Active set identification and rapid convergence for degenerate primal-dual problems

Mateo Díaz* Pedro Izquierdo Lehmann* Haihao Lu† Jinwen Yang‡

Abstract

Primal-dual methods for solving convex optimization problems with functional constraints often exhibit a distinct two-stage behavior. Initially, they converge towards a solution at a sublinear rate. Then, after a certain point, the method identifies the active set—in the sense that all subsequent iterates share the same active constraints—and the convergence enters a faster local linear regime. Theory characterizing this phenomenon spans over three decades. However, most existing work only guarantees eventual identification of the active set and relies heavily on nondegeneracy conditions, such as strict complementarity, which often fail to hold in practice. We characterize mild conditions on the problem geometry and the algorithm under which this phenomenon provably occurs. Our guarantees are entirely nonasymptotic and, importantly, do not rely on strict complementarity. Our framework encompasses several widely-used algorithms, including the proximal point method, the primal-dual hybrid gradient method, the alternating direction method of multipliers, and the extragradient method.

1 Introduction

We study convex optimization problems of the form

$$\min_{x\in\mathbf{R}^n} f(x) \qquad \text{s.t.} \qquad g_j(x) \le 0 \quad \text{for all } j \in \{1,\dots,m\}\,, \tag{1}$$

where $f\colon \mathbf{R}^n \to \mathbf{R}$ and $g_j\colon \mathbf{R}^n \to \mathbf{R}$ are convex functions. Primal-dual first-order methods, which simultaneously solve (1) and its dual, often display a marked two-stage behavior: initially, their iterates converge sublinearly towards a solution; then, after a finite number of iterations, the iterates identify the set of active constraints and their convergence switches to linear. This phenomenon is ubiquitous in practice, and it is exhibited by several first-order methods. To illustrate this phenomenon, Figure 1 shows the performance of three popular algorithms—the primal-dual hybrid gradient method (PDHG), the alternating direction method of multipliers (ADMM), and the extragradient method (EGM)—on a simple two-dimensional quadratic program with four constraints; see details in Appendix A.

There is extensive work on this phenomenon, from algorithm-specific analyses [30, 36, 37, 38, 39] to geometric accounts of the structure that drives it for broad families of algorithms [8, 23, 31, 32, 34, 56]. Although technical, the core idea is simple and directly relevant here, so we briefly recall it. It is

*Department of Applied Mathematics and Statistics and Mathematical Institute for Data Science, Johns Hopkins University, Baltimore, MD 21218, USA. MD was partially supported by NSF awards CCF 2442615 and DMS 2502377.

†Sloan School of Management, MIT, Cambridge, MA 02142, USA. HL is partially supported by AFOSR Grant No. FA9550-24-1-0051 and ONR Grant No. N000142412735.

‡Department of Statistics, University of Chicago, Chicago, IL 60637, USA. JY is partially supported by AFOSR Grant No. FA9550-24-1-0051.

convenient to reformulate (1) as an equivalent minimax problem

$$\min_{x\in\mathbf{R}^n}\max_{y\in\mathbf{R}^m}\mathcal{L}(x,y) \qquad \text{with} \qquad \mathcal{L}(x,y)=f(x)-\iota_{\mathbf{R}^m_+}(y)+\langle y,G(x)\rangle \tag{2}$$

here $\iota_{\mathbf{R}^m_+}$ denotes the indicator of the nonnegative orthant and $G\colon \mathbf{R}^n\to\mathbf{R}^m$ is the map with ith component $(G(x))_i=g_i(x)$. In turn, a pair $z^\star=(x^\star,y^\star)$ solves (2) if, and only if,

$$0\in\mathcal{F}(z^\star) \qquad \text{with} \qquad \mathcal{F}(z):=\begin{bmatrix}\partial_x\mathcal{L}(z)\\ \partial_y\left(-\mathcal{L}(z)\right)\end{bmatrix},$$

where $\partial_x\mathcal{L}$ and $\partial_y(-\mathcal{L})$ denote convex subdifferentials in x and y, respectively. Suppose we had an iterative algorithm that generates a sequence $z^k=(x^k,y^k)$ for which there exists a sequence of saddle subdifferentials $\xi^k\in\mathcal{F}(z^k)$ satisfying

$$(z^k,\xi^k)\to(z^\star,0) \qquad \text{as} \qquad k\to\infty. \tag{3}$$

Many methods satisfy this requirement, including the proximal point method (PPM), PDHG, ADMM, and EGM. Under strict complementarity, the primal-dual space decomposes locally into a manifold $\mathcal{M}$ passing through $z^\star$—encoding the active constraints—and its complement $\mathcal{M}^c$. On $\mathcal{M}$, the function $\mathcal{L}$ is smooth; off $\mathcal{M}$, the subdifferentials are bounded away from zero. Hence, small subdifferentials can arise only on $\mathcal{M}$. Consequently, if $\xi^k\to 0$, the iterates must eventually enter $\mathcal{M}$, yielding finite identification. Once on $\mathcal{M}$, the algorithm effectively solves a smooth problem on a manifold; under a local error bound, the algorithm exhibits linear convergence.

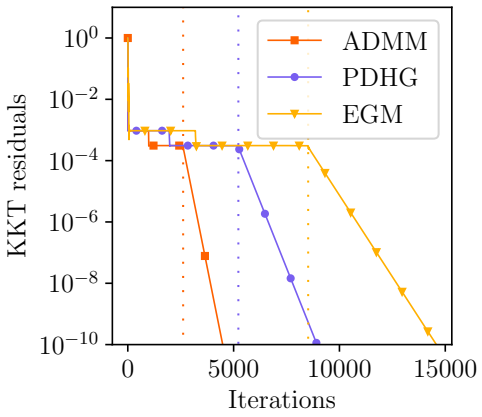


Figure 1: Distance to solution versus iteration count for several algorithms applied to a degenerate QP. The vertical dotted lines represent the last iteration at which the active set changed.

Although impressive in scope, this body of work suffers from two limitations. First, existing results rely on the strict complementarity of the limit solution, which can only be verified a posteriori. In practice, algorithms frequently converge to degenerate solutions that violate strict complementarity—as occurs for all algorithms on the problem in Figure 1. Second, most existing results guarantee only eventual active set identification without explicit finite-time bounds. These limitations motivate the central question of this work.

What problem and algorithmic properties enable nonasymptotic, finite-time identification and subsequent linear convergence without requiring strict complementarity?

To answer this question, we identify a set of mild properties that covers a broad range of algorithms and problems. Algorithmically, we require: (*i*) certain structure on the dual updates, ensuring feasibility, and (*ii*) convergence to a solution, in the sense of (3), with a sublinear decay of the saddle subgradient norm. Once again, these conditions hold for major primal-dual algorithms, such as PPM, PDHG, ADMM, and EGM. On the problem side, we require $G(\cdot)$ being Lipschitz and an error bound on the saddle subdifferential: there exists $\alpha>0$ such that

$$\alpha\,\mathrm{dist}_2(z,\mathcal{S}^\star)\le\mathrm{dist}_2(0,\mathcal{F}(z)) \qquad \text{for all } z\in\mathcal{D}\,, \tag{4}$$

where $\mathcal{S}^\star$ is the solution set and $\mathcal{D}$ is a set containing all iterates of the algorithm. This inequality is known as *metric subregularity* and has been widely studied in variational analysis [28, 53]. It

generalizes quadratic-growth conditions [15, 17] and holds broadly; for instance, it holds for linear programming (LP) problems where α can be bounded via the Hoffman constant [2, 26].

To state our nonasymptotic bounds, we need to introduce three key ingredients. The first ingredient is the correct notion of an identifiable set. When strict complementarity does not hold, we can still identify a set $\mathcal{M}$, but unlike before, this set is no longer a manifold; instead, it is a union of manifolds. In particular, assuming the algorithm converges to $z^\star$, we define the set

$$\mathcal{M} = \Big\{(x,y) \in \mathbf{R}^n \times \mathbf{R}^m_+ \mid G(x)_N < 0,\ y_N = 0,\ \text{ and }\ y_{B_a} > 0\Big\}$$

with $N := \{i \in [m] : g_i(x^\star) < 0\}$ the set of *non-active* constratints and $B_a := \{i \in [m] : g_i(x^\star) = 0 \text{ and } y_i > 0\}$ the set of *active* constraints. When strict complementary holds, $N \cup B_a = [m]$, which ensures $\mathcal{M}$ is a manifold (in fact, a subspace). When it does not hold, the set of *degenerate* constraints $B_d := [m] \setminus (N \cup B_a) = \{i \in [m] : g_i(x^\star) = 0 \text{ and } y_i = 0\}$ is non-empty and $\mathcal{M}$ can be described as a union of manifolds indexed by subsets of B_d,

$$\mathcal{M} = \bigcup_{T \subseteq B_d} \Big\{(x,y) \in \mathbf{R}^n \times \mathbf{R}^m_+ \mid G(x)_N < 0,\ y_N = 0\,, y_T = 0\,, y_{B_d \setminus T} > 0 \text{ and }\ y_{B_a} > 0\Big\}.$$

The second ingredient is the *radius of active-set stability*, defined as the smallest perturbation that flips the sign of some entry of $G(x)_N$ or y_{B_a}; formally,

$$\delta := \sup\Big\{\, t > 0 \mid G(x)_N < 0 \text{ and } y_{B_a} > 0 \text{ for all } (x,y) \in \mathbf{B}_t(z^\star)\Big\}.$$

Intuitively, δ measures how deeply $z^\star$ lies in $\mathcal{M}$; larger δ means a wider margin by which the active inequalities are satisfied. In turn, once the iterates enter $\mathbf{B}_\delta(z^\star)$, they remain there and, after a few further steps, land on $\mathcal{M}$.

We show that $\left(\max\{1, \frac{1}{\alpha}\}\frac{8\,\mathrm{dist}(z^0, \mathcal{S}^\star)}{\delta}\right)^2 + \texttt{const}$ iterations suffice to identify $\mathcal{M}$,

where α is the metric subregularity modulus in (4) and `const` is a small constant depending on algorithmic tuning. The quadratic term captures the time to reach the δ-ball; the additive constant accounts for the final steps to lock onto $\mathcal{M}$.

