \ q^\top(c - s) \\ & \quad \text{s.t. } x \in \mathcal{F}_p := V_p \cap \mathbb{R}^n_+ & & \quad \text{s.t. } s \in \mathcal{F}_d := V_d \cap \mathbb{R}^n_+ \\ & \quad\quad V_p := q + \operatorname{Null}(A) & & \quad\quad V_d := c + \operatorname{Im}(A^\top) \end{array} \tag{2.3}$$

This reformulation of the dual was, to the best of our knowledge, first proposed in [57]. Here the sets of primal and dual optima are $\mathcal{X}^\star$ and $\mathcal{S}^\star$, respectively, and we use $\mathcal{Y}^\star$ to denote the corresponding optimal solutions y associated with $\mathcal{S}^\star$.

[From a computational perspective, none of the quantities specific to the symmetric reformulation actually need to be computed, i.e., we do not need to compute q or $A^\dagger$. We present these objects as they frame our analysis and our results.]

We now briefly review optimality conditions for (2.3). Note that the duality gap in (2.3) is equal to $\text{Gap}(x,s) := c^\top x - q^\top(c-s)$, and a solution pair (x,s) is optimal for (2.3) if and only if the following conditions are met:

- Primal feasibility: $\text{Dist}(x, V_p) = 0$ and $\text{Dist}(x, \mathbb{R}^n_+) = 0$,
- Dual feasibility: $\text{Dist}(s, V_d) = 0$ and $\text{Dist}(s, \mathbb{R}^n_+) = 0$, and
- Nonpositive duality gap: $\text{Gap}(x,s) := c^\top x - q^\top(c-s) \le 0$.

The optimal primal-dual solution solution sets can be directly written as

$$\mathcal{X}^\star \times \mathcal{S}^\star := \left\{(x,s) \,\middle|\, x \in V_p,\ x \in \mathbb{R}^n_+,\ s \in V_d,\ s \in \mathbb{R}^n_+,\ \text{Gap}(x,s) \le 0\right\} .$$

In our theoretical development we will measure the error of a non-optimal pair (x,s) using the distance to optima, defined as:

$$\mathcal{E}_d(x,s) := \max\{\text{Dist}(x,\mathcal{X}^\star), \text{Dist}(s,\mathcal{S}^\star)\} . \tag{2.4}$$

(The distance to optima is not conveniently computable, and so in practice it is more typical to compute the relative error defined as $\mathcal{E}_r(x,y) := \frac{\|Ax^+ - b\|}{1+\|b\|} + \frac{\|(c-A^\top y)^-\|}{1+\|c\|} + \frac{|c^\top x^+ - b^\top y|}{1+|c^\top x^+|+|b^\top y|}$.)

It can also be observed from (2.3) that the linear subspaces associated with the primal and dual problems are orthogonal to each other. Let us denote by $\vec{V}_p$ and $\vec{V}_d$ the linear subspaces associated with the affine subspaces V_p and V_d. Then $\vec{V}_p$ and $\vec{V}_d$ are orthogonal complements. The following fact collects some other useful properties of the symmetric formulation (2.3):

Fact 2.1. *In the symmetric formulation (2.3), $\vec{V}_d$ is the orthogonal complement of $\vec{V}_p$, i.e., $\vec{V}_d = \vec{V}_p^\perp$. Furthermore, $P_{\vec{V}_p}(c) \in \vec{V}_p$ and $P_{\vec{V}_p}(c) = \arg\min_{v\in V_d} \|v\|$, and $q \in \vec{V}_d$ and $q = \arg\min_{v\in V_p} \|v\|$.*

Finally, we make the following assumption about (1.1) and its dual problem (2.1).

Assumption 1. *We assume that the LP problem (1.1) has an optimal solution, and non-optimal feasible solutions exist for the duality-paired problems (1.1) and (2.1), and equivalently for (2.3).*

2.2 Convergence properties of PDHG (without restarts) for LP

In this subsection, we review the ergodic convergence properties of PDHG (without restarts) for LP. The ergodic convergence is also known as the convergence of the average iterate. A sublinear ergodic convergence bound has been shown for PDHG in [11, 3] for general convex-concave saddlepoint problems, whose error is measured in terms of the "primal-dual gap" associated with the saddlepoint problem. We briefly review these convergence results with a focus on two issues, namely (i) the role of the primal and dual step-sizes, and (ii) the convergence (to zero) of errors for the LP problem (measured using the distances to the constraints and the duality gap) instead of gap measures of the saddlepoint problem. The material presented in this subsection will be used later in Section 3. Complete proofs of the results in this subsection are deferred to Appendix B.

One step of PDHG for LP (1.3) is defined in the function PDHGSTEP in (1.4). Note that we use $z := (x,y) \in \mathbb{R}^{m+n}$ to denote the pair of primal and dual solutions x and y, and PDHG generates iterates by $z^+ \leftarrow \text{PDHGSTEP}(z)$. We will freely use the notation z with sub/superscripts and other modifications, so that it denotes the primal and dual solutions (x,y) with the same sub/superscripts and other modifications, such as $\bar{z}^k = (\bar{x}^k, \bar{y}^k)$.

The convergence guarantees for PDHG rely on the step-sizes τ and σ in (1.4) being sufficiently small. In particular, if the following condition is satisfied:

$$M := \begin{pmatrix} \frac{1}{\tau} I_n & -A^\top \\ -A & \frac{1}{\sigma} I_m \end{pmatrix} \succeq 0 \ , \tag{2.5}$$

then PDHG's average iterates will converge to a saddlepoint of the problem (1.3) [10, 11], though the guaranteed rate of convergence is sublinear. The requirement (2.5) is equivalently written as:

$$\tau > 0,\ \sigma > 0, \ \text{ and } \tau\sigma \le \left(\frac{1}{\sigma_{\max}^{+}(A)}\right)^2 \ , \tag{2.6}$$

where $\sigma_{\max}^{+}(A)$ is the largest positive singular value of A. The matrix M defined in (2.5) turns out to be particularly useful in analyzing the convergence of PDHG through its induced inner product norm defined by $\|z\|_M := \sqrt{z^\top M z}$ (though it is only a semi-norm if M is not positive definite), which will be used extensively in the rest of this paper.

For notational convenience let us define:

$$\lambda_{\max} := \sigma_{\max}^{+}\left(A\right),\ \lambda_{\min} := \sigma_{\min}^{+}\left(A\right),\ \text{ and } \kappa := \frac{\lambda_{\max}}{\lambda_{\min}} \ . \tag{2.7}$$

Here κ is the standard condition number of the matrix A. To measure the convergence of PDHG for LP, [3] introduces the "normalized duality gap," which we have briefly reviewed in (1.6), and whose formal definition is as follows:

Definition 2.1 (Normalized duality gap, (4a) in [3])**.** *For any $z = (x, y) \in \mathbb{R}^n_+ \times \mathbb{R}^m$ and $r > 0$, define $\widetilde{B}(r; z) := \{\hat{z} := (\hat{x}, \hat{y}) : \hat{x} \ge 0$ and $\|\hat{z} - z\|_M \le r\}$. The normalized duality gap of the saddlepoint problem* (1.3) *is then defined as*

$$\rho(r; z) := \left(\frac{1}{r}\right) \max_{\hat{z} \in \widetilde{B}(r;z)} \left[L(x, \hat{y}) - L(\hat{x}, y)\right] \ . \tag{2.8}$$

Note in Definition 2.1 that $\widetilde{B}(r; z)$ is technically not a ball in the usual sense of the term, since the requirement that $\hat{x} \ge 0$ means that $\widetilde{B}(r; z)$ is not necessarily symmetric relative to its center z. The normalized duality gap serves as an upper bound for evaluating error tolerances and distances to optimality, see [3], which also presents an efficient algorithm for approximating $\rho(r; z)$.

The following lemma shows that the normalized duality gap provides an upper bound on both the distances to feasibility and the magnitude of the duality gap; this lemma is a variation of [3, Lemma 4] but measures distances instead of error tolerances.

Lemma 2.1. *For any $r > 0$, $\bar{z} := (\bar{x}, \bar{y})$ such that $\bar{x} \ge 0$, and $\bar{s} := c - A^\top \bar{y}$, the normalized duality gap $\rho(r; \bar{z})$ provides the following bounds:*

1. *Primal near-feasibiltiy:* $\operatorname{Dist}(\bar{x}, V_p) \le \frac{1}{\sqrt{\sigma}\lambda_{\min}} \cdot \rho(r; \bar{z})$ *and* $\operatorname{Dist}(\bar{x}, \mathbb{R}^n_+) = 0$,
2. *Dual near-feasibility:* $\operatorname{Dist}(\bar{s}, V_d) = 0$ *and* $\operatorname{Dist}(\bar{s}, \mathbb{R}^n_+) \le \frac{1}{\sqrt{\tau}} \cdot \rho(r; \bar{z})$*, and*
3. *Duality gap:* $\operatorname{Gap}(\bar{x}, \bar{s}) \le \max\{r, \|\bar{z}\|_M\}\rho(r; \bar{z})$.

A proof of Lemma 2.1 as well as other results in this section are given in Appendix B. Let the k-th iterate of PDHG be denoted as $z^k := (x^k, y^k)$ for $k = 0, 1, \ldots$, and let the average of the first K iterates be denoted as $\bar{z}^K = (\bar{x}^K, \bar{y}^K) := \frac{1}{K}\sum_{i=1}^K (x^i, y^i)$ for $K \ge 1$. The following lemma presents the sublinear convergence of the average iterates of PDHG for the saddlepoint LP formulation (1.3) in terms of the normalized duality gap.

Lemma 2.2. *Suppose that τ and σ satisfy* (2.6)*. Then for all $K \geq 1$ it holds that*

$$\rho(\|\bar{z}^K - z^0\|_M; \bar{z}^K) \leq \frac{4\,\mathrm{Dist}_M(z^0, \mathcal{Z}^\star)}{K} \ . \tag{2.9}$$

We remark that Lemma 2.2 can be viewed as an extension of Property 3 of [3] to the case of different primal and dual stepsizes τ and σ. By combining Lemma 2.1 and Lemma 2.2 and changing the norm, we obtain the following theorem regarding the ergodic behavior of PDHG in terms of distances to constraints and the duality gap.

Theorem 2.3. *Suppose that τ and σ satisfy* (2.6)*, and suppose that PDHG is initiated with $z^0 = (x^0, y^0) := (0, 0)$. Then for any $K \geq 1$ and $\bar{x}^K := \frac{1}{K}\sum_{i=1}^K x^i$ and $\bar{s}^K := \frac{1}{K}\sum_{i=1}^K (c - A^\top y^i)$, the following hold:*

1. *Primal near-feasibiltiy:* $\mathrm{Dist}(\bar{x}^K, V_p) \leq \frac{4\sqrt{2}}{K}\left(\frac{\mathrm{Dist}(0,\mathcal{X}^\star)}{\sqrt{\sigma\tau}\lambda_{\min}} + \frac{\mathrm{Dist}(c,\mathcal{S}^\star)}{\sigma\lambda_{\min}^2}\right)$ *and* $\mathrm{Dist}(\bar{x}^K, \mathbb{R}^n_+) = 0$,
2. *Dual near-feasibility:* $\mathrm{Dist}(\bar{s}^K, V_d) = 0$ *and* $\mathrm{Dist}(\bar{s}^K, \mathbb{R}^n_+) \leq \frac{4\sqrt{2}}{K}\left(\frac{\mathrm{Dist}(0,\mathcal{X}^\star)}{\tau} + \frac{\mathrm{Dist}(c,\mathcal{S}^\star)}{\sqrt{\sigma\tau}\lambda_{\min}}\right)$ *, and*
3. *Duality gap:* $\mathrm{Gap}(\bar{x}^K, \bar{s}^K) \leq \frac{16}{K}\left(\frac{\mathrm{Dist}(0,\mathcal{X}^\star)^2}{\tau} + \frac{\mathrm{Dist}(c,\mathcal{S}^\star)^2}{\sigma\lambda_{\min}^2} + \frac{2\,\mathrm{Dist}(0,\mathcal{X}^\star)\,\mathrm{Dist}(c,\mathcal{S}^\star)}{\sqrt{\sigma\tau}\lambda_{\min}}\right)$.

Theorem 2.3 states that the distances to the constraints and the duality gap of the average iterates of PDHG converge to zero at a sublinear rate. We note the roles of the primal and dual step-size τ and σ: generally speaking the larger τ is, the faster $\bar{s}^K$ converges to the dual feasible set, while the larger σ is, the faster $\bar{x}^K$ converges to the primal feasible set. However, a balanced choice of τ and σ is required (at least in theory) for faster convergence of the duality gap of $(\bar{x}^K, \bar{s}^K)$. Items (*1.*) and (*2.*) of Theorem 2.3 follow directly using Lemma 2.2 and Lemma 2.1 plus a norm inequality relating distances in the M-norm to Euclidean distances (Proposition 2.6).

Below we present two lemmas that are related to the performance of PDHG and are used in the proof of Theorem 2.3. The first is a well-known general nonexpansive property for PDHG (and certain other methods as well), see [53, 3, 10].

Lemma 2.4 (Nonexpansive property of PDHG, see [3])**.** *Suppose that τ, σ satisfy* (2.6)*. For any saddlepoint z^* of* (1.3)*, and for all $k \geq 0$,*

$$\|z^{k+1} - z^\star\|_M \leq \|z^k - z^\star\|_M \ . \tag{2.10}$$

Therefore under the assignment $z := z^k$ or $z := \bar{z}^k$ it holds that $\|z - z^\star\|_M \leq \left\|z^0 - z^\star\right\|_M$.

Lemma 2.5. *Suppose z^a, z^b, and z^c satisfy the nonexpansive properties: $\|z^b - z^\star\|_M \leq \|z^a - z^\star\|_M$ and $\|z^c - z^\star\|_M \leq \|z^a - z^\star\|_M$ for every $z^* \in \mathcal{Z}^*$. Then*

$$\max\{\|z^b - z^c\|_M, \|z^b\|_M\} \leq 2\,\mathrm{Dist}_M(z^a, \mathcal{Z}^\star) + \|z^a\|_M \ . \tag{2.11}$$

Lemma 2.4 is a simple extension of Proposition 2 [3] to the setting of different primal and dual step sizes. It can be proved just as in [3] by directly substituting the identical primal and dual step size η of [3] with τ and σ, respectively. Inequality (2.10) is known as the nonexpansive property and appears in many operator splitting and other related methods [31, 53, 3].

We end this section with a proposition on the relation between the M-norm used in PDHG and the Euclidean norm. While the M-norm arises naturally in the analysis of PDHG, it has the disadvantage that its quadratic form is not separable in x and y and also it implicitly includes the step-sizes τ and σ in its definition. Towards the goal of stating results in terms of the Euclidean norms on x and y, we introduce the following N-norm on $(x,y) \in \mathbb{R}^{m+n}$, whose quadratic form is separable in x and y. Define:

$$\|(x,y)\|_N := \sqrt{\frac{1}{\tau}\|x\|^2 + \frac{1}{\sigma}\|y\|^2} \;\text{ where }\; N := \begin{pmatrix} \frac{1}{\tau} I_n & \\ & \frac{1}{\sigma} I_m \end{pmatrix} \;.$$

In comparison with the M-norm, the N-norm offers advantages in both computation and analysis. Furthermore, when τ and σ are sufficiently small, the M-norm and N-norm are equivalent up to well-specified constants related to τ and σ, as follows.

Proposition 2.6. *Suppose that σ, τ satisfy (2.6). Then for any $z = (x,y) \in \mathbb{R}^{m+n}$, it holds that*

$$\sqrt{1 - \sqrt{\tau\sigma}\lambda_{\max}} \cdot \|z\|_N \leq \|z\|_M \leq \sqrt{2}\|z\|_N \;. \tag{2.12}$$

Furthermore, if $x \in \mathbb{R}^n_+$ and $s := c - A^\top y \in \mathbb{R}^n$ then

$$\mathrm{Dist}_M(z, \mathcal{Z}^\star) \leq \frac{\sqrt{2}}{\sqrt{\tau}} \mathrm{Dist}(x, \mathcal{X}^\star) + \frac{\sqrt{2}}{\sqrt{\sigma}\lambda_{\min}} \mathrm{Dist}(s, \mathcal{S}^\star) \,, \quad \textit{and} \tag{2.13}$$

$$\mathrm{Dist}_M(z, \mathcal{Z}^\star) \geq \sqrt{1 - \sqrt{\tau\sigma}\lambda_{\max}} \cdot \max\left\{ \frac{1}{\sqrt{\tau}} \mathrm{Dist}(x, \mathcal{X}^\star), \frac{1}{\sqrt{\sigma}\lambda_{\max}} \mathrm{Dist}(s, \mathcal{S}^\star) \right\} \;. \tag{2.14}$$

Appendix B presents proofs of the (uncited) results in this section.

3 Computational Guarantees for PDHG with Restarts

In this section we formally state and prove the computational guarantees that we previewed and discussed in Section 1.2. In Section 3.1 we present our new computational guarantees in two theorems: Theorem 3.1 is a more formal statement of Main Result 1 previewed in Section 1.2, and Theorem 3.2 is a formalization of the improved bound previewed in (1.17) obtained for a very special choice of stepsizes τ and σ that are primarily of theoretical interest since they are not efficiently computable. Section 3.2 is dedicated to the proofs of our key results, and Section 3.3 is comprised of statements and proofs of supporting lemmas.

Building on the basic step of PDHG described in (1.4), Algorithm 1 formally presents the framework of PDHG with restarts that we will work with. Recall that we use the notation

Algorithm 1: PDHG with general restart scheme (rPDHG)

```
Input: Initial iterate $z^{0,0} := (x^{0,0}, y^{0,0})$, $n \leftarrow 0$ ;
repeat
    initialize the inner loop: inner loop counter $k \leftarrow 0$ ;
    repeat
        conduct one step of PDHG: $z^{n,k+1} \leftarrow \textsc{PDHGstep}(z^{n,k})$ ;
        compute the average iterate in the inner loop. $\bar{z}^{n,k+1} \leftarrow \frac{1}{k+1}\sum_{i=1}^{k+1} z^{n,i}$ ;
        $k \leftarrow k+1$ ;
    until Some (verifiable) restart condition is satisfied by $\bar{z}^{n,k}$ ;
    restart the outer loop: $z^{n+1,0} \leftarrow \bar{z}^{n,k}$, $n \leftarrow n+1$ ;
until Either $z^{n,0}$ is a saddlepoint or $z^{n,0}$ satisfies some other convergence condition ;
Output: $z^{n,0}$ ( $= (x^{n,0}, y^{n,0})$)
```

$z^{k+1} \leftarrow \text{PDHGSTEP}(z^k)$ to denote an iteration of PDHG as described in (1.4). The double superscript on the variable $z^{n,k}$ indexes the outer iteration counter followed by the inner iteration counter, so that $z^{n,k}$ is the k-th inner iteration of the n-th outer loop. In order to implement Algorithm 1 it is necessary to specify a (verifiable) restart condition on the average iterate $\bar{z}^{n,k}$ in line **8** that is used to determine when to restart PDHG. We will primarily consider Algorithm 1 using the following restart condition in line **8**:

$$\rho(\|\bar{z}^{n,k} - z^{n,0}\|_M; \bar{z}^{n,k}) \le \beta \cdot \rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0}) \ , \tag{3.1}$$

for a specific value of $\beta \in (0,1)$ (in fact we will use $\beta = 1/e$ where e is the base of the natural logarithm). In this way (3.1) is nearly identical to the condition used in [3]. The condition (3.1) essentially states that the normalized duality gap shrinks by the factor β between restart values $\bar{z}^{n,k}$ and $z^{n,0}$. One of the reasons for using condition (3.1) is that the normalized duality gap can be easily approximated, see [3, Section 6].

3.1 Computational guarantees for rPDHG based on LimitingER and LP sharpness

In this section we present our main computational guarantee for rPDHG. This computational guarantee relies on the two condition measures $\theta_p^\star$ and μ_p for the primal and on their counterparts $\theta_d^\star$ and μ_d for the dual problem. Recall that we formally defined $\theta_p^\star$ and μ_p for the primal problem in Definitions 1.1 and 1.2. Now that we have established the symmetric representation of the primal and the dual problems in (2.3), we can write the formal descriptions of $\theta_d^\star$ and μ_d for the dual problem. Similar to (1.10), for $s \in V_d \setminus \mathcal{F}_d$ we define $\theta_d(s) := \frac{\text{Dist}(s,\mathcal{F}_d)}{\text{Dist}(s,\mathbb{R}^n_+)}$, and then define

$$\theta_d^\star := \lim_{\varepsilon \to 0} \left(\sup_{s \in V_d,\, \text{Dist}(s,\mathcal{S}^\star) \le \varepsilon} \theta_d(s) \right) \ , \quad \text{and} \quad \mu_d = \inf_{s \in \mathcal{F}_d \setminus \mathcal{S}^\star} \frac{\text{Dist}(s, V_d \cap H_d^\star)}{\text{Dist}(s,\ \mathcal{S}^\star)} \tag{3.2}$$

where $H_d^\star$ denotes the optimal objective hyperplane for the dual problem, namely $H_d^\star := \{\hat{s} \in \mathbb{R}^n_+ : q^\top(c-s) = q^\top(c - s^\star)\}$ for any $s^\star \in \mathcal{S}^\star$.

We consider running rPDHG (Algorithm 1) using the following simple restart condition:

Definition 3.1 (β-restart condition)**.** *For a given $\beta \in (0,1)$, the iteration (n,k) satisfies the β-restart condition if $n \ge 1$ and condition (3.1) is satisfied, or $n = 0$ and $k = 1$.*

In the following theorem we suppose for simplicity and ease of exposition that $c \in \vec{V}$. Also recall the definitions of $\lambda_{\max}$, $\lambda_{\min}$, and κ from (2.7).

Theorem 3.1. *Suppose $c \in \vec{V}$, and that Assumption 1 holds, and that Algorithm 1 (rPDHG) is run starting from $z^{0,0} = (x^{0,0}, y^{0,0}) = (0,0)$ using the β-restart condition with $\beta := 1/e$. Furthermore, let the step-sizes be chosen as follows:*

$$\tau = \frac{\|q\|}{2\kappa\|c\|} \quad \textit{and} \quad \sigma = \frac{\|c\|}{2\|q\|\lambda_{\max}\lambda_{\min}} \ . \tag{3.3}$$

Let T be the total number of PDHGSTEP *iterations that are run in order to obtain n for which $(x^{n,0}, s^{n,0})$ satisfies $\mathcal{E}_d(x^{n,0}, s^{n,0}) \le \varepsilon$. Then*

$$T \ \le \ 5e \cdot \mathcal{N} \cdot \ln\left(8e \cdot \mathcal{N} \cdot \frac{\mathcal{E}_d(x^{0,0}, s^{0,0})}{\varepsilon} \cdot \left(1 + \kappa\frac{\|c\|}{\|q\|}\right)\left(1 + \frac{\|q\|}{\|c\|}\right)\right) + 1 \ , \tag{3.4}$$

where $\mathcal{N}$ is defined as follows:

$$\mathcal{N} := 8.5\kappa \left(\frac{1}{\mu_p} + \frac{1}{\mu_d}\right)\left(\theta_p^\star + \theta_d^\star + \frac{\operatorname{Dist}(0,\mathcal{X}^\star)}{\operatorname{Dist}(0,V_p)} + \frac{\operatorname{Dist}(c,\mathcal{S}^\star)}{\operatorname{Dist}(0,V_d)}\right) . \tag{3.5}$$

In Section 1.3 we discussed key points and comments about Theorem 3.1. Here we present some detailed technical remarks. The optimality tolerance criterion used in the theorem is $\mathcal{E}_d(x^{n,0}, s^{n,0})$ (see (2.4)), which is the distance to optima of the primal and dual (slack) variable $(x^{n,0}, s^{n,0})$. Note that $\operatorname{Dist}(s^{n,0}, \mathcal{S}^\star)$ is equal to $\operatorname{Dist}_{AA^\top}(y^{n,0}, \mathcal{Y}^\star)$ – the distance of $y^{n,0}$ to $\mathcal{Y}^\star$ under the $AA^\top$-norm. The prescribed step-sizes in (3.3) are relatively easy to compute as long as estimates of the largest and smallest positive singular values of A are easy to compute. In particular neither the LP sharpness nor the LimitingER are needed in the computation of the step-sizes. See in Remark 3.5 for extensions of the theorem to other step-sizes, different values of β, and relaxations of the assumption that $c \in \vec{V}_p$. The proof of Theorem 3.1 is presented in Section 3.2, and the proofs of the supporting technical lemmas are in Section 3.3 and Appendix C.

Let us now show that the expressions (3.4) and (3.5) satisfy (1.13), (1.14), and (1.15) up to absolute constant factors, thus validating that Main Result 1 is the same as Theorem 3.1. Under the condition that $c \in \vec{V}_p$ in the theorem it follows that $\operatorname{Dist}(c, \mathcal{S}^\star) \le \|c\| + \operatorname{Dist}(0, \mathcal{S}^\star) \le 2\operatorname{Dist}(0, \mathcal{S}^\star)$, and therefore $\frac{\operatorname{Dist}(c,\mathcal{S}^\star)}{\operatorname{Dist}(0,V_d)} \le 2\frac{\operatorname{Dist}(0,\mathcal{S}^\star)}{\operatorname{Dist}(0,V_d)} = 2\frac{\operatorname{Dist}(0,\mathcal{S}^\star)}{\|c\|}$. Also it follows from Fact 2.1 that $\operatorname{Dist}(0, V_p) = \|q\| = \|b\|_Q$. Last of all we observe that the term $\ln\left(8e\left(1 + \kappa\frac{\|c\|}{\|q\|}\right)\left(1 + \frac{\|q\|}{\|c\|}\right)\right)$ in (3.4) is at most as large as $\ln(\mathcal{D})$ for the value of $\mathcal{D}$ defined in (1.15).