The third and final ingredient is a notion of 'restricted' metric subregularity as opposed to the 'global' notion. It is well known that when (4) holds, convergence of first-order algorithms gets boosted from sublinear to linear; paralleling what happens with gradient descent on strongly-convex smooth function. However, the rate depends on the modulus α, and when α is small, the linear rate can be impractically slow. Crucially, once identification kicks in, it suffices to enforce metric subregularity in a neighborhood of the identifiable set: replace $\mathcal{D}$ in (4) by $\mathcal{M}$ intersected with a δ-ball around the solution. Let $\alpha_{\mathcal{M}}$ denote the corresponding restricted modulus; this constant governs the post-identification linear rate.

After $\mathcal{M}$ is identified, $\mathcal{O}\left(\frac{1}{\alpha_{\mathcal{M}}^2}\log\left(\frac{1}{\alpha_{\mathcal{M}}\varepsilon}\right)\right)$ iterations suffice to find a z with $\mathrm{dist}(z, \mathcal{S}^\star) \le \varepsilon$.

Importantly, $\alpha_{\mathcal{M}}$ can be orders of magnitude larger than the global metric subregularity modulus and this convergence rate holds even when $\mathcal{M}$ fails to be a manifold, thereby explaining the second-phase speed-up and yielding nonasymptotic guarantees without requiring strict complementarity.

Outline. The rest of this section is devoted to related work. Section 2 briefly summarizes the necessary background. In Section 3, we describe the problem and algorithmic classes that we consider, and in Section 3.3, we verify that several popular algorithms fall within the algorithmic class we study. Section 4 presents our general guarantees. Section 5 closes the paper with numerical experiments supporting our theory. Lengthy, technical proofs are deferred to the appendix.

Related literature

Our work is closely related to that of the last two authors [43], who studied the central question of this work for the particular case of PDHG applied to linear programs (LPs). Our answer builds on their ideas but requires substantial changes. Two main obstacles to this generalization are: (*i*) the LP analysis is based on polyhedral geometry, which is unavailable for general convex problems with functional constraints; and (*ii*) [43] leverages the explicit PDHG updates, whereas here we distill the basic algorithmic conditions that still ensure the two-stage behavior. The second obstacle necessitates more delicate arguments to treat algorithms such as ADMM, which lack several of the structural properties enjoyed by PDHG.

Convex-concave primal-dual algorithms. Convex-concave saddle-point problems and their associated primal-dual algorithms have been extensively studied for decades. The saddle-point problems are also studied as instances of general variational inequalities. In 1976, Rockafellar's seminal work introduced the proximal point method (PPM) [51] for solving monotone variational inequalities, while Korpelevich proposed the extragradient method (EGM) for solving convex–concave saddle-point problems [29]. In [45], Nemirovski showed that EGM, as a mirror-prox instance, can be viewed as an approximation of PPM. For saddle-point problems with bilinear interaction terms, algorithmic development has been especially rich: the primal–dual hybrid gradient (PDHG) method [10, 11, 62] and the alternating direction method of multipliers (ADMM) [7, 21] are widely used in practice. More recently, it has been established that PDHG and ADMM can also be interpreted as approximations of PPM [24, 42].

Finite time identification. Finite-time identification of active sets refers to an algorithm's capability to identify the underlying manifold or active constraints in finite iterations [8, 22, 56]. This behavior is often analyzed under the framework of partial smoothness [22, 33, 34] and through the closely related $\mathcal{VU}$-decomposition perspective developed by [31, 44]. Roughly speaking, a function is partly smooth relative to a manifold if it is smooth along the manifold while being sharply nonsmooth in directions transverse to it. This structure, combined with strict complementarity, enables characterizations of identification and the subsequent fast local convergence of algorithms. For example, manifold identification for dual averaging has been established in [30], while forward–backward splitting methods have been shown to achieve finite-time identification under similar assumptions [35, 37]. In the context of primal-dual methods, finite-time active-set identification and local convergence of PDHG and ADMM are analyzed in [2, 38, 39, 59]. Recent work [3] showed that finite time identification and fast convergence also occur for infeasible linear programming problems, albeit with respect to an auxiliary feasible problem that characterizes the direction in which the iterates diverge. However, these guarantees hinge on nondegeneracy of the limiting solution—an assumption that is difficult to certify a priori and is frequently violated in applications [18, 19]. To our knowledge, there are two notable exceptions. First, the line of work initiated by [46, 57, 58], which develops modified sequential quadratic programming schemes designed to recover two-stage behavior even in the presence of degeneracy. In contrast, our analysis applies to standard first-order methods without any degeneracy-handling modifications. Second, the work [18]

develops a sensitivity and identification theory for (degenerate) mirror-stratifiable convex functions. Our work, instead, is not concerned with sensitivity and does not rely on mirror-stratifiability.

Metric subregularity and growth conditions. Metric subregularity imposes a linear error bound on the saddle subdifferential [13, 14, 27], serving as a unifying regularity condition across diverse problem classes. Originally introduced in the early works of Robinson [50], metric subregularity is closely related to notions such as calmness and error bounds [16, 25]. Many structured problems, including those involving piecewise linear–quadratic models such as Lasso and support vector machines, naturally satisfy subregularity on compact domains [50, 61]. Motivated by these applications, recent research has studied how first-order methods behave under this assumption. In convex minimization, metric subregularity of the subdifferential has been shown to be equivalent to quadratic growth conditions, leading to linear convergence [16]. Similar developments extend to saddle-point and primal–dual settings, such as for PDHG and ADMM [41, 60].

2 Preliminaries

In this section, we review the notation and necessary background in linear algebra and convex analysis. We defer the interested reader to the monographs [6, 52]. We use the symbols $\mathbf{N}$ and $\mathbf{R}$ to denote the set of natural (without zero) and real numbers, respectively. Further, we use $\overline{\mathbf{R}}$ to denote $\mathbf{R} \cup \{+\infty\}$. We denote the set of nonnegative reals as $\mathbf{R}_+$. We endow $\mathbf{R}^n$ with the standard dot product $\langle x, y\rangle = x^\top y$ and its induced norm $\|x\|_2 = \sqrt{\langle x, x\rangle}$. The symbols $\mathbf{B}_t(z)$ and $\overline{\mathbf{B}}_t(z)$ denote the open and closed ball of radius t centered at z, respectively, and we label $\mathbf{B} := \mathbf{B}_1(0)$ and $\overline{\mathbf{B}} := \overline{\mathbf{B}}_1(0)$. We use $\mathcal{S}^n$ to denote the set of $n \times n$ symmetric matrices. For a given $M \in \mathcal{S}^n$ we denote its eigenvalues by $\lambda_1(M) \geq \cdots \geq \lambda_n(M)$. We write $\lambda_{\max}(M)$ for the largest eigenvalue of M and $\lambda^+_{\min}(M)$ for the smallest nonzero eigenvalue of M. We use $\mathcal{S}^n_+ = \{M \in \mathcal{S}^n : \lambda_n(M) \geq 0\}$ and $\mathcal{S}^n_{++} = \{M \in \mathcal{S}^n : \lambda_n(M) > 0\}$ to denote positive semidefinite (PSD) and positive definite (PD) matrices, respectively. The symbol $\kappa(M) = \lambda_1(M)/\lambda_n(M)$ denotes the condition number of M. The symbol $M^\dagger$ denotes the pseudoinverse of M. Any $M \in \mathcal{S}^n_+$ induces a semi-inner product $\langle x, y\rangle_M = x^\top M y$ and a semi-norm $\|x\|_M = \sqrt{\langle x, x\rangle_M}$ that induces a pseudodistance. The symbol $\overline{\mathbf{B}}^M_t(z)$ denotes the set of points whose M-pseudodistance to z is less than or equal to t. Similarly, the symbol $\mathbf{B}^M_t(z)$ denotes the set of points whose M-pseudodistance to z is strictly less than t. Abusing notation, we define the M-pseudo-distance from a point x to a set $Q \subseteq \mathbf{R}^n$ via

$$\operatorname{dist}_M(x, Q) := \inf_{y \in Q} \|x - y\|_M.$$

When M corresponds to the identity, we drop the subindex and simply write dist. We define the Euclidean projection onto a closed set Q as

$$\operatorname{proj}(x) := \operatorname*{argmin}_{y \in Q} \|x - y\|_2.$$

Consider a function $f\colon \mathbf{R}^n \to \overline{\mathbf{R}}$. We let $\operatorname{epi}(f) = \{(x, t) \in \mathbf{R}^{d+1} : f(x) \leq t\}$ be the epigraph of f. The function is proper if the epigraph is nonempty. Analogously, we say that f is closed (resp. convex) if its epigraph is closed (resp. convex). Given a convex, closed set $Q \subseteq \mathbf{R}^n$, we define its indicator function as

$$\iota_Q(x) := \begin{cases} 0 & \text{if } x \in Q, \\ +\infty & \text{otherwise.} \end{cases}$$

For a closed, convex, proper function $f\colon \mathbf{R}^d \to \overline{\mathbf{R}}$ and a point $x \in \mathbf{R}^n$, the *convex subdifferential* of

f at x, denoted by $\partial f(x)$, corresponds to the set of vectors g satisfying

$$f(z) \geq f(x) + \langle g, z - x \rangle \qquad \text{for all } z \in \mathbf{R}^m.$$

Similarly, for a function $\mathcal{L}\colon \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ which is convex in its first component and concave in its second component, the *saddle subdifferential* at a point $(x, y) \in \mathbf{R}^n \times \mathbf{R}^m$ is given by

$$\mathcal{F}(x, y) = \begin{bmatrix} \partial_x \mathcal{L}(x, y) \\ \partial_y(-\mathcal{L}(x, y)) \end{bmatrix}, \tag{5}$$

where, for any fixed $y \in \mathbf{R}^m$ we use $\partial_x \mathcal{L}(x, y)$ to denote the subdifferential of the function $\mathcal{L}(\cdot, y)$ at x and an analogous definition follows for ∂_y.

3 Setting and algorithms

In this section, we formalize the setting we study, state assumptions, and present a general algorithmic template. Rather than focusing on a single method, we analyze a meta-algorithm satisfying mild conditions. At the end of this section, we show that many popular methods fit this template.

3.1 Problem class

Our departing point is a pair of primal-dual problems of the form

$$p^\star = \begin{cases} \min\limits_{x \in \mathbf{R}^n} & f(x) \\ \text{s.t.} & g_j(x) \leq 0 \quad \text{for all } j \in \{1, \dots, m\}, \end{cases} \qquad \text{and} \qquad q^\star = \begin{cases} \max\limits_{y \in \mathbf{R}^m} & h(y) \\ \text{s.t.} & y \geq 0, \end{cases} \tag{6}$$

where the functions f and g_j are assumed to be convex, and $h(y) = \min_{x \in \mathbf{R}^n} f(x) + \langle y, G(x) \rangle$ with G the map whose entries are given by $G(x)_i = g_i(x)$. This formulation subsumes linear and quadratic programming and extends well beyond these classes. Primal-dual optimal solutions $\mathcal{S}^\star$ are given by the set of pairs $(x, y) \in \mathbf{R}^{n+m}$ satisfying

$$\begin{aligned} f(x) - h(y) &\leq 0 \quad \text{(Zero duality gap)} \\ G(x) &\leq 0 \quad \text{(Primal feasibility)} \\ y &\geq 0 \quad \text{(Dual feasibility)}. \end{aligned} \tag{7}$$

As mentioned in the introduction (6) is equivalent to the minimax problem (2). For any $z \in \mathbf{R}^{n+m}$ let $\mathcal{F}(z) \subseteq \mathbf{R}^{n+m}$ denote the saddle subdifferential (5) of the constrained Lagrangian function $\mathcal{L}(x, y) = f(x) - \iota_{\mathbf{R}^m_+}(y) + \langle y, G(x) \rangle$. By the saddle point theorem for convex optimization [52], when the functions f and g_j are convex we have $\mathcal{F}^{-1}(0) = \mathcal{S}^\star$.

We will impose a standard set of assumptions on the primal-dual problem.

Assumption 1. Problem (2) satisfies the following two conditions.

1. **(Convexity)** The functions f and g_j are convex for all $j \in [m]$.
2. **(Existence of solutions)** The set of primal-dual solutions $\mathcal{S}^\star$ is nonempty.

These conditions are standard. The existence of primal-dual solutions is equivalent to strong duality $p^\star = d^\star$ with primal-dual attainment, which is implied by constraint qualification conditions such as the Slater condition.

Assumption 2 (Metric sub-regularity). For any radius $t > 0$ there exists $\alpha > 0$ such that

$$\alpha \operatorname{dist}_2(z, \mathcal{S}^\star) \leq \operatorname{dist}_2(0, \mathcal{F}(z)) \qquad \text{for all } z \in \mathcal{S}^\star + t\mathbf{B}.$$

The assumption asserts an error bound for the saddle subdifferential—known as metric subregularity in the variational analysis literature [28, 53]—that is intimately related to quadratic-growth behavior and the stability of solutions [15, 17]. The assumption holds for a broad class of primal-dual problems; in particular, all LP problems satisfy it [2, 26].

3.2 Algorithmic template

In this section, we introduce the meta-algorithm we analyze. To solve (6), the meta-algorithm maintains two sequences: the main iterates $z^k = (x^k, y^k)$ and the auxiliary iterates $\tilde{z}^k = (\tilde{x}^k, \tilde{y}^k)$. That is, the method is initialized at some z^0 and updates

$$z^{k+1}, \tilde{z}^{k+1} \leftarrow \texttt{PrimalDualStep}(z^k). \tag{8}$$

In a nutshell, $\tilde{z}^k$ is an intermediate update that we will use to describe our assumptions; yet for several algorithms it is trivially equal to z^k. We also suppose that each algorithm comes equipped with a PD matrix $P \in \mathcal{S}^{n+m}_{++}$ and its associated norm $\|z\|_P := \sqrt{z^\top P z}$, which dictates a natural geometry to measure progress of the algorithm.

Next, we introduce three assumptions on the meta-algorithm (8). Assumption 3 is necessary for eventual identification; Assumption 4 is required for local linear convergence; and Assumption 5, in tandem with the first two, delivers nonasymptotic identification. We shall see in Section 3.3 that all of these assumptions hold for several popular algorithms. For intuition, the reader might take $P = I$, in which case $\|\cdot\|_P = \|\cdot\|_2$; this choice is realized by a number of methods.

Assumption 3. There exists a positive definite matrix $P \in \mathcal{S}^n_{++}$ such that the following three hold.

(i) (**Convergence**) For any initial iterate $z^0 \in \mathbf{R}^{n+m}$ there exists $z^\star \in \mathcal{S}^\star$ such that

$$\max\left\{\|z^k - z^\star\|_P, \|\tilde{z}^k - z^\star\|_P\right\} \to 0 \quad \text{as} \quad k \to \infty\,.$$

(ii) (**Dual update**) There exists $\eta > 0$ (stepsize) such that dual update takes the form

$$y^{k+1} = \mathrm{proj}_{\mathbf{R}^m_+}(y^k + \eta G(\tilde{x}^{k+1}))\,.$$

(iii) (P-**Lipschitzness**) For each $t > 0$ and $j \in [m]$ there exist $L^x_{tj}, L^y_{tj} > 0$ such that

$$g_j(x) - g_j(x') \le L^x_{tj}\|z - z'\|_P \qquad \text{and} \qquad |y_j - y'_j| \le L^y_{tj}\|z - z'\|_P$$

for $z = (x, y), z' = (x', y') \in \mathbf{B}^P_t(z^\star)$.

Let us comment on these conditions. The first condition essentially states that the algorithm converges to an optimal solution point-wisely. The second condition is ubiquitous in primal-dual algorithms tackling (6) as we shall see in the following section. The third condition, although not standard, holds automatically provided the functions g_j are Lipschitz continuous since in Euclidean spaces all norms are equivalent. We state it in terms of the P-norm since it is a natural norm to state our guarantees. In the appendix, we state a weaker, more technical assumption that only requires P to be positive semidefinite, which is crucial to cover ADMM, whose associated P is singular. In what follows, P denotes the same matrix as in Assumption 3. We use the range of P in the statement of the next assumption, which under Assumption 3 is trivially the whole space (since P is positive definite); we keep this form to remain compatible with the semidefinite (singular) case handled in the appendix.

Assumption 4. The following two hold.

(*i*) (**Subdifferential sublinear rate**) There exists $\gamma > 0$ such that

$$\operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z^k) \cap \operatorname{range}(P)) \leq \frac{\gamma \operatorname{dist}_P(z^0, \mathcal{S}^\star)}{\sqrt{k}} \qquad \text{for any } k \in \mathbf{N}\,.$$

(*ii*) (**Closedness to solutions**) There exists $\tau > 0$ such that $\operatorname{dist}_2(z^k, \mathcal{S}^\star) \leq \tau$ for all $k \in \mathbf{N}$.

The first assumption states essentially that the subgradient $\mathcal{F}(z^k)$ converges to 0 at a sublinear rate, and the second assumption states the iterates stay in a bounded region from the optimal solution set. These conditions are common among first-order methods. In turn, a simple consequence of this assumption and metric subregularity is linear convergence. The next result formalizes this statement. This is a well-known guarantee, we include it for completeness; its proof appears in Appendix B.1.

Proposition 3.1. *Suppose Assumptions 1, 2, and 4 hold. Let z^k be the kth iterate generated by update (8). Then,*

$$0 < \alpha_G := \inf_{z \in \mathcal{D}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} \qquad \textit{with} \qquad \mathcal{D} = \mathcal{S}^\star + \tau \mathbf{B}\,. \tag{9}$$

Further, for any $k \in \mathbf{N}$ we have

$$\operatorname{dist}_2(z^k, \mathcal{S}^\star) \leq \sqrt{e}\nu \exp\left(-\frac{1}{2}\frac{k}{\lceil e\nu^2 \rceil}\right) \operatorname{dist}_2(z^0, \mathcal{S}^\star) \quad \textit{where} \quad \nu = \frac{\gamma\, \lambda_{\max}(P)}{\alpha_G}.$$

The rate depends on the global metric subregularity modulus in (9). In practice, this constant is often small, especially for badly conditioned problems, yielding impractically slow convergence. Thus, it is common to observe an first stage of active set identification; the next assumption is key in deriving bounds on this stage.

Assumption 5. The following two conditions hold.

(*i*) (**Sublinear rate**) There exists $\gamma > 0$ such that the following inequality holds:

$$\max\left\{\|z^{k+1} - z^k\|_P,\, \|z^k - \tilde{z}^k\|_P\right\} \leq \frac{\gamma \operatorname{dist}_P(z^0, \mathcal{S}^\star)}{\sqrt{k}} \qquad \text{for all} \quad k \in \mathbf{N}_0\,.$$

(*ii*) (**Star non-expansiveness**) For any $\underline{z}^\star \in \mathcal{S}^\star$ the following inequality holds

$$\|z^\star - \underline{z}^\star\|_P \leq \|z^k - \underline{z}^\star\|_P \qquad \text{for all} \quad k \in \mathbf{N}_0\,.$$

The next section shows that all these requirements are satisfied by several algorithms.

3.3 Instantiations of the meta-algorithm

In this section, we show that four classic algorithms for solving minimax problems satisfy the assumptions defining the meta-algorithm described in Section 3.2. The proofs of all results in this section are deferred to Appendix B.2.

Proximal Point Method (PPM). Fix $\eta > 0$, PPM [51] solves (2) by iteratively updating

$$(x^{k+1}, y^{k+1}) \leftarrow \arg\min_{x \in \mathbf{R}^n} \max_{y \in \mathbf{R}^m_+} f(x) + \langle y, G(x)\rangle + \frac{1}{2\eta}\|x - x^k\|_2^2 - \frac{1}{2\eta}\|y - y^k\|_2^2\,. \tag{10}$$

For this method, we trivially take $\tilde{z}_k = z_k$. We highlight that solving (10) might be just as hard as solving the original problem (6), and so, in most situations, this is not a practical algorithm. However, it serves as a clean canonical baseline for our framework. The next result shows that PPM satisfies all our assumptions; we defer its proof to Appendix B.2.1.