The step-sizes (3.3) in Theorem 3.1 are valid step-sizes that are typically relatively easy to compute. It has been observed in practice that heuristically tuning the ratio τ/σ can significantly improve the performance of PDHG, see [1, 3]. Theorem 3.2 below shows that a specially chosen step-size ratio (that is unfortunately not easy to compute in practice) results in an iteration bound that is potentially much smaller than that of Theorem 3.1.

Theorem 3.2. *Suppose $c \in \vec{V}$, and that Assumption 1 holds, and that Algorithm 1 (rPDHG) is run starting from $z^{0,0} = (x^{0,0}, y^{0,0}) = (0,0)$ using the β-restart condition with $\beta := 1/e$. Let the step-sizes be chosen as follows:*

$$\tau = \frac{\mu_d\|q\|}{2\kappa\mu_p\|c\|} \quad \text{and} \quad \sigma = \frac{\mu_p\|c\|}{2\mu_d\|q\|\lambda_{\max}\lambda_{\min}} , \tag{3.6}$$

and let T be the total number of PDHGStep *iterations that are run in order to obtain n for which $(x^{n,0}, s^{n,0})$ satisfies $\mathcal{E}_d(x^{n,0}, s^{n,0}) \le \varepsilon$. Then*

$$T \ \le\ 5e \cdot \widehat{\mathcal{N}} \cdot \ln\left(8e \cdot \widehat{\mathcal{N}} \cdot \frac{\mathcal{E}_d(x^{0,0}, s^{0,0})}{\varepsilon} \cdot \left(1 + \kappa\frac{\mu_p\|c\|}{\mu_d\|q\|}\right)\left(1 + \frac{\mu_d\|q\|}{\mu_p\|c\|}\right)\right) + 1 , \tag{3.7}$$

where $\widehat{\mathcal{N}}$ is defined as follows:

$$\widehat{\mathcal{N}} := 16\kappa\left(\frac{\theta_p^\star}{\mu_p} + \frac{\theta_d^\star}{\mu_d} + \frac{\operatorname{Dist}(0,\mathcal{X}^\star)}{\mu_d \cdot \operatorname{Dist}(0,V_p)} + \frac{\operatorname{Dist}(c,\mathcal{S}^\star)}{\mu_p \cdot \operatorname{Dist}(0,V_d)}\right) . \tag{3.8}$$

Note that $\widehat{\mathcal{N}}$ in (3.8) is potentially much smaller than (3.5) because there are fewer cross-terms involving μ_p, μ_d, $\theta_p^\star$, and $\theta_d^\star$. However, the assignment of the step-sizes (3.6) requires knowledge of the LP sharpness constants μ_p and μ_d (or just their ratio), which are likely to be neither known nor easily computable. Although the step-sizes in (3.6) are not implementable in practice, Theorem 3.2

lends some theoretical justification to the observed practical value of adaptively tuning the step-sizes in practical solvers [3, 1]. Using identical logic as presented earlier, it is straightforward to show that the value of $\widehat{\mathcal{N}}$ in (3.8) satisfies (1.17) up to an absolute constant factor. See the last paragraph of Section 1.3 for other remarks and points about Theorem 3.2. The proof of Theorem 3.2 is similar to that of Theorem 3.1, and is presented in Appendix C.

3.2 Two lemmas, the proof of Theorem 3.1, and extensions

Similar to the main complexity proofs in [3], the proof of Theorem 3.1 involves two steps. The first step is to demonstrate that if the normalized duality gap of the iterates satisfies a certain sharpness condition, then rPDHG achieves linear convergence, with the total number of PDHGStep iterations being largely determined by this sharpness condition. The second step involves analyzing and bounding the sharpness condition using the condition measures μ_p, μ_d, $\theta_p^\star$, and $\theta_d^\star$ (plus other data-related quantities). Lemma 3.3 below embodies the first step.

Lemma 3.3. *Suppose Algorithm 1 is run starting from $z^{0,0} = (x^{0,0}, y^{0,0}) = (0,0)$ using the β-restart condition, and let the step-sizes τ and σ satisfy (2.6). Suppose also that there exists a scalar $\mathcal{N}$ for which it holds that*

$$\mathrm{Dist}_M(z^{n,0}, \mathcal{Z}^\star) \ \le \ \mathcal{N} \cdot \rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0}) \tag{3.9}$$

for all $n \ge 1$. Let T be the total number of PDHGStep *iterations that are run in order to obtain n for which $z^{n,0} = (x^{n,0}, s^{n,0})$ satisfies $\mathcal{E}_d(x^{n,0}, s^{n,0}) \le \varepsilon$. Then*

$$T \ \le \ \frac{5}{\beta \ln(1/\beta)} \cdot \mathcal{N} \cdot \ln\left(\tilde{c} \cdot \mathcal{N} \cdot \left(\frac{\mathcal{E}_d(x^{0,0}, s^{0,0})}{\varepsilon}\right)\right) + 1 \ , \tag{3.10}$$

where $\tilde{c} := \frac{4\sqrt{2}}{\beta\sqrt{1-\sqrt{\tau\sigma}\lambda_{\max}}}\left(\sqrt{\tau} + \sqrt{\sigma}\lambda_{\max}\right) \cdot \left(\frac{1}{\sqrt{\tau}} + \frac{1}{\sqrt{\sigma}\lambda_{\min}}\right)$.

The proof of Lemma 3.3 is similar to the proof of [3, Theorem 2], and proceeds by establishing an upper bound on the PDHGStep iteration count between restarts. The primary difference is that we are particularly interested in the role of the primal and dual step-sizes. The complete proof of Lemma 3.3 is presented in Appendix C.

Lemma 3.4 embodies the second step described above.

Lemma 3.4. *Suppose that the initial iterate of rPDHG is $z^0 := (0,0)$, and let z^b, $z^c \in \mathbb{R}^n_+ \times \mathbb{R}^m$ satisfy $z^b \neq z^c$ and the nonexpansive inequalities $\|z^b - z^\star\|_M \le \|z^0 - z^\star\|_M$ and $\|z^c - z^\star\|_M \le \|z^0 - z^\star\|_M$ for all $z^\star \in \mathcal{Z}^\star$. Then it holds that:*

$$\mathrm{Dist}_M(z^b, \mathcal{Z}^\star) \le \widetilde{\mathcal{N}} \cdot \rho(\|z^b - z^c\|_M; z^b) \ , \tag{3.11}$$

in which

$$\widetilde{\mathcal{N}} := \left(\begin{array}{l} \quad \left(\frac{3\sqrt{2}}{\sqrt{\sigma\tau}\lambda_{\min}\mu_p} + \frac{\sqrt{2}}{\sigma\lambda_{\min}^2\mu_d} \cdot \frac{\|P_{\vec{V}_p}(c)\|}{\|q\|}\right) \cdot \left(\theta_p^\star + \frac{\|P_{\vec{V}_p^\perp}(c)\|}{\|P_{\vec{V}_p}(c)\|}\right) \\ + \left(\frac{2\sqrt{2}}{\sqrt{\sigma\tau}\lambda_{\min}\mu_d} + \frac{\sqrt{2}}{\tau\mu_p} \cdot \frac{\|q\|}{\|P_{\vec{V}_p}(c)\|}\right) \cdot \theta_d^\star \\ + \left(\frac{4}{\sqrt{\tau}\mu_p\|P_{\vec{V}_p}(c)\|} + \frac{4}{\sqrt{\sigma}\mu_d\|q\|\lambda_{\min}}\right) \cdot \left(\frac{1}{\sqrt{\tau}}\,\mathrm{Dist}(0, \mathcal{X}^\star) + \frac{1}{\lambda_{\min}\sqrt{\sigma}}\,\mathrm{Dist}(c, \mathcal{S}^\star)\right) \end{array} \right) . \tag{3.12}$$

The proof of Lemma 3.4 is presented in Section 3.3. Armed with Lemma 3.3 and Lemma 3.4, we now prove Theorem 3.1.

Proof of Theorem 3.1. For any $n \geq 1$ it holds that $z^{n,0} = \bar{z}^{n-1,K} = \frac{1}{K}\sum_{i=1}^{K} z^{n-1,i}$ where K is the total number of inner iterations of rPDHG run in the outer loop iteration $n-1$, whose initial iterate value is $z^{n-1,0}$. It then follows from the nonexpansive properties of PDHG from Lemma 2.4 that $\|z^{n,0} - z^\star\|_M \leq \|z^{n-1,0} - z^\star\|_M$ for each $z^\star \in \mathcal{Z}^\star$. By telescoping inequalities we then have

$$\|z^{n,0} - z^\star\|_M \leq \|z^{n-1,0} - z^\star\|_M \leq \cdots \leq \|z^{0,0} - z^\star\|_M$$

for each $z^\star \in \mathcal{Z}^\star$. Now notice that $z^0 := z^{0,0}$, $z^b := z^{n,0}$, and $z^c := z^{n-1,0}$ satisfy the hypotheses of Lemma 3.4, whereby it holds that

$$\mathrm{Dist}_M(z^{n,0}, \mathcal{Z}^\star) \leq \widetilde{\mathcal{N}} \cdot \rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0}) \ . \tag{3.13}$$

From the supposition of Theorem 3.1 we have $c \in \vec{V}_p$ and therefore $\|P_{\vec{V}_p}(c)\| = \|c\|$, $\|P_{\vec{V}_p^\perp}(c)\| = 0$, and $\|c\| = \mathrm{Dist}(0, V_d)$. Also by construction we have $\|q\| = \mathrm{Dist}(0, V_p)$. Substituting the step-size values (3.3) into (3.12) and using the above norm equalities in (3.12) yields:

$$\widetilde{\mathcal{N}} = 6\sqrt{2}\kappa\left(\frac{1}{\mu_p} + \frac{\frac{1}{3}}{\mu_d}\right)\theta_p^\star + 4\sqrt{2}\kappa\left(\frac{\frac{1}{2}}{\mu_p} + \frac{1}{\mu_d}\right)\theta_d^\star + 8\kappa\left(\frac{1}{\mu_p} + \frac{1}{\mu_d}\right)\left[\frac{\mathrm{Dist}(0, \mathcal{X}^\star)}{\mathrm{Dist}(0, V_p)} + \frac{\mathrm{Dist}(c, \mathcal{S}^\star)}{\mathrm{Dist}(0, V_d)}\right] \ . \tag{3.14}$$

Now notice that the value of $\mathcal{N}$ specified in (3.5) is at least as large as the value of $\widetilde{\mathcal{N}}$ above, whereby it holds that $\mathrm{Dist}_M(z^{n,0}, \mathcal{Z}^\star) \leq \mathcal{N} \cdot \rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0})$ for the value of $\mathcal{N}$ specified in (3.5). Therefore condition (3.9) of Lemma 3.3 is satisfied, and it follows from Lemma 3.3 that T satisfies (3.10) with the value of $\tilde{c}$ specified in the statement of the lemma, namely:

$$T \ \leq \ \frac{5}{\beta \ln(1/\beta)} \cdot \mathcal{N} \cdot \ln\left(\tilde{c} \cdot \mathcal{N} \cdot \left(\frac{\mathcal{E}_d(x^{0,0}, s^{0,0})}{\varepsilon}\right)\right) + 1 \ . \tag{3.15}$$

Substituting in the step-sizes (3.3) and $\beta = 1/e$ into the value of $\tilde{c}$ in Lemma 3.3 we find that $\tilde{c} = 8e \cdot \left(1 + \kappa\frac{\|c\|}{\|q\|}\right)\left(1 + \frac{\|q\|}{\|c\|}\right)$, and using $\beta = 1/e$ and the value of $\tilde{c}$ above in (3.15) we finally arrive at:

$$T \ \leq \ 5e \cdot \mathcal{N} \cdot \ln\left(8e \cdot \mathcal{N} \cdot \left(\frac{\mathcal{E}_d(x^{0,0}, s^{0,0})}{\varepsilon}\right)\left(1 + \kappa\frac{\|c\|}{\|q\|}\right)\left(1 + \frac{\|q\|}{\|c\|}\right)\right) + 1 \ , \tag{3.16}$$

which completes the proof of the theorem. □

We now present several remarks.

Remark 3.5 (Extensions: other parameter settings, and allowing $c \notin \vec{V}_p$.)**.** *Theorem 3.1 uses the specific values of the primal and dual step-sizes in (3.3), uses $\beta = 1/e$, and assumes that $c \in \vec{V}_p$. The results could have been presented without these specific conditions, albeit at the expense of more complicated expressions and more challenging exposition. Indeed, the step-sizes τ and σ could have been left unassigned – just as they are in Lemma 3.3 and Lemma 3.4. Using (3.13) directly with Lemma 3.3 and some algebraic calculations yields the proof for other step-size assignments.*

If $c \notin \vec{V}_p$, then the value of $\mathcal{N}$ in (3.5) changes based on the changed value of $\widetilde{\mathcal{N}}$ in (3.12). Observing (3.12), we see that $\|P_{\vec{V}_p^\perp}(c)\|$ appears only in the last term of the first of the three lines; if $c \notin \vec{V}_p$ then $\|P_{\vec{V}_p^\perp}(c)\| \neq 0$, and then the iteration bound in Theorem 3.1 will change by replacing $\theta_p^\star$ with $\theta_p^\star + \frac{\|P_{\vec{V}_p^\perp}(c)\|}{\|P_{\vec{V}_p}(c)\|}$ throughout, and the step-sizes in (3.3) will also change. If the step-sizes follow *some other rules, then in addition to the change of $\mathcal{N}$ based on $\widetilde{\mathcal{N}}$ in* (3.12)*, the value of $\tilde{c}$ in Lemma 3.3 will also change.*

The choice of the value of the scalar β affects only the expression $\frac{5}{\beta \ln(1/\beta)}$ outside the logarithmic term in (3.10) *of Lemma 3.3. The specified value $\beta = 1/e$ was chosen to minimize the scalar $\frac{5}{\beta \ln(1/\beta)}$, but the value of β could have been left unassigned, again at the expense of expositional simplicity.*

Remark 3.6 (Relation to the bound (1.5) from [3])**.** *The proof of Theorem 3.1 is based in part on Lemma 3.3. We note that the idea of Lemma 3.3 stems from the iteration bound* (1.5) *from [3]. Indeed the condition* (3.9) *in Lemma 3.3 is quite similar to the sharpness condition* (1.7) *from [3]; the difference is only in using $\mathcal{N}$ instead of $\frac{1}{\alpha}$ and using the M-norm distance $\operatorname{Dist}_M(z^{n,0}, \mathcal{Z}^\star)$ instead of the Euclidean distance $\operatorname{Dist}(z^{n,0}, \mathcal{Z}^\star)$.*

Remark 3.7 (Relating our results to sharpness of the normalized duality gap functional in [3])**.** *It follows from Lamma 3.4 and the proof of Theorem 3.1 that the normalized duality gap functional $\rho(r; z)$ obeys an "α-sharpness" condition $\rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0}) \geq \alpha \cdot \operatorname{Dist}_M(z^{n,0}, \mathcal{Z}^\star)$ along the algorithm iterates for*

$$\alpha := \frac{1}{\mathcal{N}} = \frac{1}{8.5\kappa \left(\frac{1}{\mu_p} + \frac{1}{\mu_d}\right) \left(\theta_p^\star + \theta_d^\star + \frac{\operatorname{Dist}(0,\mathcal{X}^\star)}{\operatorname{Dist}(0,V_p)} + \frac{\operatorname{Dist}(c,\mathcal{S}^\star)}{\operatorname{Dist}(0,V_d)}\right)} ,$$

where the expression for $\mathcal{N}$ above is from (3.5)*. Note that this is just like the α-sharpness condition* (1.7) *except it is expressed with the M-norm distance $\operatorname{Dist}_M(z^{n,0}, \mathcal{Z}^\star)$ instead of the Euclidean distance. Furthermore, it follows from Proposition 2.6 that $\operatorname{Dist}_M(z^{n,0}, \mathcal{Z}^\star) \geq \gamma \cdot \operatorname{Dist}(z^{n,0}, \mathcal{Z}^\star)$ for a certain constant γ whose expression is defined using $\tau, \sigma, \lambda_{\max}$, see the formulas in Proposition 2.6 for details.*

Remark 3.8 (Extensions to other first-order methods)**.** *The proof of Theorem 3.1 is based on the two "steps" of Lemma 3.3 and Lemma 3.4, and it is quite possible that it can be extended to certain other first-order methods. Indeed, [3] proves that some iteration bounds analogous to* (1.5) *also apply to other first-order methods, such as the alternating direction method of multipliers (ADMM) and extragradient method (EGM). For these methods, if similar results to Lemma 3.4 can be extended (which we believe likely is the case), then computational guarantees similar to Theorem 3.1 may be obtained for these other methods.*

Section 3.4 below is focused on the proof of Lemma 3.4.

3.3 Proof of Lemma 3.4

This subsection is devoted to proving Lemma 3.4. As part of this task, we will need the following two very specific lemmas.

Lemma 3.9. *For any $\tilde{z} = (\tilde{x}, \tilde{y})$, let $z^\star \in \arg\min_{z \in \mathcal{Z}^\star} \|z - \tilde{z}\|_M$, and define for all $t \geq 0$*

$$\tilde{z}_t := \tilde{z} + t \cdot (\tilde{z} - z^\star) \tag{3.17}$$

(whereby $\tilde{z}_0 = \tilde{z}$). Suppose that there exist nonnegative scalars C_1, C_2, C_3 such that the following inequality ($\mathcal{I}_t$) *holds for $t = 0$:*

$$\operatorname{Dist}_M(\tilde{z}_t, \mathcal{Z}^\star) \leq C_1 \cdot \operatorname{Dist}(\tilde{x}_t, V_p) + C_2 \cdot \operatorname{Dist}(\tilde{s}_t, \mathbb{R}^n_+) + C_3 \cdot \max\{0, \operatorname{Gap}(\tilde{x}_t, \tilde{s}_t)\} \ . \tag{$\mathcal{I}_t$}$$

Then inequality ($\mathcal{I}_t$) *holds for all $t \geq 0$.*

Proof. Let $z^\star \in \arg\min_{z\in\mathcal{Z}^\star}\|z-\tilde z\|_M$ be given. Then because $\mathcal{Z}^\star$ is a convex set, it follows that $z^\star \in \arg\min_{z\in\mathcal{Z}^\star}\|z-\tilde z_t\|_M$ for all $t\geq 0$. Therefore $\mathrm{Dist}_M(\tilde z_t,\mathcal{Z}^\star)=\|z^\star-\tilde z_t\|_M=(1+t)\cdot\|\tilde z-z^\star\|_M=(1+t)\cdot\mathrm{Dist}_M(\tilde z,\mathcal{Z}^\star)$.

Regarding the terms on the right-hand side of ($\mathcal{I}_t$), for $t\geq 0$ we have $\mathrm{Dist}(\tilde x_t,V_p)=(1+t)\cdot\mathrm{Dist}(\tilde x,V_p)$ and $\max\{0,\mathrm{Gap}(\tilde x_t,\tilde s_t)\}=(1+t)\cdot\max\{0,\mathrm{Gap}(\tilde x,\tilde s)\}$. It also holds that:

$$\mathrm{Dist}(\tilde s_t,\mathbb{R}^n_+)=\|(\tilde s_t)^-\|=\|(\tilde s+t\cdot(\tilde s-s^\star))^-\|\geq\|(1+t)(\tilde s)^-\|=(1+t)\cdot\mathrm{Dist}(\tilde s,\mathbb{R}^n_+)\ ,$$

where the inequality above follows since $s^\star\geq 0$ and hence $(\tilde s+t\cdot(\tilde s-s^\star))^-\geq(1+t)(\tilde s)^-$. Combining the above equalities and inequalities proves ($\mathcal{I}_t$) for all $t\geq 0$. □

Lemma 3.10. *Suppose that τ and σ satisfy (2.6). For any $x^0\in\mathbb{R}^n_+$, $y^0\in\mathbb{R}^m$ and $s^0:=c-A^\top y^0\in\mathbb{R}^n$, then $z^0:=(x^0,y^0)$ satisfies:*

$$\mathrm{Dist}_M(z^0,\mathcal{Z}^\star)\leq C_1\cdot\mathrm{Dist}(x^0,V_p)+C_2\cdot\mathrm{Dist}(s^0,\mathbb{R}^n_+)+C_3\cdot\max\{0,\mathrm{Gap}(x^0,s^0)\}\ , \tag{3.18}$$

where

$$\begin{aligned}
C_1 &:= \left(\theta_p^\star\|P_{\vec V_p}(c)\|+\|P_{\vec V_p^\perp}(c)\|\right)\cdot\left(\tfrac{3\sqrt2}{\sqrt\tau\mu_p\|P_{\vec V_p}(c)\|}+\tfrac{\sqrt2}{\sqrt\sigma\mu_d\|q\|\lambda_{\min}}\right)\\
C_2 &:= \theta_d^\star\|q\|\cdot\left(\tfrac{2\sqrt2}{\sqrt\sigma\mu_d\|q\|\lambda_{\min}}+\tfrac{\sqrt2}{\sqrt\tau\mu_p\|P_{\vec V_p}(c)\|}\right)\\
C_3 &:= \left(\tfrac{\sqrt2}{\sqrt\tau\mu_p\|P_{\vec V_p}(c)\|}+\tfrac{\sqrt2}{\sqrt\sigma\mu_d\|q\|\lambda_{\min}}\right)\ .
\end{aligned} \tag{3.19}$$

Proof. We presume that $z^0\notin\mathcal{Z}^\star$, for otherwise (3.18) follows trivially. Let $\tilde x^0$ be the projection of x^0 onto V_p, namely $\tilde x^0=P_{V_p}(x^0)$. Then $\tilde x^0-x^0$ is orthogonal to $\vec V_p$. Also let $\hat x$ and $\hat s$ be the projections of $\tilde x^0$ and s^0 onto $\mathcal{F}_p$ and $\mathcal{F}_d$, respectively, namely $\hat x:=\arg\min_{x\in\mathcal{F}_p}\|x-\tilde x^0\|$ and $\hat s:=\arg\min_{s\in\mathcal{F}_d}\|s-s^0\|$. Since $\hat x$ and $\hat s$ are feasible for $\mathcal{F}_p$ and $\mathcal{F}_d$, respectively, the duality gap $\mathrm{Gap}(\hat x,\hat s)$ is nonnegative. Furthermore, we have

$$\begin{aligned}
\mathrm{Gap}(\hat x,\hat s)\ &=c^\top\hat x-q^\top(c-\hat s)=\mathrm{Gap}(x^0,s^0)+c^\top(\hat x-x^0)+q^\top(\hat s-s^0)\\
&\leq\mathrm{Gap}(x^0,s^0)+c^\top(\hat x-x^0)+\|q\|\cdot\mathrm{Dist}(s^0,\mathcal{F}_d)\ ,
\end{aligned} \tag{3.20}$$

and also

$$\begin{aligned}
&c^\top(\hat x-x^0)=\left(P_{\vec V_p}(c)+P_{\vec V_p^\perp}(c)\right)^\top\left((\hat x-\tilde x^0)+(\tilde x^0-x^0)\right)=P_{\vec V_p}(c)^\top\left(\hat x-\tilde x^0\right)+P_{\vec V_p^\perp}(c)^\top\left(\tilde x^0-x^0\right)\\
&\leq\|P_{\vec V_p}(c)\|\cdot\left\|\hat x-\tilde x^0\right\|+\|P_{\vec V_p^\perp}(c)\|\cdot\left\|\tilde x^0-x^0\right\|=\|P_{\vec V_p}(c)\|\cdot\mathrm{Dist}\left(\tilde x^0,\mathcal{F}_p\right)+\|P_{\vec V_p^\perp}(c)\|\cdot\mathrm{Dist}(x^0,V_p)\ ,
\end{aligned} \tag{3.21}$$

where the second equality above is due to $\hat x-\tilde x^0\in\vec V_p$ (because both $\hat x$, $\tilde x^0\in V_p$) and $\tilde x^0-x^0\in\vec V_p^\perp$. Substituting (3.21) into (3.20) then yields:

$$\mathrm{Gap}(\hat x,\hat s)\leq\mathrm{Gap}(x^0,s^0)+\|P_{\vec V_p}(c)\|\cdot\mathrm{Dist}\left(\tilde x^0,\mathcal{F}_p\right)+\|P_{\vec V_p^\perp}(c)\|\cdot\mathrm{Dist}(x^0,V_p)+\|q\|\cdot\mathrm{Dist}(s^0,\mathcal{F}_d)\ . \tag{3.22}$$