Proposition 3.2. *Fix any stepsize $\eta > 0$ and set the auxiliary iterates to $\tilde{z}^k = z^k$. Then, PPM satisfies Assumptions 3, 4, and 5 with $P = \eta^{-1} I$ and $\gamma = 1$.*

Primal Dual Hybrid Gradient (PDHG). Assume the constraints in (6) are affine, namely $G(x) = Ax - b$ for $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$. Fix $\eta > 0$, PDHG [10] solves the corresponding minimax problems (2) by iteratively updating

$$\begin{aligned} x^{k+1} &\leftarrow \mathrm{prox}_{\eta f}(x^k - \eta A^\top y^k) \\ \tilde{x}^{k+1} &\leftarrow 2x^{k+1} - x^k \\ y^{k+1} &\leftarrow \mathrm{proj}_{\mathbf{R}^m_+}(y^k + \eta A \tilde{x}^{k+1}) \ . \end{aligned}$$

PDHG is used extensively for inverse problems arising in imaging [5, 20] and large-scale linear programming [1, 4]. The next proposition shows that PDHG satisfies all our assumptions; the proof is deferred to Appendix B.2.2.

Proposition 3.3. *Fix a stepsize $\eta > 0$ satisfying $\eta < \|A\|_{\mathrm{op}}^{-1}$ and set the dual auxiliary iterates to be $\tilde{y}^k = y^k$. Then, PDHG satisfies Assumptions 3, 4, and 5 with $P = \begin{bmatrix} \frac{1}{\eta} I & -A^\top \\ -A & \frac{1}{\eta} I \end{bmatrix}$ and $\gamma = 1$.*

Alternating Direction Method of Multipliers (ADMM). As with PDHG, suppose that the constraints are affine $G(x) = Ax - b$, $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$. Fix $\eta > 0$, ADMM [7, 21] solves (2) by iteratively updating

$$\begin{aligned} u^{k+1} &\leftarrow \operatorname*{argmin}_{u \in \mathbf{R}^m} \left(\iota_{\mathbf{R}^m_+}(u) + \langle y^k, u \rangle + \frac{\eta}{2} \|Ax^k + u - b\|^2) \right) \\ y^{k+1} &\leftarrow y^k + \eta(Ax^k - b + u^{k+1}) \\ x^{k+1} &\leftarrow \operatorname*{argmin}_{x \in \mathbf{R}^n} \left(f(x) + \langle y^{k+1}, Ax \rangle + \frac{\eta}{2} \|Ax + u^{k+1} - b\|^2 \right) \ . \end{aligned}$$

ADMM underpins widely used quadratic and convex programming solvers such as OSQP and SCS [47, 55]. The analysis of ADMM is more subtle. Indeed, unlike the other algorithms we consider, ADMM does not satisfy Assumption 3 with a strictly positive definite P. Nevertheless, it does satisfy a weaker version of it with P positive semidefinite, namely Assumption 3*. As alluded to before, this version is more technical, and so we deferred it to the appendix. Nonetheless, all the convergence guarantees we establish hold with either assumption. The next proposition shows that ADMM satisfies this slightly modified set of assumptions; the proof is deferred to Appendix B.2.3.

Proposition 3.4. *Fix any stepsize $\eta > 0$ and set the iterates of ADMM to be $z^k = \tilde{z}^k = (x^k, y^k)$. Assume there exists $\tau > 0$ such that $\mathrm{dist}_2(z^k, \mathcal{S}^\star) \leq \tau$ for all $k \in \mathbf{N}$. Then, ADMM satisfies Assumptions 3*, 4, and 5 with $P = \begin{bmatrix} \eta A^\top A & A^\top \\ A & \frac{1}{\eta} I \end{bmatrix}$ and $\gamma = 1$.*

This proposition supposes that Assumption 4 (*ii*) holds true. This is the case, for instance, when A has full rank or when the solution set $\mathcal{S}^\star$ is bounded [7].

Extragradient Method (EGM). Assume the Lagrangian function $\bar{\mathcal{L}}(x, y) = f(x) + \langle y, G(x) \rangle$, with f and G as in (1), is L-smooth, i.e., differentiable with L-Lipschitz gradients. Fix $\eta > 0$,

EGM [29] solves (2) by iteratively updating

$$\begin{aligned}
\tilde{x}^{k+1} &\leftarrow x^k - \eta(\nabla f(x^k) + J_G(x^k)^\top y^k) \\
\tilde{y}^{k+1} &\leftarrow \mathrm{proj}_{\mathbf{R}^m_+}(y^k + \eta G(x^k)) \\
x^{k+1} &\leftarrow x^k - \eta(\nabla f(\tilde{x}^{k+1}) + J_G(\tilde{x}^{k+1})^\top \tilde{y}^{k+1}) \\
y^{k+1} &\leftarrow \mathrm{proj}_{\mathbf{R}^m_+}(y^k + \eta G(\tilde{x}^{k+1}))\ ,
\end{aligned}$$

EGM adds an extrapolation step to the vanilla gradient descent-ascent method to ensure convergence. In turn, it can be EGM interpreted as an approximation of PPM [45]. EGM is used across a broad range of modern minimax and saddle-point applications, including games, machine learning, and imaging [12, 40, 49]. The next proposition shows that EGM satisfies all our assumption; the proof is deferred to Appendix B.2.4.

Proposition 3.5. *Fix $\eta < 1/L$. Then, EGM satisfies Assumptions 3, 4, and 5 with $P = I$ and $\gamma = \frac{3}{\sqrt{1-(\eta L)^2}}$.*

4 Guarantees

In this section, we present our main theoretical results. Section 4.1 provides our identification guarantees. Section 4.2 shows that the local geometry of the problem around the limit solution is better conditioned than the full problem, which yield faster linear convergence. Section 4.3 compares the metric subregularity moduli associated to the global and the local problems.

4.1 Finite time identification

Next, we state our finite-time identification results. We express these results in terms of the P-norm $\|z\|_P = \sqrt{z^\top P z}$, which differs slightly from our narrative in the introduction where for simplicity we used the Euclidean norm, i.e., $P = I$. To start, recall that the set we identify depends on the active set at the solution we converge to $z^\star$. In particular, it depends on the index sets

$$\begin{aligned}
N &= \{j \in [m] : g_j(x^\star) < 0\}, \\
B_a &= \{j \in [m] : g_j(x^\star) = 0,\ y^\star_j > 0\},\ \text{and} \\
B_d &= \{j \in [m] : g_j(x^\star) = 0,\ y^\star_j = 0\}\,.
\end{aligned} \tag{11}$$

We call N the set of *nonactive* indices, B_a the set of *active* indices and B_d the set of *degenerate* indices.[1] Further, we use the placeholder $B = B_a \cup B_d$. We say the solution $z^\star$ is *degenerate* if it does not satisfy strict complementarity, that is, if $B_d \neq \emptyset$.
With this index partition, we define the identifiable set as

$$\mathcal{M} = \Big\{(x, y) \in \mathbf{R}^n \times \mathbf{R}^m_+ \mid G(x)_N < 0,\ y_N = 0,\ \text{and}\ y_{B_a} > 0\Big\}.$$

In light of Assumption 3, this is the effective domain of our algorithms. The weakly active indices do not otherwise enter the definition; for $(x, y) \in \mathcal{M}$ they are only required to satisfy $y_{B_d} \geq 0$. Consequently, if $B_d \neq \emptyset$, the set $\mathcal{M}$ need not be a manifold—it has a border along $y_i = 0 : i \in B_d$. We define the *radius of active-set stability* as the size of the smallest perturbation of $z^\star$, with respect to the P-seminorm, that violates one constraints $G_N(x) < 0$ and $y_{B_a} > 0$; formally

$$\delta = \sup\Big\{t \in (0, \infty) \mid \text{For all } (x, y) \in \mathbf{B}^P_t(z^\star) \text{ we have } G_N(x) < 0,\ \text{and}\ y_{B_a} > 0\Big\}\,, \tag{12}$$

[1] The indices in B_a and B_d also go under the names of strongly and weakly active in the literature [46].

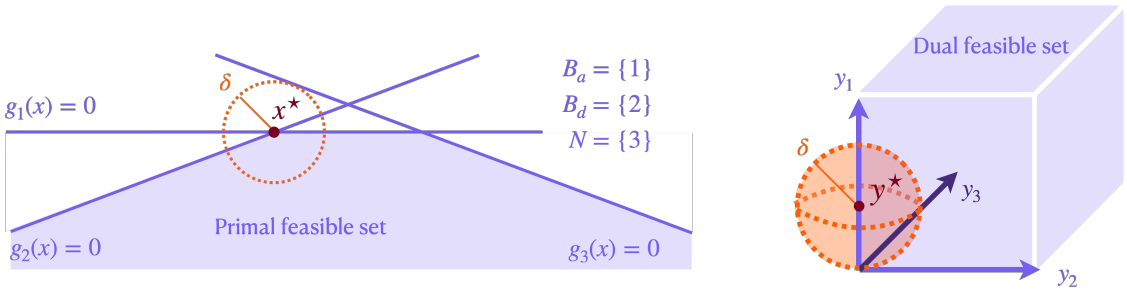


Figure 2: Illustration of the radius of active-set stability (12).

The radius of active-set stability quantifies how internal the point $z^\star$ is with respect to the set $\mathcal{M}$, modulo the requirement $y_N = 0$. Indeed, we have $\mathbf{B}_\delta^P(z^\star) \cap \mathcal{Z}_0 \subseteq \mathcal{M}$ (Proposition C.6 in Appendix C.2), where

$$\mathcal{Z}_0 = \{(x, y) \in \mathbf{R}^n \times \mathbf{R}^m_+ : y_N = 0\}. \tag{13}$$

The following result shows that converging primal-dual algorithms with a projected gradient ascent dual update eventually identify this intersection. We defer its proof to Appendix C.2.

Theorem 4.1. *Suppose Assumptions 1 and 3 hold. Let z^k be the kth iterate generated by update (8). Then, there exists $K \in \mathbf{N}$ such that $z^k \in \mathbf{B}_\delta^P(z^\star) \cap \mathcal{Z}_0 \subseteq \mathcal{M}$ for all $k \geq K$.*

We note that we only require two basic assumptions for this asymptotic result. Variations of this result have appeared in the literature before [8, 23, 56], typically without an explicit quantification of the identification neighborhood. Making this radius explicit is the starting point for nonasymptotic rates. Combining our characterization of δ with metric subregularity yields explicit rates of convergence towards the ball $\mathbf{B}_{\delta/2}^P(z^\star)$. Once inside the neighborhood, the iterates reach $\mathcal{Z}_0$ after a small number of iterations, depending only on the stepsize and Lipschitz modulus.

Theorem 4.2. *Suppose Assumptions 1, 3, 4, and 5 hold and that the k-th iterate z^k and the k-th intermediate iterate $\tilde{z}^k$ of the meta-algorithm defined in (8) are equal. Then, $z^k \in \mathbf{B}_{\delta/2}^P(z^\star) \cap \mathcal{Z}_0 \subseteq \mathcal{M}$ provided*

$$k > K := \left\lceil e \left(\frac{\gamma\, \lambda_{\max}(P)}{\alpha_G} \right)^2 \right\rceil \left[1 + 2 \ln \left(\frac{4\, \lambda_{\max}(P)^{\frac{3}{2}}\, \mathrm{dist}_2(z^0, \mathcal{S}^\star)}{\alpha_G \delta} \right) \right] + \left\lceil \max_{j \in N} \frac{L^y_{\delta j}}{\eta \overline{L^x_{\delta j}}} \right\rceil .$$

The constant $\overline{L^x_{\delta j}}$ in the definition of K is the smallest Lipschitz constant of g_j on the ball $\mathbf{B}_\delta^P(z^\star)$.[2] The constant α_G corresponds with the 'global' metric subregularity modulus in (9). The proof parallels that of Theorem 3.1 and is deferred to Appendix C.3. The first summand in the definition of K corresponds to the time to reach the δ ball around $z^\star$, the second summand bounds the number of additional iterations required to reach $\mathcal{Z}_0$. The latter depends on the choice of stepsize tuning.

As noted in the paragraph after (9), for badly conditioned problems the constant α_G can be small—we will see an explicit example in Section 4.3— so the resulting bound may be overly conservative. To derive a more meaningful bound, we use a completely different approach that does not depend on the global constant α_G, but rather on a *local* metric subregularity modulus given by

$$\alpha_L = \inf_{z \in \mathcal{D}} \frac{\mathrm{dist}_2(0, \mathcal{F}(z))}{\mathrm{dist}_2(z, \mathcal{S}_L^\star)}, \tag{14}$$

[2] Formally, $\overline{L^x_{\delta j}} := \sup_{z \in \mathbf{B}_\delta^P(z^\star) \setminus z^\star} \frac{g_j(x) - g_j(x^\star)}{\|z - z^\star\|_P}$. If g_j is constant in $\mathbf{B}_\delta^P(z^\star)$, then $\overline{L^x_{\delta j}} = 0$, whence the bound in Theorem 4.2 becomes unrealizable. In that case, as generalized in Theorem C.8, $\overline{L^x_{\delta j}}$ can be replaced by $-2g_j(x^\star)/\delta$.

where $\mathcal{S}_L^\star$ is defined as the set of primal-dual solutions to the following reduced system of equations

$$f(x) - h(y) \leq 0, \quad G(x)_B \leq 0, \quad \text{and} \quad y \geq 0\,. \tag{15}$$

The set of solutions $\mathcal{S}^\star$ of the original problem solves a bigger system of equations (7). Hence, $\mathcal{S}^\star \subseteq \mathcal{S}_L^\star$ and comparing the two definitions, we derive $\alpha_L \geq \alpha_G$. With it we obtain the following.

Theorem 4.3: Finite time identification

Suppose Assumptions 1, 3, 4, and 5 hold. Let z^k be the kth iterate generated by update (8). Then, $z^k \in \mathbf{B}^P_{\delta/2}(z^\star) \cap \mathcal{Z}_0 \subseteq \mathcal{M}$ for all

$$k > K := \left(\max\left\{1, \frac{1}{\alpha_L}\right\} \frac{8\,\lambda_{\max}(P)^{\frac{3}{2}}\,\mathrm{dist}_2(z^0, \mathcal{S}^\star)}{\delta} \right)^2 + \left\lceil \max_{j \in N} \frac{L^y_{\delta j}}{\eta \overline{L}^x_{\delta j}} \right\rceil .$$

A couple of remarks are in order. At first sight, it might appear that the bound in Theorem 4.2 gives a fast linear convergence in terms of δ, while Theorem 4.3 only yields sublinear convergence. However, the linear convergence rate relies on the conservative global constant α_G. In contrast, the sublinear convergence is dependent on the local metric subregularity modulus α_L, and can be more informative than the global constant α_G. Secondly, we would like to comment that the additive term $\left\lceil \max_{j\in N} \frac{L^y_{\delta j}}{\eta \overline{L}^x_{\delta j}} \right\rceil$ can be viewed as a rather "small" constant term that does not affect much of the order of rate. The goal for this term is to guarantee that the iterates identify the set $\mathcal{M}$ after reaching the ball $\mathbf{B}^P_{\delta/2}(z^\star)$.

4.2 Local rapid convergence

So far we have established that the meta algorithm (8) identifies the union of manifolds $\mathcal{M}$ after enough iterations. After which, the algorithm effectively solves the primal dual problem restricted to $\mathcal{M}$. This restricted problem is better conditioned than the global problem, as it eliminates the nonactive constraints at the limit solution. In turn, this speeds up the convergence from sublinear to linear. We quantify this phenomenon via a local metric subregularity modulus

$$\alpha_{\mathcal{M}} = \inf_{z \in \mathcal{D}_\delta \cap \mathcal{M}} \frac{\mathrm{dist}_2(0, \mathcal{F}(z))}{\mathrm{dist}_2(z, \mathcal{S}^\star)} \quad \text{where} \quad \mathcal{D}_\delta = \mathcal{D} \cap \mathbf{B}^P_{\delta/2}(z^\star)\,. \tag{16}$$

The following proposition demonstrates the faster local convergence after identification. More formally, it states that, suppose after K iterations, all iterates z^k for $k > K$ stay in the union of manifold M and they are not too far away from z^*, then the iterates enjoy a faster local linear convergence rate to the optimal solution set that only relies on the local sharpness constant $\alpha_{\mathcal{M}}$ instead of the global sharpness constant α_G.

Proposition 4.4. *Suppose Assumptions 1, 2 and 4. Let z^k be the kth iterate generated by the update (8). Further suppose that there exists $K > 0$ such that for all $k > K$ we have $z^k \in \mathcal{D}_\delta \cap \mathcal{M}$. Then,*

$$\mathrm{dist}_2(z^{k+K}, \mathcal{S}^\star) \leq \sqrt{e}\nu_{\mathcal{M}} \exp\left(-\frac{1}{2}\frac{k}{\lceil e\nu_{\mathcal{M}}^2 \rceil}\right) \mathrm{dist}_2(z^K, \mathcal{S}^\star) \quad \textit{with} \quad \nu_{\mathcal{M}} = \frac{\gamma\,\lambda_{\max}(P)}{\alpha_{\mathcal{M}}}\,.$$

Proposition 4.4 follows with the same proof as that of Proposition 3.1 by replacing ν with $\nu_{\mathcal{M}}$, and is therefore omitted.

Putting together the finite time identification (Theorem 4.3) and the faster linear convergence after identification (Proposition 4.4), we derive the following result, which presents the full charac-

terization of the two-stage convergence behavior of primal-dual algorithms. Its proof is deferred to Appendix C.4.

Theorem 4.5: Two-stage convergence rates

Suppose Assumptions 1, 2, 3, 4, 5 hold and fix $\varepsilon > 0$. Let z^k be the kth iterate generated by update (8). Then, we have $\operatorname{dist}_2(z^k, \mathcal{S}^\star) \leq \varepsilon$ provided that

$$k > \underbrace{\left\lceil \max\left\{1, \frac{1}{\alpha_L}\right\} \frac{8\,\lambda_{\max}(P)^{\frac{3}{2}}\operatorname{dist}_2(z^0, \mathcal{S}^\star)}{\delta} \right\rceil^2 + \left\lceil \max_{j \in N} \frac{L^y_{\delta j}}{\eta \overline{L}^x_{\delta j}} \right\rceil}_{\text{Finite time identification}} + \underbrace{\rho_{\mathcal{M}} \left\lceil 1 + 2\ln\left\lceil \frac{\gamma\,\lambda_{\max}(P)^{\frac{1}{2}}\delta}{2\alpha_{\mathcal{M}}\varepsilon} \right\rceil \right\rceil}_{\text{Fast linear convergence}},$$

with $\rho_{\mathcal{M}} = \lceil e\gamma^2\,\lambda_{\max}(P)^2/\alpha^2_{\mathcal{M}} \rceil$.

Thus, he overall complexity is $O\left(\frac{1}{\alpha_L^2\delta^2} + \frac{1}{\alpha^2_{\mathcal{M}}}\ln\left(\frac{\delta}{\alpha_{\mathcal{M}}}\right)\right)$. A few observations in order. First, both α_L and $\alpha_{\mathcal{M}}$ are local metric subregularity constants, thereby avoiding dependence on potentially conservative global constants. Second, this complexity order is consistent with that of PDHG for linear programming, as described in [43]. Our results extend the analysis in [43] to general convex optimization problems and a broader class of algorithms.

4.3 Comparison between metric subregularity moduli

In this section, we provide a comparison between the three metric subregularity moduli α_G, α_L, and $\alpha_{\mathcal{M}}$, defined in (9), (14) and (16), respectively, which underpin our theoretical results. We start by showing that under mild conditions, we have $\alpha_{\mathcal{M}} \geq \alpha_L \geq \alpha_G$.

Proposition 4.6. *The metric subregularity constants satisfy $\min\{\alpha_{\mathcal{M}}, \alpha_L\} \geq \alpha_G$. Furthermore, if the condition number $\kappa(P) \leq 4$, then, $\alpha_{\mathcal{M}} \geq \alpha_L$.*

The upper bound on $\kappa(P)$ is immaterial; except for ADMM, all algorithms we study satisfy it provided that we take the stepsize η small enough. Further, we could relax it, as it only reflects our choice of the identification radius of $\delta/2$. Specifically, we could modify Proposition 4.4 to reach the ball $\theta\delta$ with $\theta \in (0, 1/2]$ and all our rates will change by constants and the constraint here would reduce to $\kappa(P) \leq \theta^{-2}$. We decided to state this version of the results in favor of simplicity. The next example shows that the gap between these constants can be significant even in low dimensions.