Now we aim to replace the distance term involving $\tilde x^0$ in the right-hand side of (3.22) with a term involving x^0. From the definition of the ER $\theta_p(\cdot)$ we have:

$$\mathrm{Dist}(\tilde x^0,\mathcal{F}_p)\ =\theta_p(\tilde x^0)\cdot\mathrm{Dist}(\tilde x^0,\mathbb{R}^n_+)\ \leq\theta_p(\tilde x^0)\cdot\|\tilde x^0-x^0\|=\theta_p(\tilde x^0)\cdot\mathrm{Dist}(x^0,V_p)\ , \tag{3.23}$$

where the inequality uses $x^0\in\mathbb{R}^n_+$. Note that $\mathrm{Dist}(x^0,\mathcal{F}_p)\leq\mathrm{Dist}(\tilde x^0,\mathcal{F}_p)+\|x^0-\tilde x^0\|=\mathrm{Dist}(\tilde x^0,\mathcal{F}_p)+\mathrm{Dist}(x^0,V_p)$, so using (3.23) we obtain

$$\mathrm{Dist}(x^0,\mathcal{F}_p)\leq(\theta_p(\tilde x^0)+1)\cdot\mathrm{Dist}(x^0,V_p)\ . \tag{3.24}$$

Similarly, since $s^0 \in V_d$, using the ER $\theta_d(\cdot)$ we have:

$$\operatorname{Dist}(s^0, \mathcal{F}_d) \le \theta_d(s^0) \cdot \operatorname{Dist}(s^0, \mathbb{R}^n_+) \ . \tag{3.25}$$

Substituting (3.23) and (3.25) into (3.22) yields:

$$\operatorname{Gap}(\hat{x}, \hat{s}) \le \operatorname{Gap}(x^0, s^0) + \Big(\|P_{\vec{V}_p}(c)\| \theta_p(\tilde{x}^0) + \|P_{\vec{V}_p^\perp}(c)\| \Big) \cdot \operatorname{Dist}(x^0, V_p) + \|q\| \cdot \theta_d(s^0) \cdot \operatorname{Dist}(s^0, \mathbb{R}^n_+) \ . \tag{3.26}$$

Let us now use (3.26) to bound the distances to optima. Note that the duality gap $\operatorname{Gap}(\hat{x}, \hat{s})$ is an upper bound for both $c^\top \hat{x} - f^\star$ and $f^\star - q^\top (c - \hat{s})$, where $f^\star$ denotes the optimal objective value. Then because $\hat{x} \in V_p$ and $\hat{s} \in V_d$ we have:

$$\operatorname{Dist}(\hat{x}, V_p \cap \{x : c^\top x = f^\star\}) \le \frac{\operatorname{Gap}(\hat{x}, \hat{s})}{\|P_{\vec{V}_p}(c)\|} \ \text{and} \ \operatorname{Dist}(\hat{s}, V_d \cap \{s : q^\top (c - s) = f^\star\}) \le \frac{\operatorname{Gap}(\hat{x}, \hat{s})}{\|P_{\vec{V}_d}(q)\|} \ ,$$

and note that $\|P_{\vec{V}_d}(q)\| = \|q\|$ because $q \in \vec{V}_d$. Because $\operatorname{Gap}(\hat{x}, \hat{s}) \ge |c^\top \hat{x} - f^\star|$ and $\operatorname{Gap}(\hat{x}, \hat{s}) \ge |q^\top (c - \hat{s}) - f^\star|$, we have:

$$\begin{aligned} \operatorname{Dist}(\hat{x}, \mathcal{X}^\star) &\le \tfrac{\operatorname{Dist}(\hat{x}, V_p \cap \{x : c^\top x = f^\star\})}{\mu_p} \le \tfrac{1}{\mu_p} \cdot \tfrac{\operatorname{Gap}(\hat{x}, \hat{s})}{\|P_{\vec{V}_p}(c)\|} \ , \\ \operatorname{Dist}(\hat{s}, \mathcal{S}^\star) &\le \tfrac{\operatorname{Dist}(\hat{s}, V_d \cap \{s : q^\top (c - s) = f^\star\})}{\mu_d} \le \tfrac{1}{\mu_d} \cdot \tfrac{\operatorname{Gap}(\hat{x}, \hat{s})}{\|q\|} \ . \end{aligned} \tag{3.27}$$

Now since $\operatorname{Dist}(x^0, \mathcal{X}^\star) \le \|x^0 - \hat{x}\| + \operatorname{Dist}(\hat{x}, \mathcal{X}^\star) = \operatorname{Dist}(x^0, \mathcal{F}_p) + \operatorname{Dist}(\hat{x}, \mathcal{X}^\star)$, using (3.24) and (3.27) implies that:

$$\operatorname{Dist}(x^0, \mathcal{X}^\star) \le (\theta_p(\tilde{x}^0) + 1) \cdot \operatorname{Dist}(x^0, V_p) + \frac{1}{\mu_p} \cdot \frac{\operatorname{Gap}(\hat{x}, \hat{s})}{\|P_{\vec{V}_p}(c)\|} \ . \tag{3.28}$$

Combining (3.26) and (3.28) we obtain:

$$\begin{aligned} \operatorname{Dist}(x^0, \mathcal{X}^\star) \ &\le (\theta_p(\tilde{x}^0) + 1) \cdot \operatorname{Dist}(x^0, V_p) \\ &\quad + \frac{\operatorname{Gap}(x^0, s^0) + \Big(\|P_{\vec{V}_p}(c)\| \theta_p(\tilde{x}^0) + \|P_{\vec{V}_p^\perp}(c)\| \Big) \cdot \operatorname{Dist}(x^0, V_p) + \|q\| \cdot \theta_d(s^0) \cdot \operatorname{Dist}(s^0, \mathbb{R}^n_+)}{\mu_p \|P_{\vec{V}_p}(c)\|} \\ &= \frac{\operatorname{Gap}(x^0, s^0)}{\mu_p \|P_{\vec{V}_p}(c)\|} + \left(\frac{\theta_p(\tilde{x}^0)}{\mu_p} + \frac{\|P_{\vec{V}_p^\perp}(c)\|}{\mu_p \|P_{\vec{V}_p}(c)\|} + \theta_p(\tilde{x}^0) + 1 \right) \cdot \operatorname{Dist}(x^0, V_p) + \frac{\|q\| \theta_d(s^0)}{\mu_p \|P_{\vec{V}_p}(c)\|} \cdot \operatorname{Dist}(s^0, \mathbb{R}^n_+) \ . \end{aligned} \tag{3.29}$$

Note that because $\mu_p \le 1$ and $\theta_p(\tilde{x}^0) \ge 1$, it follows that (3.29) can be relaxed to:

$$\operatorname{Dist}(x^0, \mathcal{X}^\star) \le \frac{\operatorname{Gap}(x^0, s^0)}{\mu_p \|P_{\vec{V}_p}(c)\|} + \left(\frac{3\theta_p(\tilde{x}^0)}{\mu_p} + \frac{\|P_{\vec{V}_p^\perp}(c)\|}{\mu_p \|P_{\vec{V}_p}(c)\|} \right) \cdot \operatorname{Dist}(x^0, V_p) + \frac{\|q\| \theta_d(s^0)}{\mu_p \|P_{\vec{V}_p}(c)\|} \cdot \operatorname{Dist}(s^0, \mathbb{R}^n_+) \ . \tag{3.30}$$

Using almost identical logic applied to s^0 instead of x^0, we obtain:

$$\operatorname{Dist}(s^0, \mathcal{S}^\star) \le \frac{\operatorname{Gap}(x^0, s^0)}{\mu_d \|q\|} + \frac{2\theta_d(s^0)}{\mu_d} \cdot \operatorname{Dist}(s^0, \mathbb{R}^n_+) + \frac{\|P_{\vec{V}_p}(c)\| \theta_p(\tilde{x}^0) + \|P_{\vec{V}_p^\perp}(c)\|}{\mu_d \|q\|} \cdot \operatorname{Dist}(x^0, V_p) \ . \tag{3.31}$$

Combining (3.30) with (3.31) and using the right-most inequality of (2.13), it follows that

$$\mathrm{Dist}_M(z^0,\mathcal{Z}^\star) \le \bar{C}_1(z^0)\cdot\mathrm{Dist}(x^0,V_p)+\bar{C}_2(z^0)\cdot\mathrm{Dist}(s^0,\mathbb{R}^n_+)+\bar{C}_3(z^0)\cdot\max\{0,\mathrm{Gap}(x^0,s^0)\}\ , \tag{3.32}$$

where

$$\begin{aligned}
\bar{C}_1(z^0) &:= \left(\theta_p(P_{V_p}(x^0))\|P_{\vec{V}_p}(c)\|+\|P_{\vec{V}_p^\perp}(c)\|\right)\cdot\left(\tfrac{3\sqrt{2}}{\sqrt{\tau}\mu_p\|P_{\vec{V}_p}(c)\|}+\tfrac{\sqrt{2}}{\sqrt{\sigma}\mu_d\|q\|\lambda_{\min}}\right)\\
\bar{C}_2(z^0) &:= \theta_d(s^0)\|q\|\cdot\left(\tfrac{2\sqrt{2}}{\sqrt{\sigma}\mu_d\|q\|\lambda_{\min}}+\tfrac{\sqrt{2}}{\sqrt{\tau}\mu_p\|P_{\vec{V}_p}(c)\|}\right)\\
\bar{C}_3(z^0) &:= C_3\ ,
\end{aligned} \tag{3.33}$$

and notice in the definition $\bar{C}_1$ we have written $\theta_p(P_{V_p}(x^0))$ since in fact $\tilde{x}^0 := P_{V_p}(x^0)$. Now notice that (3.32) is nearly identical to (3.18), except that the constants $\bar{C}_1(z^0)$ and $\bar{C}_2(z^0)$ use $\theta_p(P_{V_p}(x^0))$ instead of $\theta_p^\star$, and use $\theta_d(s^0)$ instead of $\theta_d^\star$.

To finish the proof, let $z^\star\in\arg\min_{z\in\mathcal{Z}^\star}\|z-z^0\|_M$ be fixed, and define $z^\lambda := (1-\lambda)z^0+\lambda z^\star$ for all $\lambda\in[0,1)$. Then (3.32) holds for z^λ, namely:

$$\mathrm{Dist}_M(z^\lambda,\mathcal{Z}^\star) \le \bar{C}_1(z^\lambda)\cdot\mathrm{Dist}(x^\lambda,V_p)+\bar{C}_2(z^\lambda)\cdot\mathrm{Dist}(s^\lambda,\mathbb{R}^n_+)+\bar{C}_3(z^\lambda)\cdot\max\{0,\mathrm{Gap}(x^\lambda,s^\lambda)\}\ , \tag{3.34}$$

since z^λ satisfies the same hypotheses as z^0. And since $z^\star\in\arg\min_{z\in\mathcal{Z}^\star}\|z-z^\lambda\|_M$ we can invoke Lemma 3.9. It follows from Lemma 3.9 that for all $t\ge 0$ with $z_t^\lambda := z^\lambda+t(z^\lambda-z^\star)$ that

$$\mathrm{Dist}_M(z_t^\lambda,\mathcal{Z}^\star) \le \bar{C}_1(z^\lambda)\cdot\mathrm{Dist}(x_t^\lambda,V_p)+\bar{C}_2(z^\lambda)\cdot\mathrm{Dist}(s_t^\lambda,\mathbb{R}^n_+)+\bar{C}_3(z^\lambda)\cdot\max\{0,\mathrm{Gap}(x_t^\lambda,s_t^\lambda)\}\ . \tag{3.35}$$

Setting $t=\lambda/(1-\lambda)$ yields $z_t^\lambda = z^0$, whereby:

$$\mathrm{Dist}_M(z^0,\mathcal{Z}^\star) \le \bar{C}_1(z^\lambda)\cdot\mathrm{Dist}(x^0,V_p)+\bar{C}_2(z^\lambda)\cdot\mathrm{Dist}(s^0,\mathbb{R}^n_+)+\bar{C}_3(z^\lambda)\cdot\max\{0,\mathrm{Gap}(x^0,s^0)\}\ . \tag{3.36}$$

Now let $\lambda\to 1$, whereby $z^\lambda\to z^\star$, and so $\limsup_{\lambda\to 1}\theta_p(P_{V_p}(x^\lambda))\le\theta_p^\star$ and therefore $\limsup_{\lambda\to1}\bar{C}_1(z^\lambda)\le C_1$. Similarly $\limsup_{\lambda\to1}\theta_d(s^\lambda))\le\theta_d^\star$ and therefore $\limsup_{\lambda\to1}\bar{C}_2(z^\lambda)\le C_2$. Thus, we can conclude (3.18) from (3.36). □

Finally, we prove Lemma 3.4.

Proof of Lemma 3.4. The proof is a combination of Lemmas 2.1, 2.5, and 3.10. Setting $s^b = c-A^\top y^b$ it follows from Lemma 2.1 that

$$\begin{aligned}
&\mathrm{Dist}(x^b,V_p)\le\tfrac{1}{\sqrt{\sigma}\lambda_{\min}}\cdot\rho(\|z^b-z^c\|_M;z^b)\ ,\\
&\mathrm{Dist}(s^b,\mathbb{R}^n_+)\le\tfrac{1}{\sqrt{\tau}}\cdot\rho(\|z^b-z^c\|_M;z^b)\ ,\\
&\mathrm{Gap}(x^b,s^b)\le\max\{\|z^b-z^c\|_M,\|z^b\|_M\}\rho(\|z^b-z^c\|_M;z^b)\ .
\end{aligned} \tag{3.37}$$

Also, from Lemma 3.10 it follows that $\mathrm{Dist}_M(z^b,\mathcal{Z}^\star)$ can be bounded using the terms in the left-hand side of (3.37):

$$\mathrm{Dist}_M(z^b,\mathcal{Z}^\star)\le C_1\cdot\mathrm{Dist}(x^b,V_p)+C_2\cdot\mathrm{Dist}(s^b,\mathbb{R}^n_+)+C_3\cdot\mathrm{Gap}(x^b,s^b)\ , \tag{3.38}$$

where C_1, C_2 and C_3 are the scalars defined in (3.19). Substituting (3.37) into (3.38) yields:

$$\mathrm{Dist}_M(z^b,\mathcal{Z}^\star)\le\left(\frac{C_1}{\sqrt{\sigma}\lambda_{\min}}+\frac{C_2}{\sqrt{\tau}}+C_3\cdot\max\{\|z^b-z^c\|_M,\|z^b\|_M\}\right)\rho(\|z^b-z^c\|_M;z^b)\ . \tag{3.39}$$

From Lemma 2.5 it holds that

$$\begin{aligned}\max\{\|z^b - z^c\|_M, \|z^b\|_M\} &\le 2\,\mathrm{Dist}_M(z^0, \mathcal{Z}^\star) + \|z^0\|_M = 2\,\mathrm{Dist}_M(0, \mathcal{Z}^\star) \\ &\le \frac{2\sqrt{2}}{\sqrt{\tau}}\,\mathrm{Dist}(0, \mathcal{X}^\star) + \frac{2\sqrt{2}}{\sqrt{\sigma}\lambda_{\min}}\,\mathrm{Dist}(c, \mathcal{S}^\star)\ ,\end{aligned}$$

where the second inequality uses (2.13). Substituting this inequality into (3.39) yields

$$\mathrm{Dist}_M(z^b, \mathcal{Z}^\star) \le \left(\frac{C_1}{\sqrt{\sigma}\lambda_{\min}} + \frac{C_2}{\sqrt{\tau}} + \frac{2\sqrt{2}C_3}{\sqrt{\tau}}\,\mathrm{Dist}(0, \mathcal{X}^\star) + \frac{2\sqrt{2}C_3}{\lambda_{\min}\sqrt{\sigma}}\,\mathrm{Dist}(c, \mathcal{S}^\star)\right)\rho(\|z^b - z^c\|_M; z^b)\ . \tag{3.40}$$

The proof is completed by substituting the values of C_1, C_2, C_3 defined in (3.19) into (3.40), which then yields (3.11). □

4 Properties of the Limiting Error Ratio (LimitingER)

In this section we present some relevant properties of the limiting error ratio (LimitingER). Without loss of generality, we focus primarily on $\theta_p^\star$ and similar arguments could also be made for $\theta_d^\star$. Theorem 4.1 in Section 4.1 characterizes an upper bound on $\theta_p^\star$ that is connected to the notion of a "nicely interior" point in a convex set – which itself is critical to the complexity of separation-oracle methods [17]. Proposition 4.2 in Section 4.2 presents a convex optimization problem (actually a conic optimization problem with one second-order cone constraint) whose solution provides an upper bound on $\theta_p^\star$, thus showing that computing a bound on $\theta_p^\star$ is computationally tractable. Finally, Theorem 4.3 in Section 4.3 shows that the error ratio $\theta_p(x)$ is upper-bounded by a simple quantity involving the data-perturbation condition number DistInfeas(·) of Renegar [52]. Proofs of these results are presented in Appendix D.

Our setup once again is the LP problem (1.1) in which the feasible set is $\mathcal{F}_p$, the intersection of V_p and $\mathbb{R}^n_+$. We will assume in this subsection that $\mathcal{X}^\star$ is nonempty, and recall the definition of the limiting error ratio (LimitingER) $\theta_p^\star$ in (1.11). Let $\mathcal{F}_{++}$ denote the strictly feasible solutions of (1.1), namely $\mathcal{F}_{++} := V_p \cap \mathbb{R}^n_{++}$. We do not necessarily assume that $\mathcal{F}_{++} \neq \emptyset$.

4.1 An upper bound based on "nicely interior" feasible solutions

The following theorem presents an upper bound on the limiting error ratio $\theta_p^\star$ using the existence of a "nicely interior" point in $\mathcal{F}_{++}$. (In the theorem we use the convention that the infimum over an empty set is $+\infty$.)

Theorem 4.1. *For the LP problem (1.1), suppose that the optimal solution set $\mathcal{X}^\star$ is nonempty and bounded. Then*

$$\theta_p^\star \le \sup_{x^\star \in \mathcal{X}^\star}\ \inf_{x_{\mathrm{int}} \in \mathcal{F}_{++}} \frac{\|x^\star - x_{\mathrm{int}}\|}{\min_i (x_{\mathrm{int}})_i}\ . \tag{4.1}$$

This theorem states that if every optimal solution $x^\star$ has a nicely interior point near to it – in the sense that there exists $x_{\mathrm{int}} \in \mathcal{F}_{++}$ that is simultaneously close to $x^\star$ and far from the boundary of the nonnegative orthant $\mathbb{R}^n_+$, then the LimitingER value $\theta_p^\star$ will not be excessively large. In the case when $\mathcal{X}^\star$ is a singleton, then (4.1) simplifies to finding a single nicely interior point that balances the distance from the optimal solution (in the numerator above) with the distance to the boundary of $\mathbb{R}^n_+$ (in the denominator above). (Note that the concept of a nicely interior point is quite similar to that of a "reliable solution" in [15], see also [16] for connections to Renegar's data-perturbation condition number DistInfeas(·) [52].)

The same argument also holds for $\theta_d^\star$. We prove Theorem 4.1 in Appendix D for a generic form of LP, which encompasses both the primal and dual problems.

4.2 A computable upper bound for the limiting error ratio

Here we show how, in principle, we can use the upper bound in Theorem 4.1 to construct a computationally tractable convex optimization problem that computes an upper bound on the LimitingER $\theta_p^\star$.

We first suppose that we can compute, without too much extra computational effort, a ball that contains the optimal solution set $\mathcal{X}^\star$. That is, we suppose we can compute a point $x_a \in \mathcal{X}^\star$ and a radius value R_a such that $\mathcal{X}^\star \subset B(x_a, R_a)$, so that every optimal solution is within a distance R_a from the optimal solution x_a. If $\mathcal{X}^\star$ is a singleton, then $R_a = 0$ trivially. If $\mathcal{X}^\star$ is not a singleton, then one choice of x_a is the analytic center of $\mathcal{X}^\star$ (see Sonnevend [55], also [45]), from which one can then easily construct a bounding ellipsoid $\mathcal{E}^{\text{out}}$ that contains $\mathcal{X}^\star$ and then examine the eigenstructure $\mathcal{E}^{\text{out}}$ to compute a suitable value of R_a.

Proposition 4.2. *Suppose $x_a \in \mathcal{X}^\star$ and there exists R_a for which $\mathcal{X}^\star \subset \{x : \|x - x_a\| \le R_a\}$, then it holds that $\theta_p^\star \le G^\star$ for $G^\star$ defined as follows:*

$$G^\star := \inf_{r>0,\ x\in\mathbb{R}^n} \frac{R_a + \|x - x_a\|}{r} \quad \text{s.t. } x \in V_p,\ x \ge r \cdot e\ . \tag{4.2}$$

Furthermore, since $V_p = \{\hat{x} \in \mathbb{R}^n : A\hat{x} = b\}$, then

$$G^\star := \min_{v\in\mathbb{R}^n,\ \alpha\in\mathbb{R}} R_a\alpha + \|v - \alpha x_a\| \quad \text{s.t. } Av = \alpha b,\ v \ge e,\ \alpha \ge 0\ . \tag{4.3}$$

Proposition 1 in Section 1.3 is just a restatement of (4.2). The formulation of the upper bound in (4.3) is a convex optimization problem of essentially the same size as that of the original LP problem (1.1), and its only non-linear component is the norm term $\|v - x_a\alpha\|$ in the objective function. This can easily be handled by a single second-order cone constraint, or can be upper bounded by an ℓ_1 or ℓ_∞ norm which then can be converted to a pure LP problem.

We prove Proposition 4.2 in Appendix D for a generic form of LP, which encompasses both the primal and dual problems. As for computing the upper bound for $\theta_d^\star$, given s_a and R_a such that $s_a \in \mathcal{S}^\star$ and $\mathcal{S}^\star \subset B(s_a, R_a)$, the upper bound $G^\star$ in (4.2) could be computed by solving the optimization problem:

$$\min_{v\in\mathbb{R}^n,\ y\in\mathbb{R}^m,\ \alpha\in\mathbb{R}} R_a\alpha + \|v - \alpha s_a\| \quad \text{s.t. } A^\top y + v = \alpha c,\ v \ge e,\ \alpha \ge 0\ . \tag{4.4}$$

4.3 Relationship between the LimitingER and the distance to infeasibility

We have established the relationship between $\theta_p^\star$ and the geometric properties of the feasible sets. In this subsection, we demonstrate that this relationship also extends to the distance of the data from infeasibility. The concept of distance to infeasibility was initially utilized to assess the complexity of LP [52]. Previous research, such as [42, 24], has also studied the connections between global upper bounds of error bounds and the existence of perturbations. Here we show it also holds for the ER $\theta_p(x)$ and the LimitingER $\theta_p^\star$. We primarily focus on $\theta_p^\star$ for the primal problem for clarity, but similar arguments also hold for $\theta_d^\star$. In the appendix we will prove them for a generic form of LP, which encompasses both the primal and dual problems.

Note that V_p is given by $\{\hat{x} \in \mathbb{R}^n : A\hat{x} = b\}$ for an $m \times n$ real matrix A and a vector b in $\mathbb{R}^m$. Let $\text{SOLN}(A, b)$ denote the feasible set corresponding to (A, b), namely $\text{SOLN}(A, b) := \{\hat{x} \in \mathbb{R}^n :$

$A\hat{x} = b, x \geq 0\}$. We suppose that $\mathrm{SOLN}(A,b) \neq \emptyset$, in which case the "distance to infeasibility" of the data (A,b) is defined as follows:

$$\mathrm{DistInfeas}(A,b) := \inf \{\|\Delta A\| + \|\Delta b\| : \mathrm{SOLN}(A + \Delta A, b + \Delta b) = \emptyset\} \ ,$$

see [52]. Now we have the following general theorem about the relationship between $\theta_p(x)$ and the distance to infeasibility.

Theorem 4.3. *Suppose that $\mathcal{F}_p$ is nonempty for the LP (1.1). Then for every $x \in V_p \setminus \mathcal{F}_p$, it holds that*

$$\theta_p(x) \leq \frac{\|A\|(1 + \|u\|)}{\mathrm{DistInfeas}(A,b)} \ . \tag{4.5}$$

This theorem shows that the larger the distance to infeasibility is (namely, the larger the least data perturbation to infeasibility), the smaller the error ratio $\theta_p(x)$ must be. The inequality (4.5) looks similar in spirit to Theorem 1.1 part (1) of Renegar [52], even though the setup and context are structurally different from that considered here. From Theorem 4.3 we also have the following relationship between $\theta_p^\star$ and the distance to infeasibility.

Corollary 4.4. *Suppose that the LP (1.1) has an optimal solution. If $\mathrm{DistInfeas}(A,b) > 0$, then it holds that*

$$\theta_p^\star \leq \frac{\|A\|(1 + \max_{x \in \mathcal{X}^\star} \|x\|)}{\mathrm{DistInfeas}(A,b)} \ . \tag{4.6}$$

Both Theorem 4.3 and Corollary 4.4 imply that the farther the feasible set $\mathcal{F}_p$ is from infeasibility (namely, the larger $\mathrm{DistInfeas}(A,b)$ is), then the smaller $\theta_p(x)$ and $\theta_p^\star$ must be. It should be noted that these inequalities do not hold oppositely, because $\theta_p(x)$ and $\theta_p^\star$ are only determined by the geometry of the feasible set, while the $\mathrm{DistInfeas}(A,b)$ is affected by the data as well. For example, simultaneously rescaling a row of A and the corresponding entry of b by a small factor could decrease the value of $\mathrm{DistInfeas}(A,b)$ while keeping $\|A\|$ roughly unchanged. This would significantly increase the right-hand sides of (4.5) and (4.6), but the left-hand sides are unchanged because it does not affect the geometry of the feasible set.