Example 4.7. Let $c \in \mathbb{R}^2$ such that $c_1, c_2 > 0$ and $\|c\|_2 = 1$. Define

$$\min_{x \in \mathbf{R}^2} \langle c, x\rangle \qquad \text{s.t.} \qquad \langle c, x\rangle \geq \|c\|_1 \quad \text{and} \quad x_1, x_2 \geq 0.$$

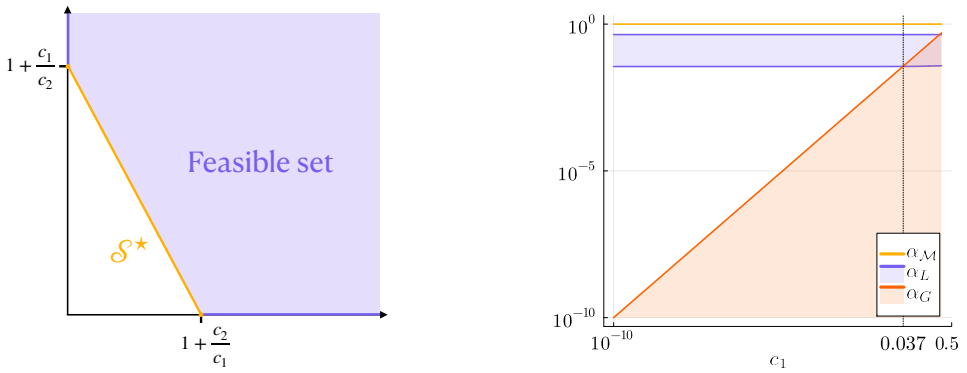
We show in Appendix C.6, using a blend of theoretical reductions and numerics, that when $\tau = 2$, $z^\star \in \operatorname{relint}(\mathcal{S}^\star)$ and $P = I$, we have

$$\alpha_G \leq \min\{c_1, c_2\}, \quad 0.037 \leq \alpha_L \leq 0.44, \quad \text{and} \quad \alpha_{\mathcal{M}} = 1.$$

Figure 3b shows that for small values of $\min\{c_1, c_2\}$, the global modulus α_G is orders of magnitude smaller than the local moduli α_L and $\alpha_{\mathcal{M}}$.

5 Experiments

In this section, we present numerical results to verify our major theoretical findings, i.e., finite-time identification with subsequent linear convergence, even in presence of degeneracy. In particular, we



(a) Primal geometry.

(b) Bounds on $\alpha_{\mathcal{M}}$, α_L, α_G.

Figure 3: Example 4.7. The left plot shows the feasible set and solutions $\mathcal{S}^\star$. The right plot displays the regions where the metric subregularity moduli could land versus c_1 (we take $c_1 \leq c_2$).

run EGM, PDHG and ADMM over linear programming (LP), convex quadratic programming (QP), and convex quadratically constrained quadratic programming (QCQP) instances. PPM is excluded as its update rule does not have closed-form solution on these instances. The code for reproducing these experiments is available at

https://github.com/pizqleh/degenerate-active-set.

Experiment setup. We use a MacBook Pro with an Apple M1 chip and 16 GB of RAM for all experiments. We test the following three classes of problems.

LP. The instances are obtained by constructing root-node linear relaxations of mixed-integer programs from MIPLIB 2017, which we write in standard form

$$\min_{x\in\mathbf{R}^n} \langle c, x\rangle \quad \text{s.t.} \quad Ax \leq b\,.$$

We initialize algorithms either at zero or at a random point of a sphere of radius 10^3, if zero is already in the rapid convergence region.

Convex QP. We select instances from Maros-Meszaros datasets written in standard form

$$\min_{x\in\mathbf{R}^n} \langle c, x\rangle + \frac{1}{2}\langle x, Qx\rangle \quad \text{s.t.} \quad Ax \leq b\,.$$

We initialize either at zero or at a random point of the sphere of radius 10^4, if zero is already in the rapid convergence region.

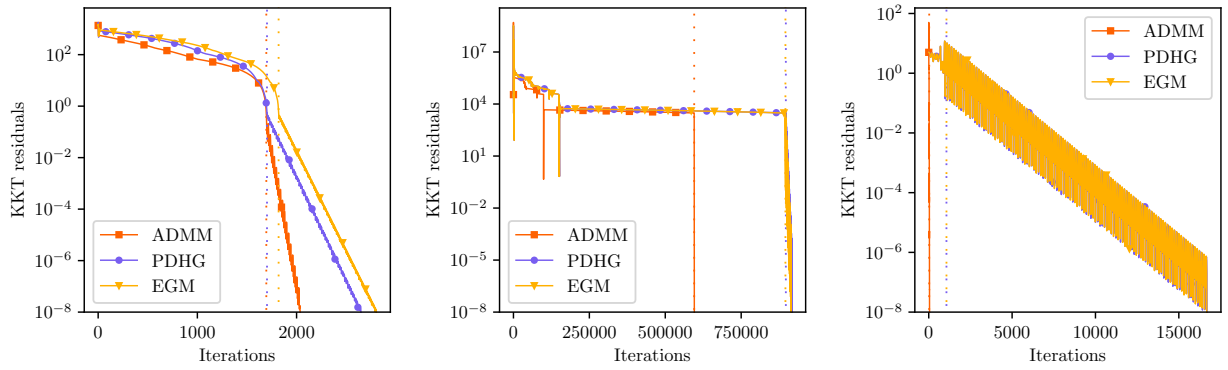
Convex QCQP. We consider instances from QPLIB dataset written in standard form

$$\min_{x\in\mathbf{R}^n} \langle c^0, x\rangle + \frac{1}{2}\langle x, Q^0 x\rangle \quad \text{s.t.} \quad \langle c^k, x\rangle + \frac{1}{2}\langle x, Q^k x\rangle \leq b^k \quad \text{for} \quad k \in \{1, \dots m\}\,.$$

We only consider EGM for this class, since all the other algorithms cannot handle quadratic constraints. We initialize EGM either at the suggested initial point z^0_{QPLIB} by QPLIB or, if that is in the rapid convergence region, at a random point of a sphere of radius 10^2 centered at z^0_{QPLIB}.

We set the iteration limit to 10^6 for all experiments. We use the stepsize $0.99 \cdot \|A\|_{\text{op}}^{-1}$ for all algorithms solving LPs and QPs.[3] For QCQP, we use the stepsize $(\|A\|_{\text{op}} + \sum_{k=0}^m \|Q^k\|_{\text{op}})^{-1}$, where

[3]This step size ensures convergence for PDHG and ADMM, but not necessarily for EGM. Nevertheless, EGM still converges in the examples we tested.



(a) `stein9inf` (b) `gt2` (c) `noswot`

Figure 4: Root-node relaxation of MIPLIB 2017 linear programs. Vertical dotted lines indicate active set identification. Figures 4b and 4c all methods converge to degenerate solutions. In 4c, ADMM identifies $\mathcal{M}$ and converges in just a couple of iterations.

A is the matrix whose kth row is c^k.[4]

Convergence criterion. We report the KKT residual as progress measure, namely,

$$\mathrm{KKT}(x,y) = \left\| \begin{bmatrix} (f(x) - h(y))_+ \\ (G(x))_+ \\ (-y)_+ \end{bmatrix} \right\|_2 .$$

The KKT residual penalizes the deviations from the KKT system (7), whence it is zero if and only if $z = (x, y) \in \mathcal{S}^\star$. We terminate the algorithms when the KKT residual of their iterates is no larger than tolerance 10^{-8}. We denote $\bar{k}$ as the index of the last iteration of the generated sequence, and we set the converging optimal solution $z^\star = z^{\bar{k}}$.

Active set identification and degeneracy. In order to account for numerical inaccuracies, we define the approximate identifiable set as

$$\mathcal{M}^\varepsilon = \{(x,y) \in \mathbf{R}^n \times \mathbf{R}^m_+ \,|\, G(x)_{N^\varepsilon} < -\varepsilon,\ |y_{N^\varepsilon}| < \varepsilon,\ \text{and}\ y_{B_a^\varepsilon} > \varepsilon\}\,,$$

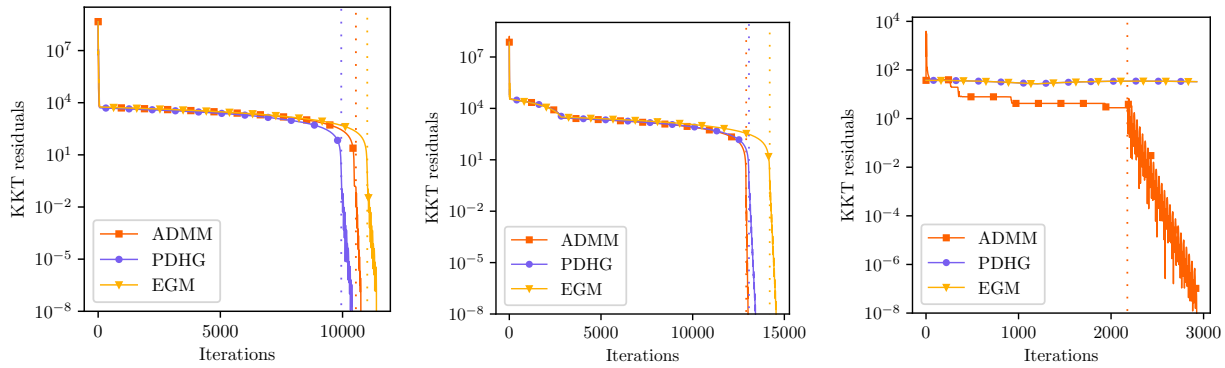
where $N^\varepsilon = \{j \in [m] : g_j(x^{\bar{k}}) < -\varepsilon,\ |y_j^{\bar{k}}| < \varepsilon\}$ and $B_a^\varepsilon = \{j \in [m] : y_j^{\bar{k}} > \varepsilon\}$. Here, we set the numerical tolerance to $\varepsilon = 10^{-10}$. We define the iteration at which the algorithm identifies the active set as

$$k^\star = \{\min k \in [\bar{k}] : z^\ell \in \mathcal{M}^\varepsilon \quad \text{for all} \quad \ell \geq k\}\,.$$

The limiting solution $z^{\bar{k}}$ is declared degenerate if the index set $B_d^\varepsilon = \{j \in [m] : |g_j(x^{\bar{k}})| < \varepsilon,\ |y_j^{\bar{k}}| < \varepsilon\}$ is nonempty.

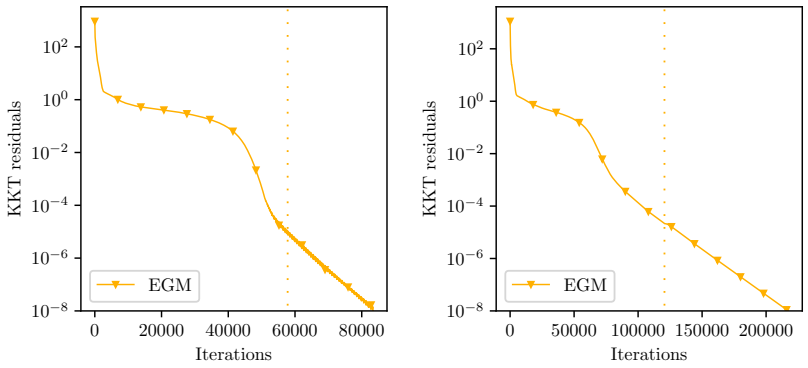
Results. Figure 4 and 5 present the behavior of EGM, PDHG and ADMM on LPs and QPs respectively, while Figure 6 displays the behavior of EGM when applied to convex QCQP instances (note that PDHG and ADMM are not applicable to QCQPs as their constraints are not affine). As shown, all algorithms under consideration exhibit the expected two-stage behavior across all instances: the iterates initially converge sublinearly toward a solution, and after identification (marked by the vertical dotted lines in the figures), they transition to a much faster linear convergence regime. This

[4]This step size does not ensure convergence for EGM either, but it works in our tested examples.



(a) HS76 (b) ZECEVIC2 (c) QRECIPE

Figure 5: Maros-Meszaros convex quadratic programs. Vertical dotted lines indicate active set identification. In 5c, the only algorithm that converged after 10^5 iterations was ADMM; PDHG and EGM showed very slow progress.



(a) 7579 (b) 3678

Figure 6: QPLIB convex quadratic programs with quadratic constraints. Vertical dotted lines indicate active set identification. In both 6a and 6b the transition toward fast linear convergence begins prior to identification, and locks in to a stable linear rate once identification occurs.

behavior persists even when the algorithms converge to a degenerate solution. In particular, in Figures 4b, 4c, 5c, and 6b, the convergent methods approach a degenerate solution. These empirical observations support the theoretical results from Section 4.

Acknowledgments

We thank Robert M. Freund and Stephen Wright for insightful conversations and pointers to relevant related work.

References

[1] D. Applegate, M. Díaz, O. Hinder, H. Lu, M. Lubin, B. O'Donoghue, and W. Schudy. Practical large-scale linear programming using primal-dual hybrid gradient. *Advances in Neural*

Information Processing Systems, 34:20243–20257, 2021.

[2] D. Applegate, O. Hinder, H. Lu, and M. Lubin. Faster first-order primal-dual methods for linear programming using restarts and sharpness. *Mathematical Programming*, 201(1):133–184, 2023.

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

[4] D. Applegate, M. Díaz, O. Hinder, H. Lu, M. Lubin, B. O'Donoghue, and W. Schudy. Pdlp: A practical first-order method for large-scale linear programming. *arXiv preprint arXiv:2501.07018*, 2025.

[5] M. Benning and M. Burger. Modern regularization methods for inverse problems. *Acta numerica*, 27:1–111, 2018.

[6] J. Borwein and A. Lewis. *Convex analysis*. Springer, 2006.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[8] J. V. Burke and J. J. Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

[9] Y. Cai, A. Oikonomou, and W. Zheng. Finite-time last-iterate convergence for learning in multi-player games. *Advances in Neural Information Processing Systems*, 35:33904–33919, 2022.

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

[12] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] A. L. Dontchev and R. T. Rockafellar. Regularity and conditioning of solution mappings in variational analysis. *Set-Valued Analysis*, 12(1):79–109, 2004.

[14] A. L. Dontchev and R. T. Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.

[15] D. Drusvyatskiy and A. S. Lewis. Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential. *SIAM Journal on Optimization*, 23(1):256–267, 2013.

[16] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of operations research*, 43(3):919–948, 2018.

[17] D. Drusvyatskiy, B. S. Mordukhovich, and T. T. Nghia. Second-order growth, tilt stability, and metric regularity of the subdifferential. *arXiv preprint arXiv:1304.7385*, 2013.

[18] J. Fadili, J. Malick, and G. Peyré. Sensitivity analysis for mirror-stratifiable convex functions. *SIAM Journal on Optimization*, 28(4):2975–3000, 2018.

[19] J. Fadili, G. Garrigos, J. Malick, and G. Peyré. Model consistency for learning with mirror-stratifiable regularizers. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1236–1244. PMLR, 2019.

[20] L. Fan, F. Zhang, H. Fan, and C. Zhang. Brief review of image denoising techniques. *Visual computing for industry, biomedicine, and art*, 2(1):7, 2019.

[21] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975.

[22] W. L. Hare and A. S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

[23] W. L. Hare and A. S. Lewis. Identifying active manifolds. *Algorithmic Operations Research*, 2 (2):75–82, 2007.

[24] B. He and X. Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

[25] R. Henrion, A. Jourani, and J. Outrata. On the calmness of a class of multifunctions. *SIAM Journal on Optimization*, 13(2):603–618, 2002.

[26] A. J. Hoffman. On approximate solutions of systems of linear inequalities. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 174–176. World Scientific, 2003.

[27] A. Ioffe. Necessary and sufficient conditions for a local minimum. 1: A reduction theorem and first order conditions. *SIAM Journal on Control and Optimization*, 17(2):245–250, 1979.

[28] A. D. Ioffe. Metric regularity and subdifferential calculus. *Russian Mathematical Surveys*, 55 (3):501, 2000.

[29] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[30] S. Lee, S. J. Wright, and L. Bottou. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(6), 2012.

[31] C. Lemaréchal, F. Oustry, and C. Sagastizábal. The $\mathcal{U}$-lagrangian of a convex function. *Transactions of the American mathematical Society*, 352(2):711–729, 2000.

[32] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13 (3):702–725, 2002.

[33] A. S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.

[34] A. S. Lewis, J. Liang, and T. Tian. Partial smoothness and constant rank. *SIAM Journal on Optimization*, 32(1):276–291, 2022.

[35] J. Liang, J. Fadili, and G. Peyré. Local linear convergence of forward–backward under partial smoothness. *Advances in neural information processing systems*, 27, 2014.

[36] J. Liang, J. Fadili, G. Peyré, and R. Luke. Activity identification and local linear convergence of douglas–rachford/admm under partial smoothness. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 642–653. Springer, 2015.

[37] J. Liang, J. Fadili, and G. Peyré. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017.

[38] J. Liang, J. Fadili, and G. Peyré. Local convergence properties of Douglas–Rachford and alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, 172(3):874–913, 2017.

[39] J. Liang, J. Fadili, and G. Peyré. Local linear convergence analysis of primal–dual splitting methods. *Optimization*, 67(6):821–853, 2018.

[40] M. Lou, K. A. Verchand, S. Fridovich-Keil, and A. Pananjady. Accurate, provable, and fast nonlinear tomographic reconstruction: A variational inequality approach. *arXiv preprint arXiv:2503.19925*, 2025.

[41] H. Lu and J. Yang. On the infimal sub-differential size of primal-dual hybrid gradient method and beyond. *arXiv preprint arXiv:2206.12061*, 2022.

[42] H. Lu and J. Yang. On a unified and simplified proof for the ergodic convergence rates of ppm, pdhg and admm. *arXiv preprint arXiv:2305.02165*, 2023.

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

[44] R. Mifflin and C. Sagastizábal. On $\mathcal{VU}$-theory for functions with primal-dual gradient structure. *SIAM Journal on Optimization*, 11(2):547–571, 2000.

[45] A. Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[46] C. Oberlin and S. J. Wright. Active set identification in nonlinear programming. *SIAM Journal on Optimization*, 17(2):577–605, 2006.

[47] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169 (3):1042–1068, June 2016. URL http://stanford.edu/~boyd/papers/scs.html.

[48] J. F. Peña. An easily computable upper bound on the hoffman constant for homogeneous inequality systems. *Computational Optimization and Applications*, 87(1):323–335, 2024.

[49] D. Quoc Tran, M. Le Dung, and V. H. Nguyen. Extragradient algorithms extended to equilibrium problems. *Optimization*, 57(6):749–776, 2008.

[50] S. M. Robinson. *Some continuity properties of polyhedral multifunctions*. Springer, 1981.

[51] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[52] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

[53] R. T. Rockafellar and R. J. Wets. *Variational analysis*. Springer, 1998.

[54] E. K. Ryu and W. Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press, 2022.

[55] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. Osqp: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.

[56] S. J. Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.

[57] S. J. Wright. Modifying sqp for degenerate problems. *SIAM Journal on Optimization*, 13(2): 470–497, 2002.

[58] S. J. Wright. Constraint identification and algorithm stabilization for degenerate nonlinear programs. *Mathematical Programming*, 95(1):137–160, 2003.

[59] Z. 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] X. Yuan, S. Zeng, and J. Zhang. Discerning the linear convergence of admm for structured convex optimization through the lens of variational analysis. *Journal of Machine Learning Research*, 21(83):1–75, 2020.

[61] X. Y. Zheng and K. F. Ng. Metric subregularity of piecewise linear multifunctions and applications to piecewise linear multiobjective optimization. *SIAM Journal on Optimization*, 24(1):154–174, 2014.

[62] M. Zhu and T. Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA Cam Report*, 34:8–34, 2008.