The main idea of the proof of Theorem 4.3 is to construct, for each $x \in V_p \setminus \mathcal{F}_p$, a suitable perturbation $(\Delta A, \Delta b)$ of (A,b) for which $\mathrm{DistInfeas}(A + \Delta A, b + \Delta b) = 0$ and $\theta_p(x) \leq \frac{\|A\|(1+\|x\|)}{\|\Delta A\|+\|\Delta b\|}$.

5 LP Sharpness, stability under perturbation, and computation

In this section we present a characterization of the LP sharpness in terms of the least relative perturbation of the objective function vector that yields a different optimal solution set that is nonempty and not a subset of the original solution set. We also present a characterization of the LP sharpness via polyhedral geometry, with implications for computing methods. We still primarily focus on μ_p for the primal problem (1.1) while similar results also hold for μ_d because of the symmetric reformulation (2.3).

5.1 Characterization of LP sharpness and the stability of optimal solutions under perturbation

Let $\mathrm{OPT}(c, \mathcal{F}_p)$ denote the set of optimal solutions of the LP problem (1.1), namely

$$\mathrm{OPT}(c, \mathcal{F}_p) := \arg\min_{x \in \mathcal{F}_p} \ c^\top x \ .$$

Also let $C_{\mathcal{X}^\star}$ denote the recession cone of the optimal solution set $\mathcal{X}^\star$, and let $C^*_{\mathcal{X}^\star}$ denote its (positive) dual cone, namely $C^*_{\mathcal{X}^\star} = \{w : w^\top x \geq 0 \text{ for all } x \in C_{\mathcal{X}^\star}\}$. Then for any $\Delta c \in \vec{V}_p^{\perp}$, $\mathrm{OPT}(c, \mathcal{F}_p) = \mathrm{OPT}(c + \Delta c, \mathcal{F}_p)$. Additionally, if $\Delta c \notin C^*_{\mathcal{X}^\star}$, then $\mathrm{OPT}(c + \Delta c, \mathcal{F}_p) = \emptyset$ (since there exists $x \in C_{\mathcal{X}^\star}$ for which $c^\top x < 0$ and hence the resulting LP instance has unbounded objective value).

Intuition suggests that the LP sharpness value μ_p should be related to objective function perturbations that alter the set of optimal solutions. Indeed, we have the following theorem that characterizes this relationship completely, namely μ_p is the smallest relative perturbation Δc of c that yields a different optimal solution set that is nonempty and not a subset of the original optimal solution set. More precisely, we have:

Theorem 5.1. *Consider the general LP problem* (1.1) *under Assumption 1, and let μ_p be the LP sharpness of* (1.1)*. Then*

$$\mu_p = \inf_{\Delta g} \left\{ \frac{\|P_{\vec{V}_p}(\Delta c)\|}{\|P_{\vec{V}_p}(c)\|} : \mathrm{OPT}(c + \Delta c, \mathcal{F}_p) \neq \emptyset \ \ \textit{and} \ \mathrm{OPT}(c + \Delta c, \mathcal{F}_p) \not\subset \mathrm{OPT}(c, \mathcal{F}_p) \right\} . \tag{5.1}$$

Note that under Assumption 1 we must have $\|P_{\vec{V}_p}(c)\| > 0$ as otherwise all feasible solutions would be optimal (which violates Assumption 1). The proof of Theorem 5.1 in its generic form (which encompasses both the primal and dual problems) is in Appendix E.

5.2 Polyhedral geometry of LP sharpness, and computation

In this subsection we present a characterization of the LP sharpness via polyhedral geometry, with implications for computing the LP sharpness. Before we go into details, we first convey our general results as follows, using the the primal problem (1.1) as an example. If the optimal solution set is a singleton, namely $\mathcal{X}^\star = \{x^\star\}$, then we will show that the LP sharpness μ_p is the smallest sharpness along all of the edges of $\mathcal{F}_p$ emanating from $x^\star$. One implication of this result is that if the LP instance is primal and dual nondegenerate, then the dual nondegeneracy implies that $\mathcal{X}^\star$ is a singleton, and the primal nondegeneracy implies that there are exactly $n - m$ edges emanating from $\mathcal{X}^\star$, and hence it will be very easy to compute the LP sharpness. The more general result that we will show is that LP sharpness is the smallest sharpness along all edges of $\mathcal{F}_p$ that intersect $\mathcal{X}^\star$ but are not subsets of $\mathcal{X}^\star$. In the absence of nondegeneracy there can be exponentially many such edges, and so computing the LP sharpness for either a primal or dual degenerate instance is not a tractable problem in general.

We now develop these results more formally. For any $x \in \mathcal{F}_p \setminus \mathcal{X}^\star$, we define the sharpness of the point x to be:

$$G(x) := \frac{\mathrm{Dist}(x, V_p \cap H_p^\star)}{\mathrm{Dist}(x,\, \mathcal{X}^\star)} = \frac{\frac{c^\top (x - x^*)}{\|P_{\vec{V}_p}(c)\|}}{\|x - x^\star\|}$$

where in the above expression $x^\star := \arg\min_{v^\star \in \mathcal{X}^\star} \|v^\star - x\|$ is the projection of x onto $\mathcal{X}^\star$. With this notation the LP sharpness (1.12) is $\mu_p = \inf_{x \in \mathcal{F}_p \setminus \mathcal{X}^\star} G(x)$. Next let us recall some notation about convex polyhedra, see [20]. An edge of a polyhedron is a 1-dimensional face of the polyhedron. And since $\mathcal{F}_p$ is a polyhedron it follows that $\mathcal{F}_p$ will have a finite number of edges. Furthermore, every edge will either be (i) a line segment joining two different vertices $v^1 \neq v^2$ of $\mathcal{F}_p$ which we denote by $\boldsymbol{e} = [v^1, v^2]$, or (ii) a half-line points $v + \theta r$ for all $\theta \geq 0$, where v is a vertex of $\mathcal{F}_p$ and r is an extreme ray of $\mathcal{F}_p$, and which we denote by $\boldsymbol{f} = [v; r]$. We will be concerned with the subset of edges of $\mathcal{F}_p$ which have one endpoint in $\mathcal{X}^\star$ but are not subsets of $\mathcal{X}^\star$, which we call edges emanating away from $\mathcal{X}^\star$ and which we denote as $\mathcal{M}$, and whose formal definition is:

$$\begin{aligned} \mathcal{M} := \mathcal{M}_1 \cup \mathcal{M}_2 \qquad \text{where} \quad & \mathcal{M}_1 := \{\boldsymbol{e} = [v^1, v^2] : \boldsymbol{e} \text{ is an edge of } \mathcal{F}_p,\ v^1 \in \mathcal{X}^\star, v^2 \notin \mathcal{X}^\star\} \\ \text{and} \quad & \mathcal{M}_2 := \{\boldsymbol{f} = [v, r] : \boldsymbol{f} \text{ is an edge of } \mathcal{F}_p,\ v \in \mathcal{X}^\star, c^\top r = 1\} \ . \end{aligned}$$

The theorem below shows that the LP sharpness can be characterized using the following two functions:

$$\begin{aligned} & R_1(\boldsymbol{e}) := G(\tilde{x}) = \frac{\operatorname{Dist}(\tilde{x}, V_p \cap H_p^\star)}{\operatorname{Dist}(\tilde{x},\ \mathcal{X}^\star)} \text{ for all edges } \boldsymbol{e} = [x^\star, \tilde{x}] \in \mathcal{M}_1 \text{ , and} \\ & R_2(\boldsymbol{f}; \bar{\varepsilon}) := G(x^\star + \bar{\varepsilon} \cdot r) = \frac{\operatorname{Dist}(x^\star + \bar{\varepsilon} \cdot r, V_p \cap H_p^\star)}{\operatorname{Dist}(x^\star + \bar{\varepsilon} \cdot r,\ \mathcal{X}^\star)} \text{ for all edges } \boldsymbol{f} = [x^\star; r] \in \mathcal{M}_2 \text{ and all } \bar{\varepsilon} > 0 \ . \end{aligned}$$

Because $\mathcal{M}$ is a finite set, we can write $\mathcal{M} = \{\boldsymbol{e}^i : i = 1, 2, \dots, m_1\} \cup \{\boldsymbol{f}^j : j = 1, 2, \dots, m_2\}$ for some integers m_1, m_2.

Theorem 5.2. *For any given $\bar{\varepsilon} > 0$, the LP sharpness is characterized as follows:*

$$\mu_p = \min\left\{R_1(\boldsymbol{e}^1), R_1(\boldsymbol{e}^2), \dots, R_1(\boldsymbol{e}^{m_1}), R_2(\boldsymbol{f}^1; \bar{\varepsilon}), R_2(\boldsymbol{f}^2; \bar{\varepsilon}), \dots, R_2(\boldsymbol{f}^{m_2}; \bar{\varepsilon})\right\} \ . \tag{5.2}$$

We now discuss the implications of Theorem 5.2 for computing the LP sharpness measure μ_p. The computation of $R_1(\boldsymbol{e})$ or $R_2(\boldsymbol{f}; \bar{\varepsilon})$ for a given edge $\boldsymbol{e}$ or $\boldsymbol{f}$ requires computing a simple algebraic projection for the numerator of R_i and requires solving a convex quadratic program associated with the projection onto $\mathcal{X}^\star$ for the denominator of R_i. Note that for a given basis of an optimal solution $x^\star$, there are at most n edges emanating from $x^\star$, and all of them have a closed-form formulation [7]. Therefore, if all optimal bases have been enumerated, then the computation of μ_p via Theorem 5.2 is itself fairly straightforward. The key issue therefore is the enumeration of all optimal bases. For a general polyhedron [26] proves that enumerating all vertices is NP-hard if the polyhedron is unbounded. However, in some cases enumerating all optimal bases is easy. For example, when there is exactly one optimal basis (which does not often occur in real-world applications but is almost surely true for randomly generated problems), then solving the LP instance yields the unique optimal basis. Furthermore, when $\mathcal{X}^\star$ is bounded and all optimal bases are primal non-degenerate, [4] provides an algorithm for enumerating all optimal bases in $O(n^2 v)$ time complexity, where v is the number of optimal bases. Hence when the number of optimal bases is small, enumerating all of them remains tractable and so computing μ_p is also tractable.

The proof of Theorem 5.2 is in Appendix E. It is noted that the $R_i(\boldsymbol{e}^i)$ and $R_2(\boldsymbol{f}; \bar{\varepsilon})$ in the right-hand side of (5.2) are both purely geometric quantities. Therefore, the characterization of μ_p in Theorem 5.2 and its implications for computing μ_p also hold for μ_d in a symmetric way on the dual problem.

6 Numerical Experiments

Here we present results of numerical experiments designed to test how consistent our theoretical bounds are with computational practice, as well as to demonstrate the value of various heuristics on practical computation, based on our theoretical results. All computation was conducted on the MIT Engaging Cluster, and each experiment used a 2.4 GHz 14 Core CPU and 32G RAM, with CentOS version 7. All experiments were implemented in Julia 1.8.5.

6.1 Simple validation experiments

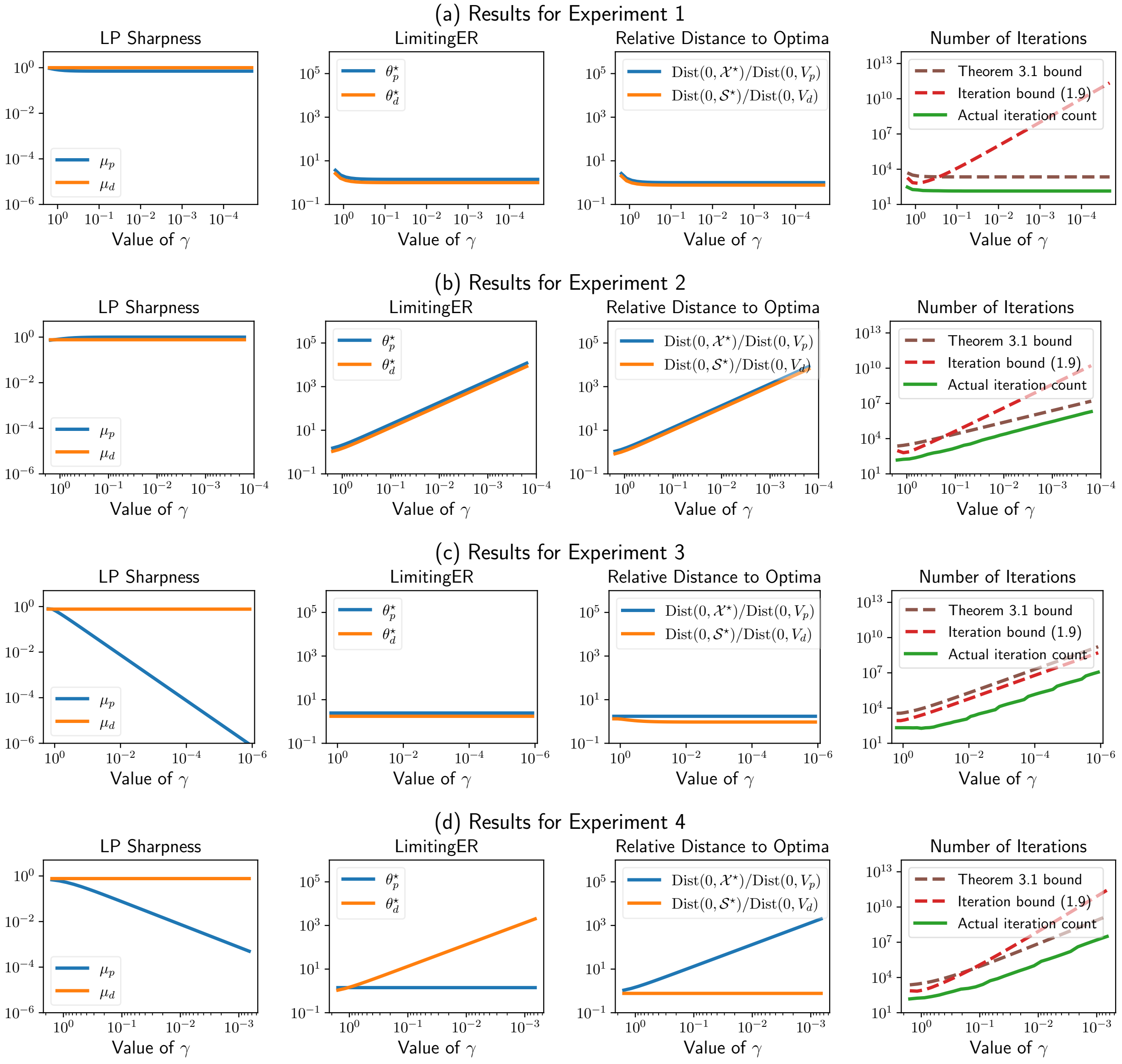
We conducted five simple experiments to test the extent to which the iteration bounds in Theorem 3.1 are "valid," by which we mean that the bounds are directionally consistent with computational practice on a specifically chosen family of test problems.

Experiment 1: Sensitivity to the Hoffman constant of the KKT system. The bounds in Theorem 3.1 are based essentially on three condition measures: LP sharpness, LimitingER, and (relative) distance to optima. This is in contrast to the analysis of [3] whose bounds are mostly based on the Hoffman constant of the KKT system of the LP instance. To test the sensitivity of Algorithm 1 to the Hoffman constant of the KKT system, we created the following family of LP instances in standard form (1.1) with $m = 1$, $n = 3$, and data (A^1, b^1, c^1) parameterized by $\gamma \in (0, 1]$ as follows: $A^1_\gamma := \left[\frac{\sin(\gamma)}{\sqrt{2}}, \cos(\gamma), \frac{\sin(\gamma)}{\sqrt{2}}\right]$, $b^1 := 1$, $c^1_\gamma := \left[\frac{\cos(\gamma)}{\sqrt{2}}, -\sin(\gamma), \frac{\cos(\gamma)}{\sqrt{2}}\right]^\top$. This family of problems was designed to have the following properties: $\|c\| = 1$, $Ac = 0$, $\|q\| = 1$, with uniform values of LP sharpness values and LimitingER for both the primal and dual problems, but with increasing values of the Hoffman constant of the KKT system as $\gamma \searrow 0$.

The first three columns of the first row of Figure 2 show the LP sharpness values (computed using the methodology in Section 5.2), LimitingER values (computed using the upper bound methodology in Section 4.2), and relative distances to optima for this simple family of problems, which are all constant over the range of $\gamma \in (0, 1]$. In the fourth column we report the actual iterations of Algorithm 1 (to obtain a solution whose Euclidean distance to the optimum is at most 10^{-10}), and the iteration bound of Theorem 3.1 as well as the iteration bound (1.9) based on [3]. (Since these two bounds are based on linear convergence rates, we report the constant outside of the logarithmic term for simplicity, and we computed the Hoffman constant for the KKT system using the algorithm and code from [49].) Notice that the bound (1.9) from [3] grows exponentially in $\ln(1/\gamma)$ while the actual number of iterations and the bound of Theorem 3.1 are constant over $\gamma \in (0, 1]$. This simple example validates the absence of the Hoffman constant from the bound in Theorem 3.1, and shows for this simple family that the actual number of iterations of Algorithm 1 is constant as suggested by Theorem 3.1.

Experiment 2: Sensitivity to LimitingER. This simple experiment is designed to test the sensitivity of Algorithm 1 to the LimitingER. Similar in approach to Experiment 1, we created the family: $A^2_\gamma := \left[\frac{\cos(\gamma)}{\sqrt{2}}, \sin(\gamma), \frac{\cos(\gamma)}{\sqrt{2}}\right]$, $b^2 = 1$, $c^2_\gamma := \left[\frac{\sin(\gamma)}{\sqrt{2}}, -\cos(\gamma), \frac{\sin(\gamma)}{\sqrt{2}}\right]^\top$, where again $\|c\| = 1$, $Ac = 0$, $\|q\| = 1$, with constant values of LP sharpness for $\gamma \in (0, 1]$, but now the LimitingER value increases as $\gamma \searrow 0$. The second row of Figure 2 shows our results. In this family of instances the LP sharpness values are constant even as the LimitingER grows. Notice that the relative distance to optima also grows similarly to the LimitingER; this must occur since the relative distances to optima are lower-bounded by the LimitingER , see [61]. The fourth column shows that for the smaller values of γ, the bound in Theorem 3.1 follows a similar pattern – including the slope in the log-log plot – as the actual iterations of Algorithm 1.

Experiment 3: Sensitivity to LP Sharpness. This simple experiment is designed to test the sensitivity of rPDHG to LP sharpness. We created the family: $A^3_\gamma := \left[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right]$, $b^3 = 1$, $c^3_\gamma := \cos(\gamma) \cdot \left[\frac{-1}{\sqrt{6}}, \frac{-1}{\sqrt{6}}, \frac{2}{\sqrt{6}}\right]^\top + \sin(\gamma) \cdot \left[\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right]^\top$. Similar to the previous experiments we have $\|c\| = 1$, $Ac = 0$, $\|q\| = 1$, with constant values of the LimitingER and the relative distances to optima for $\gamma \in (0, 1]$, but now the primal LP sharpness μ_p value decreases as $\gamma \searrow 0$. The third row of Figure 2 shows our results. Similar in spirit to Experiment 2, the fourth column shows that for the smaller values of γ that the bound in Theorem 3.1 follows a similar pattern – including the slope


(a) Results for Experiment 1

(b) Results for Experiment 2

(c) Results for Experiment 3

(d) Results for Experiment 4

Figure 2: Values of LP sharpness, LimitingER, relative distance to optima, theoretical iteration upper bound of Theorem 3.1, actual iteration count, and the iteration bound (1.9) based on [3] for the four simple validation experiments.

of the log-log plot – as the actual iterations of Algorithm 1.

Experiment 4: Sensitivity to simultaneous changes in LP sharpness and LimitingER. We created the family: $A^4_\gamma := \left[\sin(\gamma), \frac{\cos(\gamma)}{\sqrt{2}}, -\frac{\cos(\gamma)}{\sqrt{2}}\right]$, $b^4 = 1$, $c^4_\gamma := \left[0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^\top$, in which the primal LP sharpness μ_p decreases and the dual LimitingER $\theta^\star_d$ increases as $\gamma \searrow 0$, see the fourth row of Figure 2 for the computational values. Examining the fourth column of this row, we see the multiplicative effect of these two condition measures both on the theoretical bounds of Theorem 3.1 as well as a doubling of the slope of the log-log plot of actual iteration counts, which aligns well with the theoretical results.

Experiment 5: Effect of the step-size rule based on LP sharpness. Theorem 3.2 presented a step-size rule (3.3) based on knowledge of the LP sharpness measures μ_p and μ_d that leads to a structurally superior complexity bound for Algorithm 1. (However, this rule is impractical since the LP sharpness measures are neither known nor easily computable in practice.) In this experiment we test the utility of this rule using the simple family of LP instances $(A^4_\gamma, b^4, c^4_\gamma)$ described in Experiment 4, where for this simple family we know the LP sharpness measures. Figure 3 shows the theoretical upper bounds and the actual iteration numbers of Algorithm 1 with standard step-sizes (Theorem 3.1) and the step-sizes of Theorem 3.2 for this family of LP instances. The figure shows that this step-size rule reduces the actual number of iterations in line with the theory.

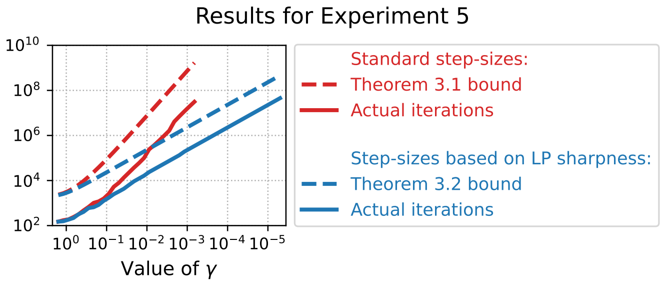


Figure 3: Theoretical upper bounds and the actual iteration numbers of Algorithm 1 with standard step-sizes (Theorem 3.1), and theoretical upper bounds and the actual iteration numbers of Algorithm 1 with the LP sharpness-based step-sizes of Theorem 3.2, for the family of LP instances described by $(A^4_\gamma, b^4, c^4_\gamma)$.

Experiment of Figure 1. Finally, we conducted an additional experiment on the family of LP instances given in (LP_γ), and the results are shown in Figure 1 in Section 1. In this experiment, the iteration bound from Theorem 3.1, the Hoffman constant, and the iteration bound (1.9) were all computed using the same approach as in the five experiments described above.

6.2 Computational evaluation of two theory-based heuristics on the MIPLIB 2017 dataset

In this subsection we introduce two heuristics that are inspired by our theoretical guarantees, and are designed to improve the practical performance of Algorithm 1. The first heuristic involves the choice of step-sizes τ and σ for Algorithm 1. It was observed in [3] that even while keeping the product of the primal and dual step-sizes τ and σ constant, heuristically modifying the ratio τ/σ had the potential to improve the computational performance of rPDHG. Theoretical justification for that observation can be seen in the computational bounds for Algorithm 1 in Theorem 3.1 using different step-sizes τ for the primal and σ for the dual in (3.3), and Theorem 3.2 shows – at least in concept –

how the complexity bound can be structurally improved by appropriately varying the ratio τ/σ while keeping the product constant, namely $\tau\sigma = 1/(4\lambda_{\max}^2)$. In the spirit of "learning from experience," our first heuristic is essentially an adaptation of the methodology in [3] to learn a reasonably good step-size ratio, and works as follows. We consider five possible choices of step-sizes, namely $(\tau, \sigma) = (40^{\ell}/2\lambda_{\max}, 40^{-\ell}/2\lambda_{\max})$ for $\ell = -1, -\frac{1}{2}, 0, \frac{1}{2}, 1$. For each of these step-size pairs we run Algorithm 1 for $5,000$ iterations from the same initial point $(x^{0,0}, y^{0,0}) = (0,0)$, and then choose which of the five step-sizes to use based on the smallest relative error $\mathcal{E}_r(x,y) := \frac{\|Ax^+ - b\|}{1+\|b\|} + \frac{\|(c-A^\top y)^-\|}{1+\|c\|} + \frac{|c^\top x^+ - b^\top y|}{1+|c^\top x^+|+|b^\top y|}$. The heuristic essentially spends $20,000$ iterations exploring/testing for a better step-size ratio. (We note that the relative error $\mathcal{E}_r(x,y)$ is upper-bounded by the distance to optima $\mathcal{E}_r d(x,s)$ (2.4), see Remark A.1.)

The second heuristic is also motivated by the computational bound in Theorem 3.1 where we observe in (3.4) that the bound grows at least linearly in the condition number κ of the matrix A (recall the definition of κ in (2.7)). The heuristic is to compute and apply a (full-rank) row-preconditioner $D \in \mathbb{R}^{m\times m}$ to the equality constraints $Ax = b$ to yield the equivalent system $DAx = Db$ for which the condition number $\kappa' := \kappa(DA) := \frac{\lambda^+_{\max}(DA)}{\lambda^+_{\min}(DA)}$ is reduced. Notice that for any such D, the preconditioned LP instance

$$\min_{x\in\mathbb{R}^n} \; c^\top x \quad \text{s.t. } DAx = Db, \; x \geq 0 \tag{6.1}$$

and its dual problem have the identical duality-paired symmetric format (2.3) as the original LP instance; and so the LP sharpness, the LimitingER, and the relative distance to optima are unchanged by the preconditioner. Indeed the only quantity in the iteration bound (3.4) that is changed is the matrix condition number $\kappa(DA)$. In our second heuristic we work with the "complete" pre-conditioner $D := (AA^\top)^{-1/2}$, for which $\kappa' = \kappa(DA) = 1$, which requires one (potentially expensive) matrix factorization. Other first-order methods for LP, such as [32, 47], also compute and use a single matrix factorization throughout all iterations. (When the problem is very large and computing even one matrix factorization is not tractable, [1] proposed to use a diagonal preconditioner D, but there were no theoretical guarantees.)

We tested the usefulness of the two heuristics using the LP relaxations of the MIPLIB 2017 dataset [18], which is a collection of mixed-integer programs from real applications. We took the LP relaxations of the problems in the dataset and converted them to standard form so that Algorithm 1 can be directly applied. We ran Algorithm 1 to compare the following choice of heuristic strategies for step-sizes and preconditioners:

- **Simple Step-size**: this is the simple step-size rule originally used in the proofs in [3], namely $\tau = \sigma = 1/(2\lambda_{\max})$,
- **Learned Step-size**: use an extra $20,000$ iterations to heuristically learn the best of five step-sizes as described above,
- **Preconditioner**: apply the preconditioner $D = (AA^\top)^{-1/2}$ as described above, and
- **Learned Step-size+Preconditioner**: apply both of the above heuristics.

Figure 4 illustrates the individual effects of the two heuristics on three representative problems, namely `nu120-pr9`, `n2seq36f` and `n3705`. The horizontal axis is the number of iterations and the vertical axis is the relative error $\mathcal{E}_r(x,y) := \frac{\|Ax^+ - b\|}{1+\|b\|} + \frac{\|(c-A^\top y)^-\|}{1+\|c\|} + \frac{|c^\top x^+ - b^\top y|}{1+|c^\top x^+|+|b^\top y|}$ computed using the original data of the LP instance for consistency. (The rather chaotic pattern of the early iterations of the Learned Step-size heuristic is due to the fact that the first $25,000$ iterations are used to test five different step-sizes.) For most of the LP instances in the MIPLIB 2017 we observed that the Learned Step-size heuristic enables much faster linear convergence, though `n3705` is an exception to this observation. We also observed that the preconditioner improves convergence significantly across all problems.

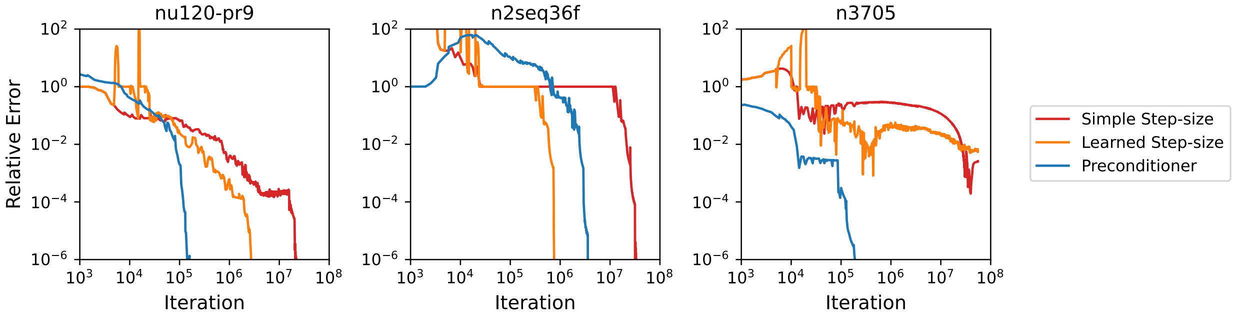


Figure 4: Performance of Algorithm 1 using two heuristic strategies, on problems `nu120-pr9`, `n2seq36f` and `n3705`.

Last of all, we tested all four combinations of heuristics on a large subset of the MIPLIB 2017 dataset, namely all LP relaxation problems in which $mn \le 10^9$, of which there are 574 such problems in total. For this evaluation we consider an LP instance to be "solved" if Algorithm 1 computes a solution (x, y) for which $\mathcal{E}_r(x, y) \le 10^{-4}$. Figure 5 shows the fraction of solved problems (of the 574 instances) on the horizontal axis, and the maximum iterations (leftmost plot) and the maximum runtime (rightmost plot). Notice that these two techniques both help the rPDHG solve more problems in a shorter time. Among the two heuristics, the preconditioner plays a prominent role in reducing the number of iterations, and also in reducing runtimes. Moreover, applying both heuristics is also valuable. Finally, it bears mentioning that if the problem is so large that the cost of working a matrix factorization is prohibitive, it is still possible to apply diagonal preconditioners to potentially improve the value of κ, see [1].

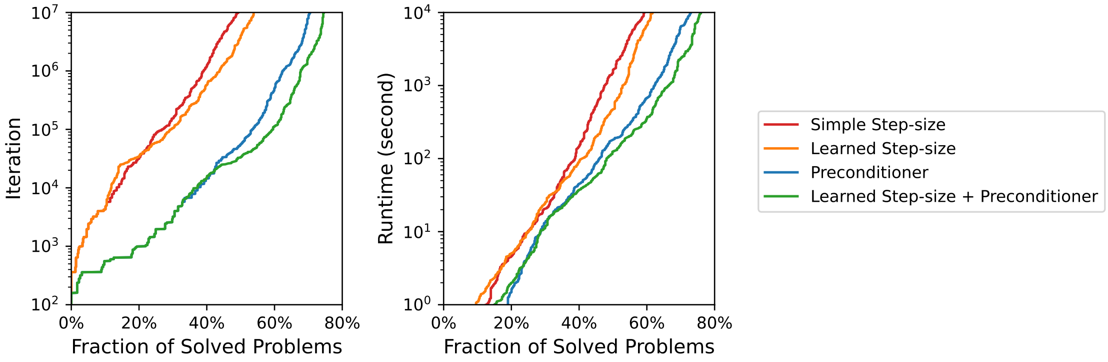


Figure 5: Performance of Algorithm 1 combined with heuristic strategies, on the 574 LP relaxation instances from the MIPLIB 2017 dataset.

The above experiments show that these two heuristics – which are motivated by our theoretical results – have clear potential to improve the practical performance of Algorithm 1, which also highlights the value of the theoretical understanding in the development of practical improvements in solution methods.

Appendix

A From the Distance to Optima to the Relative Error

Here we show that the relative error is upper-bounded by the distance to optima up to a scalar factor. For the pair (x, y) and $s := c - A^\top y$, defining $\mathcal{E}_r(x, s) := \mathcal{E}_r(x, y)$, it follows that $\mathcal{E}_r(x, s) = \frac{\|Ax^+ - b\|}{1+\|b\|} + \frac{\|s^-\|}{1+\|c\|} + \frac{|c^\top x^+ - q^\top(c-s)|}{1+|c^\top x^+| + |q^\top(c-s)|}$, in which $q := A^\top(AA^\top)^\dagger b$.

Remark A.1 (Relative error bounded by distance to optima)**.** *There exists a scalar constant $\bar{c}$ depending only on the data instance (A, b, c), such that for any solution pair (x, s), the relative error of (x, s) is upper bounded by the distance to optima by a factor of $\bar{c}$, namely $\mathcal{E}_r(x, s) \le \bar{c} \cdot \mathcal{E}_d(x, s)$ for any (x, s). One such value of $\bar{c}$ is $\bar{c} = c_0 := \frac{2\|A\|}{1+\|b\|} + 2\|c\| + \|q\| + 1$.*

Proof. The error $\mathcal{E}_r(x, s)$ is comprised of three parts, namely $\frac{\|Ax^+ - b\|}{1+\|b\|}$, $\frac{\|s^-\|}{1+\|c\|}$ and $\frac{|c^\top x^+ - q^\top(c-s)|}{1+|c^\top x^+| + |q^\top(c-s)|}$. For any $x^\star \in \mathcal{X}^\star$ and $s^\star \in \mathcal{S}^\star$, because $\|x - x^\star\| \ge \|x - x^+\|$, we have

$$\frac{\|Ax^+ - b\|}{1 + \|b\|} \le \frac{\|Ax - Ax^+\| + \|Ax - Ax^\star\|}{1 + \|b\|} \le \frac{\|A\| \cdot (\|x - x^\star\| + \|x - x^+\|)}{1 + \|b\|} \le \frac{2\|A\| \cdot \|x - x^\star\|}{1 + \|b\|}$$

and

$$\frac{|c^\top x^+ - q^\top(c - s)|}{1 + |c^\top x^+| + |q^\top(c - s)|} \le |c^\top x^+ - q^\top(c - s)| = |c^\top x^+ - c^\top x^\star - q^\top(c - s) + q^\top(c - s^\star)|$$
$$\le \|c\| \cdot (\|x - x^\star\| + \|x - x^+\|) + \|q\| \cdot \|s - s^\star\| \le 2\|c\| \cdot \|x - x^\star\| + \|q\| \cdot \|s - s^\star\| \ .$$

Similarly, because $\|s - s^+\| \le \|s - s^\star\|$, we have $\frac{\|s^-\|}{1+\|c\|} = \frac{\|s - s^+\|}{1+\|c\|} \le \frac{\|s - s^\star\|}{1+\|c\|} \le \|s - s^\star\|$. Now let $x^\star := \arg\min_{\hat{x} \in \mathcal{X}^\star} \|x - \hat{x}\|$ and $s^\star := \arg\min_{\hat{s} \in \mathcal{S}^\star} \|s - \hat{s}\|$, then combining the above three inequalities implies that $\mathcal{E}_r(x, s) \le \bar{c} \cdot \mathcal{E}_d(x, s)$. □

B Proofs for Section 2.2

B.1 Proof of Lemma 2.1

Proof of Lemma 2.1. We first examine primal near-feasibility. It holds trivially from the supposition that $\bar{x} \ge 0$ that $\mathrm{Dist}(\bar{x}, \mathbb{R}^n_+) = 0$. Let us now show the upper bound on $\mathrm{Dist}(\bar{x}, V_p)$. We assume that $\mathrm{Dist}(\bar{x}, V_p) > 0$ as otherwise the upper bound holds trivially. Let $\hat{x} = \arg\min_{x \in V_p} \|x - \bar{x}\|$ and hence $\mathrm{Dist}(\bar{x}, V_p) = \|\hat{x} - \bar{x}\|$. Note from the standard optimality conditions that $\hat{x} - \bar{x} \in \mathrm{Im}(A^\top)$ and hence there exists w such that $\hat{x} - \bar{x} = A^\top w$ and also $w \in \mathrm{Im}(A)$. It further holds that $A^\top w \ne 0$, since $\mathrm{Dist}(\bar{x}, V_p) > 0$.

From the definition of $\rho(r; \cdot)$ we have:

$$L(\bar{x}, y) - L(x, \bar{y}) \le r\rho(r; \bar{z}) \quad \text{for any } z \in \widetilde{B}(r; \bar{z}) \ . \tag{B.1}$$

Define $y := \bar{y} + \sqrt{\sigma} r \cdot w/\|w\|$ and set $z := (\bar{x}, y)$, whereby $z \in \widetilde{B}(r; \bar{z})$ and hence from (B.1) we have

$$r\rho(r; \bar{z}) \ge L(\bar{x}, y) - L(\bar{x}, \bar{y}) = (b - A\bar{x})^\top (y - \bar{y}) = (\hat{x} - \bar{x})^\top A^\top (y - \bar{y}) = w^\top AA^\top w \sqrt{\sigma} r/\|w\| \ .$$

It then follows that

$$\mathrm{Dist}(\bar{x}, V_p) = \|\hat{x} - \bar{x}\| = \|A^\top w\| \le \frac{\rho(r; \bar{z})}{\sqrt{\sigma}} \cdot \frac{\|w\|}{\|A^\top w\|} = \frac{\rho(r; \bar{z})}{\sqrt{\sigma}} \cdot \frac{\|w\|}{\|w\|_{AA^\top}} \le \frac{\rho(r; \bar{z})}{\sqrt{\sigma}\lambda_{\min}} \ ,$$

where the last inequality above follows since $\lambda_{\min} = \min_{v\in \mathrm{Im}(A)} \frac{\|v\|_{AA^\top}}{\|v\|}$. This proves item *1.*

Let us now examine dual near-infeasibility. Notice that by definition it holds that $\bar{s} \in V_d$. Define $x := \bar{x} + \sqrt{\tau} r \cdot (\bar{s})^- / \|(\bar{s})^-\|$ and set $z := (x, \bar{y})$, whereby $z \in \widetilde{B}(r; \bar{z})$ and hence from (B.1) we have

$$r\rho(r;\bar{z}) \geq L(\bar{x},\bar{y}) - L(x,\bar{y}) = (c - A^\top \bar{y})^\top(\bar{x} - x) = -\bar{s}^\top(\bar{s})^- \sqrt{\tau} r / \|(\bar{s})^-\| = \sqrt{\tau} r \|(\bar{s})^-\| \ ,$$

and hence $\mathrm{Dist}(\bar{s}, \mathbb{R}^n_+) = \|(\bar{s})^-\| \leq \frac{1}{\sqrt{\tau}} \cdot \rho(r;\bar{z})$. This proves item *2.*

Lastly, we examine the duality gap $\mathrm{Gap}(\bar{x},\bar{s}) = c^\top \bar{x} - b^\top \bar{y}$, and we consider two cases, namely $\bar{z} = 0$ and $\bar{z} \neq 0$. If $\bar{z} = 0$, then $\mathrm{Gap}(\bar{x},\bar{s}) = c^\top \bar{x} - b^\top \bar{y} = 0$, which satisfies the duality gap bound trivially. If $\bar{z} \neq 0$, then define $z := \bar{z} - \min\{\frac{r}{\|\bar{z}\|_M}, 1\}\bar{z}$, which satisfies $\|z - \bar{z}\|_M \leq r$. Substituting the $z := \bar{z} - \min\{\frac{r}{\|\bar{z}\|_M}, 1\}\bar{z}$ in (B.1) yields:

$$r\rho(r;\bar{z}) \geq L(\bar{x},y) - L(x,\bar{y}) = \min\left\{\frac{r}{\|\bar{z}\|_M}, 1\right\}(c^\top \bar{x} - b^\top \bar{y}) \ , \tag{B.2}$$

which simplifies to

$$c^\top \bar{x} - b^\top \bar{y} \leq \max\{r, \|\bar{z}\|_M\}\rho(r;\bar{z}) \ . \tag{B.3}$$

This proves the desired bound in item *3.* □

B.2 Proof of Lemma 2.2

We first recall the convergence result for PDHG in Remark 2 of [11].

Lemma B.1 (Sublinear convergence of PDHG (Remark 2 of [11]))**.** *Suppose that τ and σ satisfy* (2.6)*. For all $K \geq 1$, and for all $x \geq 0$ and y and $z = (x,y)$, it holds that*

$$L(\bar{x}^K, y) - L(x, \bar{y}^K) \leq \frac{\|z - z^0\|_M^2}{2K} \ . \tag{B.4}$$

The actual result in Remark 2 of [11] is slightly different than above, but the logic of the proof leads to (B.4) in our set-up for PDHG for LP.

Proof of Lemma 2.2. From the triangle inequality it holds that

$$\|z - z^0\|_M^2 \leq \left(\|z - \bar{z}^K\|_M + \|\bar{z}^K - z^0\|_M\right)^2 \ ,$$

which then implies via Lemma B.1 that every $z \in \widetilde{B}(\|\bar{z}^K - z^0\|_M; \bar{z}^K)$ satisfies

$$L(\bar{x}^K, y) - L(x, \bar{y}^K) \leq \frac{1}{2K}\|z - z^0\|_M^2 \leq \frac{\left(\|z - \bar{z}^K\|_M + \|\bar{z}^K - z^0\|_M\right)^2}{2K} \leq \frac{2}{K}\|\bar{z}^K - z^0\|_M^2 \ .$$

Therefore

$$\rho(\|\bar{z}^K - z^0\|_M; \bar{z}^K) \leq \frac{2\|\bar{z}^K - z^0\|_M}{K} \ . \tag{B.5}$$

For any $z^\star \in \mathcal{Z}^\star$, it follows from Lemma 2.4 that $\|\bar{z}^K - z^0\|_M \leq \|\bar{z}^K - z^\star\|_M + \|z^\star - z^0\|_M \leq 2\|z^\star - z^0\|_M$, which combines with (B.5) to prove the lemma. □

B.3 Proof of Lemma 2.5

Proof of Lemma 2.5. Let $z^\star := \arg\min_{z\in\mathcal{Z}^\star} \|z - z^a\|_M$, we have

$$\begin{array}{rl} & \max\{\|z^b - z^c\|_M, \|z^b\|_M\} \\ \leq & \max\{\|z^b - z^\star\|_M + \|z^c - z^\star\|_M, \|z^b - z^\star\|_M + \|z^\star\|_M\} \\ \leq & \max\{\|z^a - z^\star\|_M + \|z^a - z^\star\|_M, \|z^a - z^\star\|_M + \|z^\star\|_M\} \\ \leq & \max\{\|z^a - z^\star\|_M + \|z^a - z^\star\|_M, \|z^a - z^\star\|_M + \|z^a - z^\star\|_M + \|z^a\|_M\} \\ = & 2\|z^a - z^\star\|_M + \|z^a\|_M \ , \end{array} \tag{B.6}$$

where the second inequality uses the nonexpansive property (Lemma 2.4). This completes the proof. □

B.4 Proof of Proposition 2.6

We first prove the following elementary inequality:

Proposition B.2. *For any y and $s = c - A^\top y$, it holds that* $\mathrm{Dist}(y, \mathcal{Y}^\star) \leq \mathrm{Dist}(s, \mathcal{S}^\star) \cdot \frac{1}{\lambda_{\min}}$.

Proof. First observe that:

$$\mathrm{Dist}(s, \mathcal{S}^\star) = \mathrm{Dist}(c-A^\top y, \mathcal{S}^\star) = \mathrm{Dist}(c-A^\top y, c-A^\top(\mathcal{Y}^\star)) = \mathrm{Dist}(A^\top y, A^\top(\mathcal{Y}^\star)) = \mathrm{Dist}_{AA^\top}(y, \mathcal{Y}^\star) \ .$$

Let $AA^\top = PD^2P^\top$ denote the thin eigendecomposition of $AA^\top$, so that $P^\top P = I$ and D is the diagonal matrix of positive singular values of A, whereby $D_{ii} \geq \min_j D_{jj} = \lambda_{\min}$ for each i. Now let $y^\star$ solve the shortest distance problem from y to $\mathcal{Y}^\star$ in the norm $\|\cdot\|_{AA^\top}$, hence $y^* \in \mathcal{Y}^*$ and $\mathrm{Dist}_{AA^\top}(y, \mathcal{Y}^\star) = \|y - y^\star\|_{AA^\top}$, and let us write $y - y^\star = u + v$ where $u \in \mathrm{Im}(A)$ and $v \in \mathrm{Null}(A^\top)$. Then setting $\tilde{y} = y^\star + v$ and noting that $\tilde{y} \in \mathcal{Y}^\star$, we have:

$$\mathrm{Dist}_{AA^\top}(y, \mathcal{Y}^\star) \leq \|y - \tilde{y}\|_{AA^\top} = \|u\|_{AA^\top} \ . \tag{B.7}$$

Next notice that since $u \in \mathrm{Im}(A) = \mathrm{Im}(AA^\top)$, there exists π for which $u = AA^\top\pi$, and define $\lambda = D^2P^\top\pi$. It then follows that $u = P\lambda$, $\lambda = P^\top u$, and $\|u\| = \|\lambda\|$. We therefore have:

$$\begin{aligned} \mathrm{Dist}_{AA^\top}(y, \mathcal{Y}^\star)^2 &= (u+v)^\top AA^\top(u+v) \\ &= u^\top AA^\top u = \lambda^\top P^\top PD^2P^\top P\lambda = \lambda^\top D^2\lambda \geq \lambda_{\min}^2\|\lambda\|^2 \ , \end{aligned} \tag{B.8}$$

and hence $\mathrm{Dist}(s, \mathcal{S}^\star) = \mathrm{Dist}_{AA^\top}(y, \mathcal{Y}^\star) \geq \lambda_{\min}\|\lambda\| = \lambda_{\min}\|u\| \geq \lambda_{\min}\,\mathrm{Dist}(y, \mathcal{Y}^\star)$, where the second inequality uses (B.7). Rearranging completes the proof. □

Proof of Proposition 2.6. We first prove (2.12). For a given $\alpha > 0$, the statement "$\|z\|_M \geq \alpha\|z\|_N$ for any z" is equivalent to

$$\|z\|_M^2 - \alpha^2\|z\|_N^2 = z^\top Q_\alpha z \geq 0 \ , \quad \text{where } Q_\alpha := \begin{pmatrix} \frac{1-\alpha^2}{\tau}I_n & -A^\top \\ -A & \frac{1-\alpha^2}{\sigma}I_m \end{pmatrix} \ ,$$

and hence $\|z\|_M \geq \alpha\|z\|_N$ for any z if and only if $Q_\alpha \succeq 0$. A Schur complement argument then establishes that $Q_\alpha \succeq 0$ if and only if $(1-\alpha^2)^2/\sigma\tau \geq \lambda_{\max}^2$, which rearranges to $\alpha \leq \sqrt{1 - \sqrt{\tau\sigma}\lambda_{\max}}$ (where the right-hand side is well defined due to (2.6)). This establishes the first inequality in (2.12). For the second inequality, note that the statement "$\|z\|_M \leq \sqrt{2}\|z\|_N$ for any z" holds if and only $1/(\tau\sigma) \geq \lambda_{\max}^2$, which is satisfied due to (2.6)), completing the proof of (2.12).

Let us now prove (2.13). We have

$$\begin{aligned}\mathrm{Dist}_M(z,\mathcal{Z}^\star) &= \min_{\tilde z\in\mathcal{X}^\star\times\mathcal{Y}^\star}\|z-\tilde z\|_M \le \sqrt{2}\cdot\min_{\tilde z\in\mathcal{X}^\star\times\mathcal{Y}^\star}\|z-\tilde z\|_N \\ &\le \frac{\sqrt{2}}{\sqrt{\tau}}\mathrm{Dist}(x,\mathcal{X}^\star)+\frac{\sqrt{2}}{\sqrt{\sigma}}\mathrm{Dist}(y,\mathcal{Y}^\star) \le \frac{\sqrt{2}}{\sqrt{\tau}}\mathrm{Dist}(x,\mathcal{X}^\star)+\frac{\sqrt{2}}{\sqrt{\sigma}\lambda_{\min}}\mathrm{Dist}(s,\mathcal{S}^\star)\ ,\end{aligned} \tag{B.9}$$

where the first inequality utilities (2.12) and the third inequality uses Proposition B.2.

We next prove (2.14). Again using (2.12), we have $\mathrm{Dist}_M(z,\mathcal{Z}^\star)\ge\sqrt{1-\sqrt{\tau\sigma}\lambda_{\max}}\cdot\mathrm{Dist}_N(z,\mathcal{Z}^\star)$, and it also holds that $\mathrm{Dist}_N(z,\mathcal{Z}^\star)\ge\max\cdot\left\{\frac{1}{\sqrt{\tau}}\mathrm{Dist}(x,\mathcal{X}^\star),\frac{1}{\sqrt{\sigma}}\mathrm{Dist}(y,\mathcal{Y}^\star)\right\}$. Furthermore, $\mathrm{Dist}(s,\mathcal{S}^\star)=\mathrm{Dist}(A^\top y,A^\top(\mathcal{Y}^\star))\le\|A\|\cdot\mathrm{Dist}(y,\mathcal{Y}^\star)$ and $\|A\|=\lambda_{\max}$, which combined with the above two inequalities yields the proof of (2.14). □

B.5 Proof of Theorem 2.3

Proof of Theorem 2.3. Before proving Theorem 2.3, we first prove the following claim:

1. Primal near-feasibiltiy: $\mathrm{Dist}(\bar x^K,V_p)\le\frac{1}{\sqrt{\sigma}\lambda_{\min}}\cdot\frac{4\,\mathrm{Dist}_M(0,\mathcal{Z}^\star)}{K}$ and $\mathrm{Dist}(\bar x^K,\mathbb{R}^n_+)=0$,
2. Dual near-feasibility: $\mathrm{Dist}(\bar s^K,V_d)=0$ and $\mathrm{Dist}(\bar s^K,\mathbb{R}^n_+)\le\frac{1}{\sqrt{\tau}}\cdot\frac{4\,\mathrm{Dist}_M(0,\mathcal{Z}^\star)}{K}$, and
3. Duality gap: $\mathrm{Gap}(\bar x^K,\bar s^K)\le\frac{8\,\mathrm{Dist}_M(0,\mathcal{Z}^\star)^2}{K}$.

The upper bounds for the primal near-feasibility and dual near-feasibility follow directly from Lemmas 2.1 and 2.2. To prove the bound on the duality gap, let $r=\|\bar z^K-z^0\|_M$, and let $z^\star\in\mathcal{Z}^\star$. Then it follows from the duality gap bound in Lemma 2.1 that

$$\mathrm{Gap}(\bar x^K,\bar s^K)\ \le\max\{\|\bar z^K-z^0\|_M,\|\bar z^K\|_M\}\cdot\rho(\|\bar z^K-z^0\|_M;\bar z^K)\le\|\bar z^K\|_M\cdot\frac{4\|z^0-z^\star\|_M}{K}\ ,$$

where the second inequality above uses Lemma 2.2 and $z^0=(0,0)$. Now we apply Lemma 2.5 with $z^a=z^0=(0,0)$ and $z^b=z^c=\bar z^K$, which yields $\|\bar z^K\|_M\le 2\,\mathrm{Dist}_M(0,\mathcal{Z}^\star)$, and we obtain $\mathrm{Gap}(\bar x^K,\bar s^K)\le\frac{8\,\mathrm{Dist}_M(0,\mathcal{Z}^\star)^2}{K}$, which completes the proof of the claim.

Proposition 2.6 states that $\mathrm{Dist}_M(0,\mathcal{Z}^\star)$ is upper bounded by $\frac{\sqrt{2}}{\sqrt{\tau}}\mathrm{Dist}(0,\mathcal{X}^\star)+\frac{\sqrt{2}}{\sqrt{\sigma}\lambda_{\min}}\mathrm{Dist}(c,\mathcal{S}^\star)$. Substituting this upper bound into the claim proves Theorem 2.3. □

C Proofs for Section 3

C.1 Proof of Lemma 3.3

Proof of Lemma 3.3. We first bound the number of inner iterations k of rPDHG between restarts in the outer loop. When $n=0$ we have $k=1$. For $n\ge1$ we show that $k\le 5\mathcal{N}/\beta$. To see this, note that it follows from Lemma 2.2 that

$$\rho(\|\bar z^{n,k}-z^{n,0}\|_M;\bar z^{n,k})\le\frac{4\,\mathrm{Dist}_M(z^{n,0},\mathcal{Z}^\star)}{k}\ . \tag{C.1}$$

We may presume that $\rho(\|z^{n,0}-z^{n-1,0}\|_M;z^{n,0})\neq0$, for otherwise it follows from Lemma 2.1 that $z^{n,0}\in\mathcal{Z}^\star$, and Algorithm 1 would have terminated already in line **10**. Let us rewrite (C.1) as:

$$\frac{\rho(\|\bar z^{n,k}-z^{n,0}\|_M;\bar z^{n,k})}{\rho(\|z^{n,0}-z^{n-1,0}\|_M;z^{n,0})}\le\frac{4}{k}\cdot\frac{\mathrm{Dist}_M(z^{n,0},\mathcal{Z}^\star)}{\rho(\|z^{n,0}-z^{n-1,0}\|_M;z^{n,0})}\ . \tag{C.2}$$

It then follows from (C.2) and (3.9) that $k=\lceil 4\mathcal{N}/\beta\rceil$ suffices to ensure that condition (3.1) is satisfied. Since $\mathcal{N}\ge1$ and $\beta\in(0,1)$, such a k is no larger than $5\mathcal{N}/\beta$.

Next we prove an upper bound on the number of outer iterations. When $n = 0$ rPDHG restarts when $k = 1$, and it follows from Lemma 2.2 and inequality (2.13) that the initial normalized duality gap is upper bounded as follows:

$$\rho(\|\bar{z}^{0,1} - z^{0,0}\|_M; \bar{z}^{0,1}) \le 4\,\mathrm{Dist}_M(z^{0,0}, \mathcal{Z}^\star) \le 4\left(\frac{\sqrt{2}}{\sqrt{\tau}} + \frac{\sqrt{2}}{\sqrt{\sigma}\lambda_{\min}}\right)\mathcal{E}_d(x^{0,0}, s^{0,0})\ . \tag{C.3}$$

Now note from (2.14) that

$$\begin{aligned}\mathrm{Dist}_M(z^{n,0}, \mathcal{Z}^\star) &\ge \gamma \cdot \max\left\{\frac{\mathrm{Dist}(x^{n,0}, \mathcal{X}^\star)}{\sqrt{\tau}}, \frac{\mathrm{Dist}(s^{n,0}, \mathcal{S}^\star)}{\sqrt{\sigma}\lambda_{\max}}\right\} \\ &\ge \gamma \cdot \min\left\{\frac{1}{\sqrt{\tau}}, \frac{1}{\sqrt{\sigma}\lambda_{\max}}\right\} \cdot \mathcal{E}_d(x^{n,0}, s^{n,0})\ ,\end{aligned}$$

where $\gamma := \sqrt{1 - \sqrt{\sigma\tau}\lambda_{\max}}$. Substituting this inequality back into (3.9) yields:

$$\mathcal{E}_d(x^{n,0}, s^{n,0}) \le \frac{\mathcal{N}}{\gamma} \cdot \max\{\sqrt{\tau}, \sqrt{\sigma}\lambda_{\max}\} \cdot \rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0})\ . \tag{C.4}$$

According to the restart condition, we have $\rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0}) \le \beta \cdot \rho(\|z^{n-1,0} - z^{n-2,0}\|_M; z^{n-1,0})$ for each $n \ge 2$. And noting that $z^{1,0} = \bar{z}^{0,1}$, it follows that:

$$\begin{aligned}\rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0}) \ &\le \beta^{n-1} \cdot \rho(\|z^{1,0} - z^{0,0}\|_M; z^{1,0}) \\ &= \beta^{n-1} \cdot \rho(\|\bar{z}^{0,1} - z^{0,0}\|_M; \bar{z}^{0,1}) \le 4\beta^{n-1}\left(\frac{\sqrt{2}}{\sqrt{\tau}} + \frac{\sqrt{2}}{\sqrt{\sigma}\lambda_{\min}}\right)\mathcal{E}_d(x^{0,0}, s^{0,0})\ ,\end{aligned} \tag{C.5}$$

where the second inequality uses (C.3). Combining (C.4) and (C.5) yields:

$$\begin{aligned}\mathcal{E}_d(x^{n,0}, s^{n,0}) &\le \frac{\mathcal{N}}{\gamma} \cdot \max\{\sqrt{\tau}, \sqrt{\sigma}\lambda_{\max}\} \cdot 4\beta^{n-1} \cdot \left(\frac{\sqrt{2}}{\sqrt{\tau}} + \frac{\sqrt{2}}{\sqrt{\sigma}\lambda_{\min}}\right)\mathcal{E}_d(x^{0,0}, s^{0,0}) \\ &\le 4\beta^{n-1} \cdot \mathcal{N} \cdot \frac{\sqrt{2}}{\gamma} \cdot \left(\sqrt{\tau} + \sqrt{\sigma}\lambda_{\max}\right) \cdot \left(\frac{1}{\sqrt{\tau}} + \frac{1}{\sqrt{\sigma}\lambda_{\min}}\right) \cdot \mathcal{E}_d(x^{0,0}, s^{0,0})\ .\end{aligned} \tag{C.6}$$

Note that $\gamma = \sqrt{1 - \sqrt{\sigma\tau}\lambda_{\max}}$, whereby (C.6) implies that for any $\varepsilon > 0$, $\mathcal{E}_d(x^{n,0}, s^{n,0}) \le \varepsilon$ for all

$$n \ge \left\lceil \frac{\ln\left(4\mathcal{N} \cdot \frac{\sqrt{2}}{\sqrt{1 - \sqrt{\sigma\tau}\lambda_{\max}}} \cdot (\sqrt{\tau} + \sqrt{\sigma}\lambda_{\max}) \cdot \left(\frac{1}{\sqrt{\tau}} + \frac{1}{\sqrt{\sigma}\lambda_{\min}}\right) \cdot \mathcal{E}_d(x^{0,0}, s^{0,0}) \cdot \varepsilon^{-1}\right)}{\ln(\beta^{-1})} \right\rceil + 1\ , \tag{C.7}$$

which must be true when

$$n \ge \frac{\ln\left(4\mathcal{N} \cdot \frac{\sqrt{2}}{\sqrt{1 - \sqrt{\sigma\tau}\lambda_{\max}}} \cdot (\sqrt{\tau} + \sqrt{\sigma}\lambda_{\max}) \cdot \left(\frac{1}{\sqrt{\tau}} + \frac{1}{\sqrt{\sigma}\lambda_{\min}}\right) \cdot \mathcal{E}_d(x^{0,0}, s^{0,0}) \cdot \varepsilon^{-1} \cdot \beta^{-1}\right)}{\ln(\beta^{-1})} + 1\ . \tag{C.8}$$

The upper bound for the total number of PDHGStep iterations now follows from (C.8) and noting that the inner loop at $n = 0$ uses $k = 1$ iteration whereas for all $n \ge 1$ we have bounded the number of PDHGStep iterations by $k \le 5\mathcal{N}/\beta$. □

C.2 Proof of Theorem 3.2

Proof of Theorem 3.2. Substituting in the step-sizes (3.6) and using the norm equalities $\|P_{\vec{V}_p}(c)\| = \|c\|$, $\|P_{\vec{V}_p^\perp}(c)\| = 0$, $\|q\| = \mathrm{Dist}(0, V_p)$, and $\|c\| = \mathrm{Dist}(0, V_d)$ into the value of $\widetilde{\mathcal{N}}$ in (3.13) yields:

$$\widetilde{\mathcal{N}} = 2\sqrt{2}\kappa \left(\frac{4\theta_p^\star + 4\sqrt{2}\cdot \frac{\mathrm{Dist}(c,\mathcal{S}^\star)}{\mathrm{Dist}(0,V_d)}}{\mu_p} + \frac{3\theta_d^\star + 4\sqrt{2}\cdot \frac{\mathrm{Dist}(0,\mathcal{X}^\star)}{\mathrm{Dist}(0,V_p)}}{\mu_d} \right) . \tag{C.9}$$

Due to (3.13), the value of $\mathcal{N}$ specified in (3.8) is at least as large as $\widetilde{\mathcal{N}}$ in (C.9), whereby it holds that $\mathrm{Dist}_M(z^{n,0}, \mathcal{Z}^\star) \le \mathcal{N} \cdot \rho(\|z^{n,0} - z^{n-1,0}\|_M; z^{n,0})$ for the value of $\mathcal{N}$ specified in (3.8). Therefore condition (3.9) of Lemma 3.3 is satisfied, and it follows from Lemma 3.3 that T satisfies (3.10) with the value of $\tilde{c}$ specified in the statement of the lemma, namely:

$$T \ \le\ \frac{5}{\beta \ln(1/\beta)} \cdot \mathcal{N} \cdot \ln\left(\tilde{c} \cdot \mathcal{N} \cdot \left(\frac{\mathcal{E}_d(x^{0,0}, s^{0,0})}{\varepsilon} \right) \right) + 1 \ . \tag{C.10}$$

Substituting in the step-sizes (3.6) and $\beta = 1/e$ into the value of $\tilde{c}$ in Lemma 3.3 we find that: $\tilde{c} = 8e \cdot \left(1 + \kappa \frac{\mu_p \|c\|}{\mu_d \|q\|}\right)\left(1 + \frac{\mu_d \|q\|}{\mu_p \|c\|}\right)$, and using $\beta = 1/e$ we finally arrive at:

$$T \ \le\ 5e \cdot \mathcal{N} \cdot \ln\left(8e \cdot \mathcal{N} \cdot \left(\frac{\mathcal{E}_d(x^{0,0}, s^{0,0})}{\varepsilon} \right) \cdot \left(1 + \kappa \frac{\mu_p \|c\|}{\mu_d \|q\|} \right) \left(1 + \frac{\mu_d \|q\|}{\mu_p \|c\|} \right) \right) + 1 \ , \tag{C.11}$$

which completes the proof of the theorem. □

D Proofs for Section 4

First of all, we introduce the following generic LP format:

$$\mathcal{U}^\star := \arg\min_{u \in \mathbb{R}^n} \ g^\top u \quad \text{s.t. } u \in \mathcal{F} := V \cap \mathbb{R}^n_+ \ , \tag{D.1}$$

which generalizes the duality-paired LPs in (2.3) as specific instances. Let $\mathcal{U}^\star$, u, g, $\mathcal{F}$ and V be $\mathcal{X}^\star$, x, c, $\mathcal{F}_p$ and V_p, respectively, then (D.1) is the primal problem of (2.3). Let $\mathcal{U}^\star$, u, g, $\mathcal{F}$ and V be $\mathcal{S}^\star$, s, $-q$, $\mathcal{F}_d$ and V_d, respectively, then (D.1) is the dual problem of (2.3). We let $\mathcal{F}_{++}$ denote the strictly feasibible solutions of (D.1), let $f^\star$ denote the optimal objective value, let $\theta(u)$ denote the error ratio of $\mathcal{F}$ at u, and let $\theta^\star$ denote the LimitingER of (D.1).

Section 4 presents some properties of the LimitingER $\theta_p^\star$ for the primal problem (1.1). In Appendix D we prove their generalizations to the generic LP (D.1).

D.1 Proof of Theorem 4.1

In this subsection, we prove Theorem 4.1 by proving its generalization to the generic LP:

Theorem D.1. *For the generic LP (D.1), suppose that the optimal solution set $\mathcal{U}^\star$ is nonempty and bounded. Then*

$$\theta^\star \le \sup_{u^\star \in \mathcal{U}^\star} \ \inf_{u_{\mathrm{int}} \in \mathcal{F}_{++}} \frac{\|u^\star - u_{\mathrm{int}}\|}{\min_i (u_{\mathrm{int}})_i} \ . \tag{D.2}$$

Before proving Theorem D.1, we first introduce a more general result about $\theta(u)$. Suppose $u \in V \setminus \mathcal{F}$ and $u_{\text{int}} \in \mathcal{F}_{++}$, then the line segment from u to u_{int} will contain a unique point that lies on the boundary of $\mathcal{F}$, and let us denote this point by $\mathcal{F}(u; u_{\text{int}})$. More formally we have

$$\mathcal{F}(u; u_{\text{int}}) := \arg\min_{\tilde{u}} \{\|u - \tilde{u}\| : \tilde{u} \in \mathcal{F} \ , \ \tilde{u} := \lambda u_{\text{int}} + (1-\lambda)u \text{ for some } \lambda \in \mathbb{R}\} \ . \tag{D.3}$$

The following lemma will be used in our proof of Theorem D.1.

Lemma D.2. *For the general LP presentation* (D.1)*, suppose that Assumption 1 holds. Then for $u \in V \setminus \mathcal{F}$ and $u_{\text{int}} \in \mathcal{F}_{++}$, it holds that*

$$\theta(u) \le \frac{\|\mathcal{F}(u; u_{\text{int}}) - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \le \frac{\|u - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i}, \tag{D.4}$$

where $\mathcal{F}(u; u_{\text{int}})$ is given by (D.3)*.*

Proof. Suppose that u_{int} is any given strictly feasible point in $\mathcal{F}_{++}$. Let $r := \min_i (u_{\text{int}})_i$, and so $r > 0$. In the line segment connecting u_{int} and u, let $v := \mathcal{F}(u; u_{\text{int}})$ defined in (D.3). Then because $r > 0$ and $u \in V \setminus \mathcal{F}$, there exists $\lambda \in (0,1)$ for which $v = \lambda u_{\text{int}} + (1-\lambda)u$. Also we have $v \in \partial \mathbb{R}^n_+$ and there exists $i \in [n]$ such that $v_i = 0$, whereby $0 = v_i = \lambda (u_{\text{int}})_i + (1-\lambda)u_i$ and $u_i < 0$ and so:

$$\frac{(u_{\text{int}})_i}{|u_i|} = \frac{1-\lambda}{\lambda} = \frac{\|v - u_{\text{int}}\|}{\|v - u\|} \ . \tag{D.5}$$

And since $(u_{\text{int}})_i \ge r$ it follows that

$$\frac{\|v - u\|}{|u_i|} = \frac{\|v - u_{\text{int}}\|}{(u_{\text{int}})_i} \le \frac{\|v - u_{\text{int}}\|}{r} \ . \tag{D.6}$$

On the left-most term of (D.6) it follows from $u_i < 0$ that

$$\frac{\|v - u\|}{|u_i|} \ge \frac{\|v - u\|}{\|(u)^-\|} = \frac{\|v - u\|}{\text{Dist}(u, \mathbb{R}^n_+)} \ . \tag{D.7}$$

Combining (D.6) and (D.7) yields

$$\frac{\|u - \mathcal{F}(u; u_{\text{int}})\|}{\text{Dist}(u, \mathbb{R}^n_+)} = \frac{\|u - v\|}{\text{Dist}(u, \mathbb{R}^n_+)} \le \frac{\|v - u_{\text{int}}\|}{(u_{\text{int}})_i} \le \frac{\|v - u_{\text{int}}\|}{r} = \frac{\|\mathcal{F}(u; u_{\text{int}}) - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \ . \tag{D.8}$$

Noting that the numerator of the right-most term of (D.8) is bounded by

$$\|\mathcal{F}(u; u_{\text{int}}) - u_{\text{int}}\| \ \le \|\mathcal{F}(u; u_{\text{int}}) - u_{\text{int}}\| + \|\mathcal{F}(u; u_{\text{int}}) - u\| = \|u - u_{\text{int}}\| \ , \tag{D.9}$$

and the left-most term (D.8) satisfies $\frac{\|u - \mathcal{F}(u; u_{\text{int}})\|}{\text{Dist}(u, \mathbb{R}^n_+)} \ge \theta(u)$, we therefore have $\theta(u) \le \frac{\|\mathcal{F}(u; u_{\text{int}}) - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i}$. Last of all, since $\|u - u_{\text{int}}\| \ge \|\mathcal{F}(u; u_{\text{int}}) - u_{\text{int}}\|$, $\theta(u)$ is also further upper bounded by $\frac{\|u - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i}$, which completes the proof. □

Lemma D.2 shows that the error ratio $\theta(u)$ is upper-bounded by the ratio of the distance from u to u_{int} to the distance of u_{int} to the boundary of the nonnegative orthant. We now use Lemma D.2 to prove Theorem D.1.

Proof of Theorem D.1. If $\mathcal{F}_{++} = \emptyset$, then the right-hand side of (D.2) is equal to $+\infty$ so (D.2) is trivially true. We therefore consider the case when $\mathcal{F} \neq \emptyset$. For any optimal solution $u^\star \in \mathcal{U}^\star$ and a given associated strictly feasible solution $u_{\text{int}} \in \mathcal{F}_{++}$, let the $\{u^k\}_{k=1}^\infty$ be a sequence in V that converges to $u^\star$, whereby from Lemma D.2 it holds that

$$\theta(u^k) \leq \frac{\|u^k - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \leq \frac{\|u^\star - u_{\text{int}}\| + \|u^\star - u^k\|}{\min_i (u_{\text{int}})_i} \ .$$

Taking the limit as $k \to \infty$ on both sides, and noting that $\lim_{k\to\infty} \|u^\star - u^k\| = 0$, it thus follows that $\limsup_{k\to\infty} \theta(u^k) \leq \frac{\|u^\star - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i}$. And since u_{int} is any strictly feasible point, taking the infimum over all such u_{int} yields

$$\limsup_{k\to\infty} \theta(u^k) \leq \inf_{u_{\text{int}} \in \mathcal{F}_{++}} \frac{\|u^\star - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \ . \tag{D.10}$$

We now seek to prove:

$$\theta^\star := \lim_{\varepsilon \to 0} \sup_{u \in V,\, \text{Dist}(u, \mathcal{U}^\star) \leq \varepsilon} \theta(u) \leq \sup_{u^\star \in \mathcal{U}^\star} \inf_{u_{\text{int}} \in \mathcal{F}_{++}} \frac{\|u^\star - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \ . \tag{D.11}$$

If this were false, there would exist $\delta > 0$ and a sequence $\{\bar{u}^k\}_{k=1}^\infty$ in V such that $\text{Dist}(\bar{u}^k, \mathcal{U}^\star) \leq 1/k$ and $\theta(\bar{u}^k) \geq \delta + \sup_{u^\star \in \mathcal{U}^\star} \inf_{u_{\text{int}} \in \mathcal{F}_{++}} \frac{\|u^\star - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i}$. Note that the points in the sequence $\{\bar{u}^k\}_{k=1}^\infty$ all lie in the compact set $\{u : \text{Dist}(u, \mathcal{U}^\star) \leq 1\}$ as $\mathcal{U}^\star$ is convex, closed and bounded. Therefore there exists a subsequence of $\{\bar{u}^k\}_{k=1}^\infty$ that converges to a limit point in $\mathcal{U}^\star$. This violates (D.10), and so provides a contradiction, whereby (D.11) is true, thus completing the proof. □

Therefore, Theorem 4.1 follows as a special case of Theorem D.1, when (D.1) is taken to be (1.1).

D.2 Proof of Proposition 4.2

In this subsection, we prove Proposition 4.2 by proving its generalization to (D.1):

Proposition D.3. *Suppose $u_a \in \mathcal{U}^\star$ and there exists R_a for which $\mathcal{U}^\star \subset \{u : \|u - u_a\| \leq R_a\}$, then it holds that $\theta^\star \leq G^\star$ for $G^\star$ defined as follows:*

$$G^\star := \inf_{r > 0,\ u \in \mathbb{R}^n} \frac{R_a + \|u - u_a\|}{r} \quad \text{s.t. } \ u \in V,\ u \geq r \cdot e \ . \tag{D.12}$$

Furthermore, let $V = \{\hat{u} \in \mathbb{R}^n : A\hat{u} = b\}$, then

$$G^\star := \min_{v \in \mathbb{R}^n,\ \alpha \in \mathbb{R}} R_a \alpha + \|v - \alpha u_a\| \quad \text{s.t. } \ Av = \alpha b,\ v \geq e,\ \alpha \geq 0 \ . \tag{D.13}$$

Proof. From Theorem D.1 we have:

$$\begin{aligned}
\theta^\star &\leq \sup_{u^\star \in \mathcal{U}^\star} \inf_{u_{\text{int}} \in \mathcal{F}_{++}} \frac{\|u^\star - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \leq \sup_{u \in B(u_a, R_a)} \inf_{u_{\text{int}} \in \mathcal{F}_{++}} \frac{\|u - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \\
&\leq \sup_{u \in B(u_a, R_a)} \inf_{u_{\text{int}} \in \mathcal{F}_{++}} \frac{\|u - u_a\| + \|u_a - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \leq \inf_{u_{\text{int}} \in \mathcal{F}_{++}} \frac{R_a + \|u_a - u_{\text{int}}\|}{\min_i (u_{\text{int}})_i} \\
&= \inf_{r > 0,\ u \in \mathbb{R}^n} \frac{R_a + \|u - u_a\|}{r} \quad \text{s.t. } \ u \in V,\ u \geq r \cdot e \ ,
\end{aligned} \tag{D.14}$$

and notice that the final right-hand side is precisely $G^\star$, which proves (D.12). Next notice that, if $V = \{\hat{u} \in \mathbb{R}^n : A\hat{u} = b\}$, (D.12) and (D.13) are equivalent via the elementary projective transformations $u = v/\alpha$ and $(v, \alpha) = (u/r, 1/r)$ if we add the additional constraint $\alpha > 0$ to (D.13). However, since we are only interested in the optimal objective value of (D.12) and (D.13), solving the (D.13) yields the same optimal objective value as (D.12). □

D.3 Proofs of Theorem 4.3 and Corollary 4.4

In this section, we prove Theorem 4.3 and Corollary 4.4 by proving their generalizations to (D.1). Suppose that V in the generalized LP (D.1) is represented by $V = \{\hat{u} : Ku = h\}$ for $K \in \mathbb{R}^{m\times n}$ and $h \in \mathbb{R}^m$, and then the generalizations are as follows:

Theorem D.4. *Suppose that $\mathcal{F}$ is nonempty for* (D.1)*. Then for every $u \in V \setminus \mathcal{F}$, it holds that*

$$\theta(u) \leq \frac{\|K\|(1+\|u\|)}{\mathrm{DistInfeas}(K,h)} \ . \tag{D.15}$$

Corollary D.5. *Suppose that* (D.1) *has an optimal solution. If* $\mathrm{DistInfeas}(K,h) > 0$*, then it holds that*

$$\theta^\star \leq \frac{\|K\|(1+\max_{u\in\mathcal{U}^\star}\|u\|)}{\mathrm{DistInfeas}(K,h)} \ . \tag{D.16}$$

We first prove Corollary D.5 using Theorem D.4:

Proof of Corollary 4.4. By the definition of $\theta^\star$ and Theorem D.4, we have

$$\theta^\star = \lim_{\varepsilon\to 0}\left(\sup_{u\in V,\,\mathsf{Dist}(u,\mathcal{U}^\star)\leq\varepsilon} \theta(u)\right) \leq \lim_{\varepsilon\to 0}\left(\sup_{u\in V,\,\mathsf{Dist}(u,\mathcal{U}^\star)\leq\varepsilon}\left(\frac{\|K\|(1+\|u\|)}{\mathrm{DistInfeas}(K,h)}\right)\right) \ , \tag{D.17}$$

which is exactly (D.16). □

Then we prove Theorem D.4 through the approach of constructing, for each $u \in V \setminus \mathcal{F}$, a suitable perturbation $(\Delta K, \Delta h)$ of (K,h) for which $\mathrm{DistInfeas}(K+\Delta K, h+\Delta h) = 0$ and $\theta(u) \leq \frac{\|K\|(1+\|u\|)}{\|\Delta K\|+\|\Delta h\|}$. Before presenting the formal proof of the theorem, we establish several key properties of points $u \in V \setminus \mathcal{F}$.

Let $\bar{u} \in V \setminus \mathcal{F}$ be fixed and given, and let $\hat{u}$ be the projection of $\bar{u}$ onto $\mathcal{F}$, denoted as $\hat{u} := P_{\mathcal{F}}(\bar{u})$. Then $\hat{u}$ solves the following convex quadratic program:

$$\min_{u\in\mathbb{R}^n} \ \tfrac{1}{2}\|u-\bar{u}\|^2 \ , \quad \mathrm{s.t.}\ Ku = h,\ u \geq 0 \ , \tag{D.18}$$

whereby there exist multipliers $\hat{y}$ and $\hat{s}$ that together with $\hat{u}$ satisfy the KKT optimality conditions:

$$K\hat{u} = h,\ \hat{u} \geq 0,\ \hat{u} - \bar{u} = K^\top\hat{y} + \hat{s},\ \hat{s} \geq 0,\ \hat{u}^\top\hat{s} = 0 \ . \tag{D.19}$$

Note that since $\bar{u} \in V \setminus \mathcal{F}$, then $K\bar{u} = K\hat{u} = h$. In addition, the following proposition holds for $(\hat{u}, \hat{s}, \hat{y})$:

Proposition D.6. *For any $u \in \mathbb{R}^n_+$ it holds that $-\|\hat{u}-\bar{u}\|^2 = \hat{s}^\top\bar{u} < \hat{s}^\top\hat{u} = 0 \leq \hat{s}^\top u$.*

Proof. This first equality follows from (D.19) since $\|\hat{u}-\bar{u}\|^2 = (K^\top y + \hat{s})^\top(\hat{u}-\bar{u})$, $K(\hat{u}-\bar{u}) = 0$, and $\hat{u}^\top\hat{s} = 0$, whereby $\|\hat{u}-\bar{u}\|^2 = -\hat{s}^\top\bar{u}$. The first inequality follows trivially since $\hat{s}^\top\bar{u} = -\|\hat{u}-\hat{u}\|^2 < 0$ and $\hat{u}^\top\hat{s} = 0$. And the last inequality follows since $u \geq 0$ and $\hat{s} \geq 0$. □

Let H be the hyperplane defined as $H := \{u : \hat{s}^\top u = 0\}$. Proposition D.6 implies that H separates $\bar{u}$ and $\mathbb{R}^n_+$, and $\hat{u} \in H$. We denote the projection of $\bar{u}$ onto H as $\check{u}$, namely $\check{u} = P_H(\bar{u})$ which has closed form $\check{u} = \bar{u} - \frac{\hat{s}^\top\bar{u}}{\|\hat{s}\|^2}\cdot\hat{s}$. For simplicity of exposition we use a to denote $\check{u} - \bar{u}$ and use b to denote $\hat{u} - \bar{u}$, namely

$$a := \check{u} - \bar{u} = -\frac{\hat{s}^\top\bar{u}}{\|\hat{s}\|^2}\cdot\hat{s} \ , \quad b := \hat{u} - \bar{u} = K^\top\hat{y} + \hat{s} \ . \tag{D.20}$$

Proposition D.7. *For a and b defined in (D.20) it holds that*

1. $a = \frac{\|b\|^2}{\|\hat{s}\|^2} \cdot \hat{s}$ *,*
2. $\|b\| \geq \|a\| > 0$ *, and*
3. $a - \frac{\|a\|^2}{\|b\|^2} \cdot b = K^\top w$*, where* $w := -\frac{\|b\|^2}{\|\hat{s}\|^2} \cdot \hat{y}$ *.*

Proof. To prove item 1, note from Proposition D.6 that $-\|b\|^2 = -\|\hat{u} - \bar{u}\|^2 = \hat{s}^\top \bar{u}$, whereby $a = -\frac{\hat{s}^\top \bar{u}}{\|\hat{s}\|^2} \cdot \hat{s} = \frac{\|b\|^2}{\|\hat{s}\|^2} \cdot \hat{s}$.

To prove item 2, notice that since $Kb = (K\hat{u} - K\bar{u}) = h - h = 0$, it follows that $\|\hat{s}\|^2 = \|b - K^\top \hat{y}\|^2 = \|b\|^2 + \hat{y}^\top K K^\top \hat{y} \geq \|b\|^2$. And since we have from item 1 that $\|a\| = \frac{\|b\|^2}{\|\hat{s}\|}$, this implies $\|a\|/\|b\| = \|b\|/\|\hat{s}\| \leq 1$ which proves the first inequality in item 2. To prove that $\|a\| > 0$, note that we cannot have $\|b\| = 0$ since $\bar{u} \in V \setminus \mathcal{F}$ and $\hat{u} \in \mathcal{F}$, and we cannot have $\hat{s} = 0$, for otherwise Proposition D.6 would also imply that $b = \hat{u} - \bar{u} = 0$. Therefore from item 1 we have $\|a\| > 0$.

To prove item 3, we first show that $\|a\|^2 = b^\top a$. From item 1 and the definition of b, we have $a^\top b = \frac{\|b\|^2}{\|\hat{s}\|^2} \cdot \hat{s}^\top (\hat{s} + K^\top \hat{y})$. Since $K(\hat{s} + K^\top \hat{y}) = K\hat{u} - K\bar{u} = h - h = 0$, it follows that

$$\hat{s}^\top (\hat{s} + K^\top \hat{y}) = \|\hat{s}\|^2 + \hat{s}^\top K^\top \hat{y} + \hat{y}^\top K (\hat{s} + K^\top \hat{y}) = \|\hat{s} + K^\top \hat{y}\|^2 = \|b\|^2 \ , \tag{D.21}$$

and therefore $a^\top b = \frac{\|b\|^4}{\|\hat{s}\|^2} = \|a\|^2$ (from item 1). This proves $\|a\|^2 = b^\top a$ and then we have

$$\begin{aligned} a - \frac{\|a\|^2}{\|b\|^2} \cdot b = a - \frac{b^\top a}{\|b\|^2} \cdot b = \frac{\|b\|^2}{\|\hat{s}\|^2} \cdot \hat{s} - \frac{\|b\|^2 \cdot b^\top \hat{s}}{\|\hat{s}\|^2 \|b\|^2} \cdot b = \frac{\|b\|^2}{\|\hat{s}\|^2} \cdot \hat{s} - \frac{\|b\|^2 \cdot (\hat{s} + K^\top \hat{y})^\top \hat{s}}{\|\hat{s}\|^2 \|b\|^2} \cdot (\hat{s} + K^\top \hat{y}) \\ = \frac{\|b\|^2}{\|\hat{s}\|^2} \cdot \hat{s} \cdot \left(1 - \frac{(\hat{s} + K^\top \hat{y})^\top \hat{s}}{\|b\|^2}\right) - \frac{(\hat{s} + K^\top \hat{y})^\top \hat{s}}{\|\hat{s}\|^2} \cdot K^\top \hat{y} \ . \end{aligned} \tag{D.22}$$

Here, the second equality follows from item 1, and the third equality uses $b = \hat{s} + K^\top \hat{y}$. Substituting (D.21) into (D.22) yields $a - \frac{\|a\|^2}{\|b\|^2} \cdot b = -\frac{\|b\|^2}{\|\hat{s}\|^2} \cdot K^\top \hat{y} = K^\top w$, thus proving item 3. □

With the above propositions established, we now prove Theorem D.4.

Proof of Theorem D.4. Let $\bar{u} \in V \setminus \mathcal{F}$ be given. We will use all of the notation developed earlier in this subsection, including $\hat{u} = P_{\mathcal{F}}(\bar{u})$, the KKT multipliers $(\hat{y}, \hat{s})$, $H := \{u : \hat{s}^\top u = 0\}$, and $\check{u} = P_H(\bar{u}) = \bar{u} - \frac{\hat{s}^\top \bar{u}}{\|\hat{s}\|^2} \cdot \hat{s}$. Additionally, let a and b be as given in (D.20) and $w = -\frac{\|b\|^2}{\|\hat{s}\|^2} \cdot \hat{y}$ as specified in Proposition D.7.

We first examine the case when $\hat{y} \neq 0$, and hence $w \neq 0$. Let us consider the following perturbations of K and h:

$$\Delta K := \frac{\|a\|^2}{\|w\|^2 \|b\|^2} \cdot w (b - a)^\top, \quad \Delta h := \Delta K \bar{u} - \varepsilon w \ , \tag{D.23}$$

where $\varepsilon > 0$ is a small positive scalar. Then:

$$\begin{aligned} (K + \Delta K)^\top w = K^\top w + \Delta K^\top w = a - \frac{\|a\|^2}{\|b\|^2} \cdot b + \frac{\|a\|^2}{\|w\|^2 \|b\|^2} \cdot (b - a) \cdot w^\top w \\ = \left(1 - \frac{\|a\|^2}{\|b\|^2}\right) a = \left(1 - \frac{\|a\|^2}{\|b\|^2}\right) \cdot \frac{\|b\|^2}{\|\hat{s}\|} \cdot \hat{s} \geq 0 \ , \end{aligned} \tag{D.24}$$

where the second and the fourth equalities are due to Proposition D.7, and the final inequality is also due to Proposition D.7. Furthermore, we have

$$w^\top (h + \Delta h) = w^\top (K + \Delta K) \bar{u} - \varepsilon \|w\|^2 = \left(1 - \frac{\|a\|^2}{\|b\|^2}\right) \cdot \frac{\|b\|^2}{\|\hat{s}\|} \cdot \hat{s}^\top \bar{u} - \varepsilon \|w\|^2 < 0 \ , \tag{D.25}$$

where the strict inequality follows since $\hat{s}^\top \bar{u} < 0$ from Proposition D.6, $\|a\| \le \|b\|$ from Proposition D.7, and $\|w\| > 0$ by supposition for this case. Examining (D.24) and (D.25) yields $(K + \Delta K)^\top w \ge 0$ and $w^\top(h + \Delta h) < 0$, which implies via Farkas' lemma that $\mathrm{SOLN}(K + \Delta K, h + \Delta h) = \emptyset$, and hence $\mathrm{DistInfeas}(K, h) \le \|\Delta K\| + \|\Delta h\|$.

Let us now bound the size of $\|\Delta K\|$ and $\|\Delta h\|$. From (D.23) we have

$$\|\Delta K\| \le \frac{\|a\|^2}{\|w\|\|b\|^2} \cdot \|b - a\| \ . \tag{D.26}$$

Since $a - \frac{\|a\|^2}{\|b\|^2} \cdot b = K^\top w$, we also have $\|K\| \ge \left\| a - \frac{\|a\|^2}{\|b\|^2} \cdot b \right\| \cdot \frac{1}{\|w\|}$. Therefore

$$\|K\| \ge \left\| \|a\| \cdot \frac{a}{\|a\|} - \frac{\|a\|^2}{\|b\|} \cdot \frac{b}{\|b\|} \right\| \cdot \frac{1}{\|w\|} = \left\| \frac{\|a\|^2}{\|b\|} \cdot \frac{a}{\|a\|} - \|a\| \cdot \frac{b}{\|b\|} \right\| \cdot \frac{1}{\|w\|} = \frac{\|a\|}{\|b\|} \cdot \|b - a\| \cdot \frac{1}{\|w\|} \ , \tag{D.27}$$

where the first equality above follows from squaring both sides and rearranging terms using Proposition D.7. Combining (D.26) and (D.27) yields $\frac{\|\Delta K\|}{\|K\|} \le \frac{\|a\|}{\|b\|} = \frac{\|\check{u} - \bar{u}\|}{\|\hat{u} - \bar{u}\|} = \frac{\mathrm{Dist}(\bar{u}, H)}{\mathrm{Dist}(\bar{u}, \mathcal{F})}$. From Proposition D.6, because H separates $\hat{u}$ and $\mathbb{R}^n_+$, it follows that $\mathrm{Dist}(\bar{u}, H) \le \mathrm{Dist}(\bar{u}, \mathbb{R}^n_+)$, whereby $\frac{\|\Delta K\|}{\|K\|} \le \frac{\mathrm{Dist}(\bar{u}, \mathbb{R}^n_+)}{\mathrm{Dist}(\bar{u}, \mathcal{F})} = \frac{1}{\theta(\bar{u})}$. Moreover, since $\Delta h := \Delta K \bar{u} - \varepsilon w$, it follows that $\|\Delta h\| \le \|\Delta K\| \cdot \|\bar{u}\| + \varepsilon\|w\| \le \frac{\|\bar{u}\|}{\theta(\bar{u})}\|K\| + \varepsilon\|w\|$. Finally, we can add the inequalities $\theta(\bar{u})\|\Delta K\| \le \|K\|$ and $\theta(\bar{u})\|\Delta h\| \le \|u\|\|K\| + \theta(\bar{u})\varepsilon\|w\|$ which yields after rearranging $\theta(u) \le \frac{\|K\|(1 + \|u\| + \varepsilon \cdot \theta(\bar{u})\|w\|/\|K\|)}{\|\Delta K\| + \|\Delta h\|} \cdot$. And since $\mathrm{DistInfeas}(K, h) \le \|\Delta K\| + \|\Delta h\|$ we have $\theta(u) \le \frac{\|K\|(1 + \|u\| + \varepsilon \cdot \theta(\bar{u})\|w\|/\|K\|)}{\mathrm{DistInfeas}(K,h)}$. Taking the limit as $\varepsilon \to 0$ then proves the result in the case when $\hat{y} \neq 0$.

Next we consider the case when $\hat{y} = 0$. It follows from (D.19) that $\hat{u} - \bar{u} = \hat{s}$. Let $I := \{i : \hat{u}_i > 0\}$ and $J := [n] \setminus I$. Then we have $\hat{s}_I = 0$ and $\hat{u}_I = \bar{u}_I$, and $\hat{s}_J = -\bar{u}_J$. This implies that $\hat{u} = \bar{u}^+ = P_{\mathbb{R}^n_+}(\bar{u})$, and hence $\mathrm{Dist}(\bar{u}, \mathcal{F}) = \|\hat{u} - \bar{u}\| = \|\hat{s}\| = \mathrm{Dist}(\bar{u}, \mathbb{R}^n_+)$, and hence $\theta(\bar{u}) = \mathrm{Dist}(\bar{u}, \mathcal{F}) / \mathrm{Dist}(\bar{u}, \mathbb{R}^n_+) = 1$. Now let $\Delta K = -K$, and for any $\varepsilon > 0$ let Δh be any vector satisfying $\|\Delta h\| \le \varepsilon$ and $h + \Delta h \neq 0$. Then $(K + \Delta K, h + \Delta h) = (0, h + \Delta h)$ whereby $\mathrm{SOLN}(K + \Delta K, h + \Delta h) = \emptyset$. Therefore $\mathrm{DistInfeas}(K, h) \le \|\Delta K\| + \|\Delta h\| \le \|K\| + \varepsilon$ for all $\varepsilon > 0$, and thus $\mathrm{DistInfeas}(K, h) \le \|K\|$. Finally, we have in this case that $\theta(\bar{u}) = 1 \le \frac{\|K\|}{\mathrm{DistInfeas}(K,h)} \le \frac{\|K\|(1 + \|\bar{u}\|)}{\mathrm{DistInfeas}(K,h)}$, which completes the proof. □

Therefore, Theorem 4.3 and Corollary 4.4 follows as special cases of Theorem D.4 and Corollary D.5, when (D.1) is taken to be (1.1).

E Proofs for Section 5

E.1 Proof of Theorem 5.1

In this subsection, we prove Theorem 5.1 by proving its generalization to the generic LP (D.1). We let μ denote the LP sharpness of (D.1). The generalization is as follows:

Theorem E.1. *Consider the the generic LP* (D.1) *under Assumption 1, and let μ be the LP sharpness of* (D.1)*. Then*

$$\mu = \inf_{\Delta g} \left\{ \frac{\|P_{\vec{V}}(\Delta g)\|}{\|P_{\vec{V}}(c)\|} : \mathrm{OPT}(g + \Delta g, \mathcal{F}_g) \neq \emptyset \ \ \textit{and} \ \ \mathrm{OPT}(g + \Delta g, \mathcal{F}) \not\subset \mathrm{OPT}(g, \mathcal{F}) \right\} \ . \tag{E.1}$$

The proof of Theorem E.1 is divided into two parts, where each part proves an inequality version of (5.1) in one of the two possible directions of the inequality. The following lemma proves the "$\leq$" version of (5.1).

Lemma E.2. *Consider the general LP problem* (D.1) *under Assumption* 1*, and let* μ *be the LP sharpness of* (D.1)*. Then*

$$\mu \leq \inf_{\Delta g} \left\{ \frac{\|P_{\vec{V}}(\Delta g)\|}{\|P_{\vec{V}}(g)\|} : \mathrm{OPT}(g+\Delta g, \mathcal{F}) \neq \emptyset \ \ \textit{and}\ \mathrm{OPT}(g+\Delta g, \mathcal{F}) \not\subset \mathrm{OPT}(g, \mathcal{F}) \right\} \ . \tag{E.2}$$

Proof. Let Δg satisfy $\mathrm{OPT}(g+\Delta g,\mathcal{F}) \neq \emptyset$ and $\mathrm{OPT}(g+\Delta g,\mathcal{F}) \not\subset \mathrm{OPT}(g,\mathcal{F})$, and let $\bar{u} \in \mathrm{OPT}(g+\Delta g,\mathcal{F}) \setminus \mathrm{OPT}(g,\mathcal{F})$. Denote the optimal objective value hyperplane by $H^\star := \{u : g^\top u = f^\star\}$. Let $\breve{u} := P_{V\cap H^\star}(\bar{u}) = \bar{u} - \frac{g^\top \bar{u} - f^\star}{\|P_{\vec{V}}(g)\|^2} \cdot P_{\vec{V}}(g)$ and $\hat{u} := P_{\mathcal{U}^\star}(\bar{u})$. For simplicity, we use the notation $a := \breve{u} - \bar{u}$ and $b := \hat{u} - \bar{u}$. From Definition 1.2 we have:

$$\mu \leq \frac{\mathsf{Dist}(\bar{u}, V \cap H^\star)}{\mathsf{Dist}(\bar{u}, \mathcal{U}^\star)} = \frac{\|\breve{u} - \bar{u}\|}{\|\hat{u} - \bar{u}\|} = \frac{\|a\|}{\|b\|} \ . \tag{E.3}$$

Our goal then is to prove that

$$\frac{\|a\|}{\|b\|} \leq \frac{\|P_{\vec{V}}(\Delta g)\|}{\|P_{\vec{V}}(g)\|} \ , \tag{E.4}$$

and combining (E.4) with (E.3) will yield the proof.

Because $\breve{u} \in H^\star$ and $\hat{u} \in H^\star$, we have $g^\top \breve{u} = g^\top \hat{u}$, which implies that $g^\top a = g^\top b$. Furthermore, since $\bar{u} \in \mathrm{OPT}(g+\Delta g, \mathcal{F})$, we have $(g+\Delta g)^\top \bar{u} \leq (g+\Delta g)^\top \hat{u}$, which implies that $(g+\Delta g)^\top b \geq 0$. Substituting $g^\top a = g^\top b$ into $(g+\Delta g)^\top b \geq 0$, we obtain $\Delta g^\top b \geq -g^\top a$. It follows directly from Assumption 1 that $\|P_{\vec{V}}(g)\| > 0$, and dividing both sides of this last inequality by $\|P_{\vec{V}}(g)\|$ yields

$$\left(\frac{\Delta g}{\|P_{\vec{V}}(g)\|} \right)^\top b \geq -\left(\frac{g}{\|P_{\vec{V}}(g)\|} \right)^\top a \ . \tag{E.5}$$

Regarding the right-hand side of (E.5), note that $\breve{u} = u - \frac{g^\top \bar{u} - f^\star}{\|P_{\vec{V}}(g)\|^2} \cdot P_{\vec{V}}(g)$ and $a = -\frac{g^\top \bar{u} - f^\star}{\|P_{\vec{V}}(g)\|^2} \cdot P_{\vec{V}}(g)$, whereby:

$$-\left(\frac{g}{\|P_{\vec{V}}(g)\|} \right)^\top a = -\left(\frac{P_{\vec{V}}(g)}{\|P_{\vec{V}}(g)\|} \right)^\top a = \|a\| \ . \tag{E.6}$$

Regarding the left-hand side of (E.5), since $b = \hat{u} - \bar{u} \in \vec{V}$, it follows that $(\Delta g)^\top b = (P_{\vec{V}}(\Delta g))^\top b \leq \|P_{\vec{V}}(\Delta g)\| \|b\|$. Substituting this inequality and (E.6) back into (E.5) yields (E.4), which as noted earlier combines with (E.3) to complete the proof. □

Before proving the "$\geq$" direction, we first establish a simple proposition. For a convex set $\mathcal{S}$ let $C_{\mathcal{S}}$ denote the recession cone of S, and let $C^*_{\mathcal{S}}$ denote the corresponding (positive) dual cone.

Proposition E.3. *Let* $\bar{u} \in \mathcal{F} \setminus \mathcal{U}^\star$*, and let* $\hat{u} := P_{\mathcal{U}^\star}(\bar{u})$*, then*

- $(\hat{u} - \bar{u})^\top (u^\star - \hat{u}) \geq 0$ *for any* $u^\star \in \mathcal{U}^\star$*, and*
- $\hat{u} - \bar{u} \in C^*_{\mathcal{U}^\star}$.

Proof. The first assertion follows directly from the optimality conditions for the projection of $\bar{u}$ onto $\mathcal{U}^\star$. For the second assertion, observe that for any $v \in C_{\mathcal{U}^\star}$ and any $u \in \mathcal{U}^\star$ we have $u + \lambda v \in \mathcal{U}^\star$ for all $\lambda \geq 0$, whereby it follows from the first assertion that $(\hat{u} - \bar{u})^\top (u + \lambda v - \hat{u}) \geq 0$ for all $\lambda \geq 0$ and hence $(\hat{u} - \bar{u})^\top v \geq 0$. Since v is an arbitrary point in $C_{\mathcal{U}^\star}$ it holds that $\hat{u} - \bar{u} \in C^*_{\mathcal{U}^\star}$. □

Lemma E.4. *Consider the general LP problem* (D.1) *under Assumption 1, and let μ be the LP sharpness of* (D.1)*. Then*

$$\mu \geq \inf_{\Delta g} \left\{ \frac{\|P_{\vec{V}}(\Delta g)\|}{\|P_{\vec{V}}(g)\|} : \mathrm{OPT}(g+\Delta g, \mathcal{F}) \neq \emptyset \;\; \textit{and} \;\, \mathrm{OPT}(g+\Delta g, \mathcal{F}) \not\subset \mathrm{OPT}(g, \mathcal{F}) \right\} \ . \tag{E.7}$$

Proof. Note that by setting $\Delta g := -g$ that $\mathrm{OPT}(g+\Delta g,\mathcal{F}) = \mathcal{F} \not\subset \mathrm{OPT}(g,\mathcal{F}) = \mathcal{U}^\star$ under Assumption 1, and therefore the right-hand side of (E.7) is at most 1. Recall from the definition of LP sharpness that $\mu \leq 1$. Therefore in the special case when $\mu = 1$ then (E.7) holds trivially.

Let us therefore consider the case $\mu < 1$. For any given $\bar{u} \in \mathcal{F} \setminus \mathcal{U}^\star$, we will construct a perturbation Δg for which $\mathrm{OPT}(g+\Delta g,\mathcal{F}) \neq \emptyset$ and $\mathrm{OPT}(g+\Delta g,\mathcal{F}) \not\subset \mathrm{OPT}(g,\mathcal{F})$, and

$$\frac{\|P_{\vec{V}}(\Delta g)\|}{\|P_{\vec{V}}(g)\|} \leq \frac{\mathrm{Dist}(\bar{u}, V \cap H^\star)}{\mathrm{Dist}(\bar{u}, \mathcal{U}^\star)} \ , \tag{E.8}$$

which then implies (E.7). We proceed as follows. Let $\bar{u} \in \mathcal{F} \setminus \mathcal{U}^\star$ be given, and define $\check{u} := P_{V \cap H^\star}(\bar{u}) = \bar{u} - \frac{g^\top \bar{u} - f^\star}{\|P_{\vec{V}}(g)\|^2} \cdot P_{\vec{V}}(g)$ and $\hat{u} := P_{\mathcal{U}^\star}(u)$. Similar to the notation used in the proof of Lemma E.2, let $a := \check{u} - \bar{u}$ and $b := \hat{u} - \bar{u}$.

To construct the perturbation Δg we first define

$$\bar{b} := \frac{-g^\top b}{\|b\|^2} \cdot b \tag{E.9}$$

and construct the perturbation $\Delta g := t \cdot \bar{b}$, where t is the optimal objective value of the following optimization problem:

$$t \ := \ \max_{\tau} \ \tau \qquad \mathrm{s.t.} \ \ \hat{u} \in \mathrm{OPT}(g + \tau\bar{b}, \mathcal{F}\} \ . \tag{E.10}$$

We aim to show that $t \in (0,1]$. Towards the proof of this inclusion, let $\mathcal{E}1$ be the set of extreme points of $\mathcal{F}$ that are in $\mathcal{U}^\star$, and let $\mathcal{E}2$ be the set of extreme points of $\mathcal{F}$ that are not in $\mathcal{U}^\star$. Similarly, let $\mathcal{R}1$ be the set of extreme rays of $\mathcal{F}$ that are also extreme rays of $\mathcal{U}^\star$, and let $\mathcal{R}2$ be the set of extreme rays of $\mathcal{F}$ that are not also extreme rays of $\mathcal{U}^\star$. Note that $\mathcal{E}1$ and $\mathcal{E}2$ are finite sets, and $\mathcal{R}1$ and $\mathcal{R}2$ are also finite sets. We can therefore rewrite (E.10) as:

$$\begin{aligned}
\mathrm{OP}: \qquad t \ := \ \max_{\tau} \quad & \tau && && \text{(E.11)}\\
\mathrm{s.t.} \quad & \tau \cdot \bar{b}^\top (\hat{u} - v^i) && \leq -\, g^\top (\hat{u} - v^i) && \text{for each } v^i \in \mathcal{E}1 \qquad \text{(E.12)}\\
& \tau \cdot \bar{b}^\top (\hat{u} - v^i) && \leq -\, g^\top (\hat{u} - v^i) && \text{for each } v^i \in \mathcal{E}2 \qquad \text{(E.13)}\\
& \tau \cdot \bar{b}^\top r^i && \geq -\, g^\top r^i && \text{for each } r^i \in \mathcal{R}1 \qquad \text{(E.14)}\\
& \tau \cdot \bar{b}^\top r^i && \geq -\, g^\top r^i && \text{for each } r^i \in \mathcal{R}2 \qquad \text{(E.15)}
\end{aligned}$$

First observe that $-g^\top b = -g^\top(\hat{u} - \bar{u}) > 0$, whereby $\bar{b}$ is a positive scaling of b. Also notice that $\tau = 0$ is feasible for OP, because $\hat{u} \in \mathrm{OPT}(g,\mathcal{F})$. It implies that the right-hand sides of (E.12) and (E.13) are nonnegative and the right-hand sides of (E.14) and (E.15) are nonpositive. Next we show that $t \leq 1$. To see this, note that if $\tau > 1$ then

$$(g + \tau \cdot \bar{b})^\top (\hat{u} - \bar{u}) = (g + \tau \cdot \bar{b})^\top b = (g + \bar{b})^\top b + (\tau - 1)\bar{b}^\top b = (\tau - 1) \cdot (-g^\top b) > 0 \ ,$$

whereby $\hat{u} \notin \mathrm{OPT}(g + \tau\bar{b}, \mathcal{F})$ and thus τ is not feasible for (E.10).

It thus remains to show that $t > 0$, which we will demonstrate by examining the constraints of OP. Examining the constraints (E.12), when $v^i \in \mathcal{E}1$ the corresponding right-hand side is equal to 0 while $\bar{b}^\top(\hat{u} - v^i) \leq 0$ (from Proposition E.3), so these constraints are satisfied for all $\tau \geq 0$. Examining the constraints (E.13), when $v^i \in \mathcal{E}2$ the corresponding right-hand side is strictly positive so (E.13) is satisfied for all sufficiently small $\tau > 0$. Examining the constraints (E.14), when $r^i \in \mathcal{R}1$ the corresponding right-hand side is equal to 0 while $\bar{b}^\top r^i \geq 0$ (from Proposition E.3), so these constraints are satisfied for all $\tau \geq 0$. And examining (E.15), when $r^i \in \mathcal{R}2$ the corresponding right-hand side is strictly negative, so (E.15) is satisfied for all sufficiently small $\tau > 0$. Therefore, there exists $au > 0$ that satisfies all the constraints of OP, which implies that $t > 0$.

Now let us show that $\mathrm{OPT}(g + \Delta g, \mathcal{F}) \neq \emptyset$ and $\mathrm{OPT}(g + \Delta g, \mathcal{F}) \not\subset \mathrm{OPT}(g, \mathcal{F})$. It follows from (E.10) that $(g + \Delta g)^\top(\hat{u} - u) \leq 0$ for any $u \in \mathcal{F}$ and therefore $\hat{u} \in \mathrm{OPT}(g + \Delta g, \mathcal{F})$ and therefore $\mathrm{OPT}(g + \Delta g, \mathcal{F}) \neq \emptyset$. Now notice that when τ is optimal for (E.10) (and its equivalent formulation OP), there exists either $v^i \in \mathcal{E}2$ for which the corresponding constraint in (E.13) is active, or $r^i \in \mathcal{R}2$ for which the corresponding constraint in (E.15) is active (or both). In the former case, $v^i \in \mathrm{OPT}(g + \Delta g, \mathcal{F})$ and in the latter case $\hat{u} + r^i \in \mathrm{OPT}(g + \Delta g, \mathcal{F})$. And in either case, we have $\mathrm{OPT}(g + \Delta g, \mathcal{F}) \not\subset \mathrm{OPT}(g, \mathcal{F})$.

Last of all, because $g^\top a = g^\top b$, $a \in \vec{V}$, and $\bar{b} \in \vec{V}$, we have

$$\|P_{\vec{V}}(\Delta g)\| = t \cdot \|\bar{b}\| = t \cdot \frac{-g^\top b}{\|b\|} \leq \frac{-g^\top b}{\|b\|} = \frac{-g^\top a}{\|b\|} = \frac{-P_{\vec{V}}(g)^\top a}{\|b\|} \leq \frac{\|a\|}{\|b\|} \cdot \|P_{\vec{V}}(g)\| = \frac{\mathrm{Dist}(\bar{u}, V \cap H^\star)}{\mathrm{Dist}(\bar{u}, \mathcal{U}^\star)} \cdot \|P_{\vec{V}}(g)\| \, .$$

This shows (E.8) and completes the proof. □

Last of all, Theorem E.1 follows by combining Lemmas E.2 and E.4. Theorem 5.1 follows by applying Theorem E.1 to (1.1).

E.2 Proof of Theorem 5.2

Our proof of Theorem 5.2 relies on the following two elementary propositions. For any $\varepsilon \geq 0$ define S_ε to be the level set whose objective function value is exactly ε larger than the optimal objective value, namely $S_\varepsilon := \mathcal{F}_p \cap \{x : c^\top x = f^\star + \varepsilon\}$.

Proposition E.5. *If $\varepsilon_2 \geq \varepsilon_1 > 0$ and $S_{\varepsilon_i} \neq \emptyset$ for $i = 1, 2$, then $\inf_{x \in S_{\varepsilon_2}} G(x) \geq \inf_{x \in S_{\varepsilon_1}} G(x)$.*

Proof. For any given $x \in S_{\varepsilon_2}$, let $\hat{x} := P_{\mathcal{X}^\star}(x)$ and $t := \varepsilon_1/\varepsilon_2$. Then for $x_t := tx + (1 - t)\hat{x}$ it holds that $x_t \in S_{\varepsilon_1}$. Also notice that $\mathrm{Dist}(x_t, \ \mathcal{X}^\star) = \|x_t - \hat{x}\| = t \cdot \|x - \hat{x}\| = t \cdot \mathrm{Dist}(x, \mathcal{X}^\star)$ and $\mathrm{Dist}(x_t, V_p \cap H_p^\star) = (1 - t) \cdot \mathrm{Dist}(\hat{x}, V_p \cap H_p^\star) + t \cdot \mathrm{Dist}(x, V_p \cap H_p^\star) = t \cdot \mathrm{Dist}(x, V_p \cap H_p^\star)$. Therefore, $G(x_t) = \frac{t \cdot \mathrm{Dist}(x, V_p \cap H_p^\star)}{t \cdot \mathrm{Dist}(x, \ \mathcal{X}^\star)} = G(x)$. Since this equality holds for all $x \in S_{\varepsilon_2}$, it follows that $\inf_{x \in S_{\varepsilon_2}} G(x) \geq \inf_{x \in S_{\varepsilon_1}} G(x)$, which proves the proposition. □

Proposition E.6. *Let $x^\star \in \mathcal{X}^\star$ and $v \in \vec{V}$ satisfy $c^\top v > 0$. If $t_2 \geq t_1 > 0$ and $x^\star + t_i \cdot v \in \mathcal{F}_p$ for $i = 1, 2$, then $G(x^\star + t_1 \cdot v) \geq G(x^\star + t_2 \cdot v)$.*

Proof. For $t \geq 0$ define $g_1(t) := \mathrm{Dist}(x^\star + t \cdot v, \ V_p \cap H_p^\star)$ and $g_2(t) := \mathrm{Dist}(x^\star + t \cdot v, \ \mathcal{X}^\star)$. Then for $t > 0$ we have $G(x^\star + t \cdot v) = g_1(t)/g_2(t)$. Notice that $g_1(t) = \frac{t \cdot c^\top v}{\|P_{\vec{V}_p}(c)\| \|v\|}$ which is a nonnegative increasing linear function of t with $g_1(0) = 0$. Also notice that $g_2(t)$ is convex and nonnegative for $t \geq 0$, and $g_2(0) = 0$, whereby $g_2(t)$ is a monotonically increasing nonnegative convex function for $t \geq 0$. Therefore $g_2(t)/g_1(t)$ is monotonically increasing on $t > 0$, whereby $G(t) = g_1(t)/g_2(t)$ is monotonically decreasing on $t > 0$, which proves the lemma. □

Proof of Theorem 5.2. First notice from the definition of μ that

$$\mu_p \leq \min\left\{R_1(\boldsymbol{e}^1), R_1(\boldsymbol{e}^2), \ldots, R_1(\boldsymbol{e}^{m_1}), R_2(\boldsymbol{f}^1;\bar{\varepsilon}), R_2(\boldsymbol{f}^2;\bar{\varepsilon}), \ldots, R_2(\boldsymbol{f}^{m_2};\bar{\varepsilon})\right\} \ . \tag{E.16}$$

For $\varepsilon > 0$ let us consider the level set S_ε and suppose that $S_\varepsilon \neq \emptyset$, and let $\mathcal{ES}_\varepsilon$ denote the extreme points of S_ε. Then we claim that

$$\begin{aligned}\mu_p = \inf_{x\in\mathcal{F}_p\setminus\mathcal{X}^\star} G(x) &= \inf_{\varepsilon>0}\left(\inf_{x\in S_\varepsilon} G(x)\right) = \lim_{\varepsilon\to 0}\left(\inf_{x\in S_\varepsilon} G(x)\right) \\ &= \lim_{\varepsilon\to 0}\left(\frac{\frac{\varepsilon}{\|P_{\vec{V}_p}(c)\|}}{\sup_{x\in S_\varepsilon}\operatorname{Dist}(x,\mathcal{X}^\star)}\right) = \lim_{\varepsilon\to 0}\left(\frac{\frac{\varepsilon}{\|P_{\vec{V}_p}(c)\|}}{\max_{v^i\in\mathcal{ES}_\varepsilon}\operatorname{Dist}(v^i,\mathcal{X}^\star)}\right) \ .\end{aligned}$$

Here the first equality is the definition of μ_p, the second equality is a restatement of the first expression, and the third equality is due to the monotonicity property of the sharpness function on the level sets S_ε from Proposition E.5. To prove the fourth equality, observe that the numerator of $G(x)$ is $\operatorname{Dist}(x, V_p \cap H_p^\star)$ which equals the constant $\frac{\varepsilon}{\|P_{\vec{V}_p}(c)\|}$ for $x \in S_\varepsilon$, and the denominator of $G(x)$ is $\operatorname{Dist}(x,\mathcal{X}^\star)$. For the fifth equality, observe that $\operatorname{Dist}(\cdot,\mathcal{X}^\star)$ is convex in x and bounded from above and below on S_ε, and so attains its maximum at an extreme point of S_ε.

Next notice that since $\mathcal{F}_p$ is a polyhedron and S_ε is a level set of $\mathcal{F}_p$, there exists $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0,\bar{\varepsilon})$ the extreme points of S_ε all lie in the edges of $\mathcal{F}_p$ emanating away from $\mathcal{X}^\star$, namely $\mathcal{M} \cap S_\varepsilon = \mathcal{ES}_\varepsilon$. Therefore using the definition of $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$ we have:

$$\mu_p = \lim_{\varepsilon\to 0}\left(\frac{\frac{\varepsilon}{\|P_{\vec{V}_p}(c)\|}}{\max_{v^i\in\mathcal{ES}_\varepsilon}\operatorname{Dist}(v^i,\mathcal{X}^\star)}\right) = \min\left\{\min_{\boldsymbol{e}\in\mathcal{M}_1}\left(\inf_{x\in\boldsymbol{e}:c^\top x\leq f^\star+\bar{\varepsilon}} G(x)\right),\ \min_{\boldsymbol{f}\in\mathcal{M}_2}\left(\inf_{x\in\boldsymbol{f}:c^\top x\leq f^\star+\bar{\varepsilon}} G(x)\right)\right\} \ . \tag{E.17}$$

It follows from Proposition E.6 that $G(x)$ is decreasing on any edge emanating away from $\mathcal{X}^\star$. Thus for $\boldsymbol{e} = [v^1, v^2] \in \mathcal{M}_1$ with $v^1 \in \mathcal{X}^\star$ and $v^2 \notin \mathcal{X}^\star$ we have $\inf_{x\in\boldsymbol{e}:c^\top x\leq f^\star+\bar{\varepsilon}} G(x) \geq G(v^2) = R_1(\boldsymbol{e})$, and similarly for $\boldsymbol{f} = [v; r] \in \mathcal{M}_2$ with $v \in \mathcal{X}^\star$ and r being an extreme ray of $\mathcal{F}$ we have $\inf_{x\in\boldsymbol{f}:c^\top x\leq f^\star+\bar{\varepsilon}} G(x) = G(v+\bar{\varepsilon}r) = R_2(\boldsymbol{f},\bar{\varepsilon})$. Substituting these inequalities back into (E.17) yields

$$\mu_p \geq \min\left\{R_1(\boldsymbol{e}^1), R_1(\boldsymbol{e}^2), \ldots, R_1(\boldsymbol{e}^{m_1}), R_2(\boldsymbol{f}^1;\bar{\varepsilon}), R_2(\boldsymbol{f}^2;\bar{\varepsilon}), \ldots, R_2(\boldsymbol{f}^{m_2};\bar{\varepsilon})\right\} \ ,$$

which combined with (E.16) yields the proof. □

Acknowledgements

The authors are grateful to Kim-Chuan Toh, Haihao Lu, Jinwen Yang, and Guanghao Ye for discussions and input about the ideas in this paper and related research ideas. The author also thank the anonymous referees for helpful comments and suggestions that improved the quality of the manuscript.

References

[1] David Applegate, Mateo Díaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O'Donoghue, and Warren Schudy. Practical large-scale linear programming using primal-dual hybrid gradient. In *Advances in Neural Information Processing Systems*, volume 34, pages 20243–20257. Curran Associates, Inc., 2021.

[2] David Applegate, Mateo Díaz, Haihao Lu, and Miles Lubin. Infeasibility detection with primal-dual hybrid gradient for large-scale linear programming. *SIAM Journal on Optimization*, 34(1):459–484, 2024.

[3] David Applegate, Oliver Hinder, Haihao Lu, and Miles Lubin. Faster first-order primal-dual methods for linear programming using restarts and sharpness. *Mathematical Programming*, 201(1-2):133–184, 2023.

[4] David Avis and Komei Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete & Computational Geometry*, 8(3):295–313, 1992.

[5] Kinjal Basu, Amol Ghoting, Rahul Mazumder, and Yao Pan. Eclipse: An extreme-scale linear program solver for web-applications. In *International Conference on Machine Learning*, pages 704–714. PMLR, 2020.

[6] C. Bergthaller and Ivan Singer. The distance to a polyhedron. *Linear Algebra and its Applications*, 169:111–129, 1992.

[7] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization.* Athena Scientific, Belmont, MA, 1997.

[8] George W Brown and Tjalling C Koopmans. Computational suggestions for maximizing a linear function subject to linear inequalities. *Activity Analysis of Production and Allocation*, pages 377–380, 1951.

[9] James V Burke and Michael C Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

[10] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011.

[11] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[12] Kaihuang Chen, Defeng Sun, Yancheng Yuan, Guojun Zhang, and Xinyuan Zhao. HPR-LP: An implementation of an HPR method for solving linear programming. *arXiv preprint arXiv:2408.12179*, 2024.

[13] George Dantzig. *Linear Programming and Extensions.* Princeton University Press, 1963.

[14] Qi Deng, Qing Feng, Wenzhi Gao, Dongdong Ge, Bo Jiang, Yuntian Jiang, Jingsong Liu, Tianhao Liu, Chenyu Xue, Yinyu Ye, and Chuwen Zhang. An enhanced alternating direction method of multipliers-based interior point method for linear and conic optimization. *INFORMS Journal on Computing*, 2024.

[15] Marina Epelman and Robert M Freund. Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system. *Mathematical Programming*, 88(3):451–485, 2000.

[16] Robert M Freund and Jorge R Vera. Some characterizations and properties of the "distance to ill-posedness" and the condition measure of a conic linear system. *Mathematical Programming*, 86(2):225–260, 1999.

[17] Robert M Freund and Jorge R Vera. Equivalence of convex problem geometry and computational complexity in the separation oracle model. *Mathematics of Operations Research*, 34(4):869–879, 2009.

[18] Ambros Gleixner, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp Christophel, Kati Jarck, Thorsten Koch, Jeff Linderoth, Marco Lübbecke,

Hans D. Mittelmann, Derya Ozyurt, Ted K. Ralphs, Domenico Salvagnin, and Yuji Shinano. MIPLIB 2017: data-driven compilation of the 6th mixed-integer programming library. *Mathematical Programming Computation*, 13(3):443–490, 2021.

[19] Jean-Louis Goffin. The relaxation method for solving systems of linear inequalities. *Mathematics of Operations Research*, 5(3):388–414, 1980.

[20] Branko Grünbaum, Victor Klee, Micha A Perles, and Geoffrey Colin Shephard. *Convex Polytopes*, volume 16. Springer, 1967.

[21] Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2023.

[22] Oliver Hinder. Worst-case analysis of restarted primal-dual hybrid gradient on totally unimodular linear programs. *Operations Research Letters*, 57:107199, 2024.

[23] Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4), 1952.

[24] Hui Hu. Perturbation analysis of global error bounds for systems of linear inequalities. *Mathematical Programming*, 88:277–284, 2000.

[25] Yicheng Huang, Wanyu Zhang, Hongpei Li, Dongdong Ge, Huikang Liu, and Yinyu Ye. Restarted primal-dual hybrid conjugate gradient method for large-scale quadratic programming. *arXiv preprint arXiv:2405.16160*, 2024.

[26] Leonid Khachiyan, Endre Boros, Konrad Borys, Vladimir Gurvich, and Khaled Elbassioni. Generating all vertices of a polyhedron is hard. *Discrete & Computational Geometry*, 39:174–190, 2008.

[27] Carlton E Lemke. The constrained gradient method of linear programming. *Journal of the Society for Industrial and Applied Mathematics*, 9(1):1–17, 1961.

[28] Adrian S Lewis and Jong-Shi Pang. Error bounds for convex inequality systems. In *Generalized Convexity, Generalized Monotonicity: Recent Results*, pages 75–110. Springer, 1998.

[29] Bingheng Li, Linxin Yang, Yupeng Chen, Senmiao Wang, Haitao Mao, Qian Chen, Yao Ma, Akang Wang, Tian Ding, Jiliang Tang, and Ruoyu Sun. PDHG-unrolled learning-to-optimize method for large-scale linear programming. In *Proceedings of the 41st International Conference on Machine Learning*, pages 29164–29180, 2024.

[30] Xudong Li, Defeng Sun, and Kim-Chuan Toh. An asymptotically superlinearly convergent semismooth Newton augmented Lagrangian method for linear programming. *SIAM Journal on Optimization*, 30(3):2410–2440, 2020.

[31] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159:403–434, 2016.

[32] Tianyi Lin, Shiqian Ma, Yinyu Ye, and Shuzhong Zhang. An ADMM-based interior-point method for large-scale linear programming. *Optimization Methods and Software*, 36(2-3):389–424, 2021.

[33] Zhenwei Lin, Zikai Xiong, Dongdong Ge, and Yinyu Ye. PDCS: A primal-dual large-scale conic programming solver with GPU enhancements. *arXiv preprint arXiv:2505.00311*, 2025.

[34] Haihao Lu and Jinwen Yang. Nearly optimal linear convergence of stochastic primal-dual methods for linear programming. *arXiv preprint arXiv:2111.05530*, 2021.

[35] Haihao Lu and Jinwen Yang. On the infimal sub-differential size of primal-dual hybrid gradient method. *arXiv preprint arXiv:2206.12061*, 2022.

[36] Haihao Lu and Jinwen Yang. cuPDLP. jl: a GPU implementation of restarted primal-dual hybrid gradient for linear programming in Julia. *arXiv preprint arXiv:2311.12180*, 2023.

[37] Haihao Lu and Jinwen Yang. A practical and optimal first-order method for large-scale convex quadratic programming. *arXiv preprint arXiv:2311.07710*, 2023.

[38] Haihao Lu and Jinwen Yang. On the geometry and refined rate of primal–dual hybrid gradient for linear programming. *Mathematical Programming*, pages 1–39, 2024.

[39] Haihao Lu and Jinwen Yang. PDOT: A practical primal-dual algorithm and a GPU-based solver for optimal transport. *arXiv preprint arXiv:2407.19689*, 2024.

[40] Haihao Lu and Jinwen Yang. Restarted Halpern PDHG for linear programming. *arXiv preprint arXiv:2407.16144*, 2024.

[41] Haihao Lu, Jinwen Yang, Haodong Hu, Qi Huangfu, Jinsong Liu, Tianhao Liu, Yinyu Ye, Chuwen Zhang, and Dongdong Ge. cuPDLP-C: A strengthened implementation of cuPDLP for linear programming by C language. *arXiv preprint arXiv:2312.14832*, 2023.

[42] Zhi-Quan Luo and Paul Tseng. Perturbation analysis of a condition number for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 15(2):636–660, 1994.

[43] Olvi L Mangasarian. A condition number for linear inequalities and linear programs. In *Proceedings of 6th Symposium über Operations Research, Universität Augsburg*, pages 3–15, 1981.

[44] Vahab Mirrokni. Google Research, 2022 & beyond: Algorithmic advances. 2023. https://ai.googleblog.com/2023/02/google-research-2022-beyond-algorithmic.html.

[45] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[46] Brendan O'Donoghue. Operator splitting for a homogeneous embedding of the linear complementarity problem. *SIAM Journal on Optimization*, 31(3):1999–2023, 2021.

[47] Brendan O'donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169:1042–1068, 2016.

[48] Jong-Shi Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1-3):299–332, 1997.

[49] Javier Pena, Juan Vera, and Luis Zuluaga. An algorithm to compute the Hoffman constant of a system of linear constraints. *arXiv preprint arXiv:1804.08418*, 2018.

[50] Javier Pena, Juan C Vera, and Luis F Zuluaga. New characterizations of hoffman constants for systems of linear constraints. *Mathematical Programming*, 187:79–109, 2021.

[51] Boris Teodorovic Polyak. Sharp minima. In *Proceedings of the IIASA Workshop on Generalized Lagrangians and Their Applications, Laxenburg, Austria. Institute of Control Sciences Lecture Notes, Moscow*, 1979.

[52] James Renegar. Some perturbation theory for linear programming. *Mathematical Programming*, 65(1-3):73–91, 1994.

[53] Ernest K Ryu and Wotao Yin. *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge University Press, 2022.

[54] Michel Schubiger, Goran Banjac, and John Lygeros. GPU acceleration of ADMM for large-scale quadratic programming. *Journal of Parallel and Distributed Computing*, 144:55–67, 2020.

[55] Gy. Sonnevend. An 'analytic' center for polyhedrons and new classes of global algorithms for linear (smooth, convex) optimization. Technical report, Department of Numerical Analysis, Institute of Mathematics, Eötvös University, 1088, Budapest, Muzeum Körut 6-8, 1985. Preprint.

[56] Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Stephen Boyd. OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.

[57] Michael J Todd and Yinyu Ye. A centered projective algorithm for linear programming. *Mathematics of Operations Research*, 15(3):508–529, 1990.

[58] Stephen J Wright. *Primal-Dual Interior-Point Methods.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997.

[59] Zikai Xiong. Accessible theoretical complexity of the restarted primal-dual hybrid gradient method for linear programs with unique optima. *arXiv preprint arXiv:2410.04043*, 2024.

[60] Zikai Xiong. High-probability polynomial-time complexity of restarted PDHG for linear programming. *arXiv preprint arXiv:2501.00728*, 2025.

[61] Zikai Xiong and Robert M Freund. On the relation between LP sharpness and limiting error ratio and its complexity implications for restarted PDHG. *arXiv preprint arXiv:2312.13773*, 2023.

[62] Zikai Xiong and Robert M Freund. The role of level-set geometry on the performance of PDHG for conic linear optimization. *arXiv preprint arXiv:2406.01942*, 2024.

[63] Tianbao Yang and Qihang Lin. RSG: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(1):236–268, 2018.

[64] Guus Zoutendijk. *Methods of Feasible Directions: A Study in Linear and Non-Linear Programming.* Elsevier, Amsterdam, 1960.