A Missing details from Section 1

In this section, we show the details of the QP from Figure 1. The QP is written in standard form

$$\min_{x \in \mathbf{R}^n} \langle c, x \rangle + \frac{1}{2}\langle x, Qx \rangle \quad \text{s.t.} \quad Ax \leq b$$

where the objective is given by

$$c = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \quad \text{and} \quad Q = UDU^T, \quad \text{where} \quad D = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\text{and} \quad U = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix},$$

with $\theta = \pi/64$, and the constraints are defined by

$$A = \begin{bmatrix} 1 & 1/\kappa \\ -1 & 1/\kappa \\ 0 & 1 \\ -\zeta & 1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 1 \\ \kappa - \widehat{\delta} \\ k - \widehat{\delta}\left(1 - \frac{\zeta}{\kappa}\right) \end{bmatrix}, \tag{17}$$

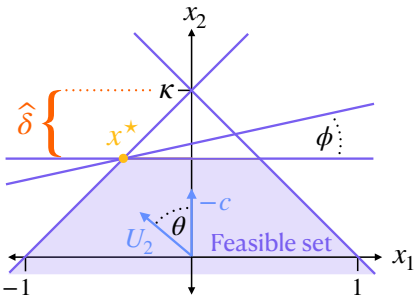


Figure 7: Primal geometry. The constant $\widehat{\delta}$ is a proxy for the radius of active set stability δ. The constant ζ in (17) satisfies $\zeta = \arctan(\phi)$. The vector U_2 is the second column of U and satisfies $\ker(Q) = \text{span}\{U_2\}$.

with $\zeta = 1/6$, $\kappa = 1/2$, and $\widehat{\delta} = 1/2^{10}$. To solve this QP, we initialize all algorithms at zero, and we use the following stepsizes: for PDHG, we use $0.99\|A\|_{\text{op}}^{-1}$; for ADMM, we use $2 \cdot 0.99\|A\|_{\text{op}}^{-1}$; and, for EGM, we use $0.99\sqrt{(\|Q\|_{\text{op}} + \|A\|_{\text{op}})^2 + \|A\|_{\text{op}}^2}^{-1}$. These choices ensure the convergence of their respective algorithms. The criteria for declaring 'convergence to the active set' and 'degeneracy' are the same as in Section 5. The convergence tolerance is set to 10^{-10}, while the active-set and degeneracy tolerances are set to 10^{-8}. The solution $(x^\star, y^\star)$ to which the algorithms converge satisfies

$$Ax^\star - b = (-3.906 \cdot 10^{-3}, 0, 0, 0) \quad \text{and} \quad y^\star = (0, 0, 0.863, 0.135)\,.$$

One can verify that $B_d = \{2\} \neq \emptyset$, hence this solution is degenerate.

B Missing proofs from Section 3

In this section, we present the missing proofs from Section 3.

B.1 Proof of Proposition 3.1

In this section, we derive the following slightly more general version of Proposition 3.1.

Proposition B.1 (Generalization of Proposition 3.1)**.** *Suppose Assumptions 1, 2, and 4 hold. Let z^k be the kth iterate generated by update (8). Then,*

$$0 < \alpha_G := \inf_{z \in \mathcal{D}} \frac{\text{dist}_2(0, \mathcal{F}(z))}{\text{dist}_2(z, \mathcal{S}^\star)} \qquad \textit{with} \qquad \mathcal{D} = \mathcal{S}^\star + \tau \mathbf{B}\,.$$

Further, for any $k \in \mathbf{N}$ we have

$$\max\left\{\text{dist}_2(z^k, \mathcal{S}^\star), \frac{\text{dist}_P(z^k, \mathcal{S}^\star)}{\lambda_{\max}(P)^{\frac{1}{2}}}\right\} \leq \sqrt{e}\nu \exp\left(-\frac{1}{2}\frac{k}{\lceil e\nu^2 \rceil}\right) \min\left\{\text{dist}_2(z^0, \mathcal{S}^\star), \frac{\text{dist}_P(z^0, \mathcal{S}^\star)}{\lambda_{\max}(P)^{\frac{1}{2}}}\right\}, \tag{18}$$

where $\nu = \gamma\, \lambda_{\max}(P)/\alpha_G$.

Notice that in this convergence statement, we simultaneously bound the P-seminorm and the ℓ_2-norm, which recovers the original statement. We prove this slightly stronger statement since it will be used later on in other proofs. Further, as Lemma B.2 below shows the minimum on the right-hand-side of (18) is always attained by the P-seminorm term. We included this redundancy as it makes it clear that this statement generalizes Proposition 3.1.

Before we delve into the proof of this result, we derive two auxiliary lemmas that we will use. The first lemma establishes the equivalence between the P-seminorm and the ℓ_2-norm on the range of P. The second lemma shows that 'Euclidean' metric subregularity implies P-norm metric subregularity.

Lemma B.2. *For any positive semidefinite matrix $P \in \mathcal{S}_+^{n+n}$ and point $z \in \text{range}(P)$, the following inequalities hold:*

$$\sqrt{\lambda_{\min}^+(P)}\|z\|_2 \leq \|z\|_P \leq \sqrt{\lambda_{\max}(P)}\|z\|_2$$
$$\frac{1}{\sqrt{\lambda_{\max}(P)}}\|z\|_2 \leq \|z\|_{P^\dagger} \leq \frac{1}{\sqrt{\lambda_{\min}^+(P)}}\|z\|_2$$

Proof. Define $p = \dim(\text{range}(P))$, $Q = \text{proj}_{\text{range}(P)} \in \mathbf{R}^{(n+m)\times(n+m)}$ and $U \in \mathbf{R}^{(n+m)\times p}$ such that $Q = UU^\top$ and $U^\top U = I$. Notice that since P is symmetric, $P = U\Sigma U^\top$ where $\Sigma \in \mathbf{R}^{p\times p}$ is a

diagonal matrix with non-negative entries. For any $z \in \operatorname{range}(P)$ we have

$$\|z\|_2^2 = \|Qz\|_2^2 = z^\top Q^\top Q z = z^\top U U^\top U U^\top z = z^\top U U^\top z = \|U^\top z\|^2 = \|\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}U^\top z\|_2^2,$$

where we used the invariance under projection of z. On the other hand,

$$\|\Sigma^{\frac{1}{2}}U^\top z\|_2^2 = \|P^{\frac{1}{2}}z\|_2^2 = \langle P^{\frac{1}{2}}z, P^{\frac{1}{2}}z\rangle = \langle z, Pz\rangle = \|z\|_P^2.$$

Also, by definition

$$\lambda_{\max}(\Sigma^{-\frac{1}{2}})^2 = \lambda_{\max}(\Sigma^{-1}) = \frac{1}{\lambda_{\min}^+(\Sigma)} = \frac{1}{\lambda_{\min}^+(P)}, \quad \text{and}$$
$$\lambda_{\min}^+(\Sigma^{-\frac{1}{2}})^2 = \lambda_{\min}^+(\Sigma^{-1}) = \frac{1}{\lambda_{\max}(\Sigma)} = \frac{1}{\lambda_{\max}(P)}.$$

Applying these identities in tandem with Cauchy-Schwarz gives

$$\|z\|_2^2 = \|\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}U^\top z\|_2^2 \le \lambda_{\max}(\Sigma^{-\frac{1}{2}})^2\|\Sigma^{\frac{1}{2}}U^\top z\|_2^2 = \frac{1}{\lambda_{\min}^+(P)}\|z\|_P^2, \quad \text{and}$$
$$\|z\|_2^2 = \|\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}U^\top z\|_2^2 \ge \lambda_{\min}^+(\Sigma^{-\frac{1}{2}})^2\|\Sigma^{\frac{1}{2}}U^\top z\|_2^2 = \frac{1}{\lambda_{\max}(P)}\|z\|_P^2.$$

The result for $\|z\|_{P^\dagger}$ follows analogously. □

Lemma B.3. *Let $P \in \mathcal{S}_+^{n+m}$ be a positive semidefinite matrix, $S \subseteq \mathbf{R}^{n+m}$ a set, and $z \in \mathbb{R}^{n+m}$ such that $\alpha \operatorname{dist}_2(z, S) \le \operatorname{dist}_2(0, \mathcal{F}(z))$ for some $\alpha > 0$. Then,*

$$\frac{\alpha}{\lambda_{\max}(P)^{\frac{1}{2}}} \max\left\{\operatorname{dist}_w(z, S), \frac{\operatorname{dist}_P(z, S)}{\lambda_{\max}(P)^{\frac{1}{2}}}\right\} \le \operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z) \cap \operatorname{range}(P)).$$

Proof. Let $\bar{z} \in S$ be arbitrary. Then,

$$\begin{aligned}
\|z - \bar{z}\|_2 &\ge \|\operatorname{proj}_{\operatorname{range}(P)}(z - \bar{z})\|_2 && \text{(Non-expansiveness)}\\
&\ge \lambda_{\max}(P)^{-\frac{1}{2}}\|\operatorname{proj}_{\operatorname{range}(P)}(z - \bar{z})\|_P && \text{(Lemma B.2)}\\
&= \lambda_{\max}(P)^{-\frac{1}{2}}\|z - \bar{z}\|_P,
\end{aligned}$$

where for the last line we used that $\operatorname{proj}_{\operatorname{range}(P)} P \operatorname{proj}_{\operatorname{range}(P)} = P$. Since $\bar{z} \in S$ is arbitrary, it follows that $\operatorname{dist}_P(z, S) \le \lambda_{\max}(P)^{\frac{1}{2}} \operatorname{dist}_2(z, S)$. On the other hand,

$$\begin{aligned}
\alpha \operatorname{dist}_2(z, S) &\le \operatorname{dist}_2(0, \mathcal{F}(z)) && \text{(By assumption)}\\
&\le \operatorname{dist}_2(0, \mathcal{F}(z) \cap \operatorname{range}(P)) && (\mathcal{F}(z) \cap \operatorname{range}(P) \subseteq \mathcal{F}(z))\\
&\le \lambda_{\max}(P)^{\frac{1}{2}} \operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z) \cap \operatorname{range}(P)). && \text{(Lemma B.2)}
\end{aligned}$$

The result follows immediately by combining these inequalities. □

We are ready to prove a generalization of Proposition 3.1 under our regularity assumptions.

Proof of Proposition B.1. Let us start by establishing a few consequences of metric subregularity. From Assumption 2 we obtain $\alpha_G > 0$. Furthermore, from Assumption 4 (ii) we have $z^k \in \mathcal{D}$ for all $k \in \mathbf{N}$. Then, for any $k \in \mathbf{N}$ we have $\alpha_G \operatorname{dist}_2(z^k, \mathcal{S}^\star) \le \operatorname{dist}_2(0, \mathcal{F}(z^k))$. Hence, invoking Lemma B.3, yields that for any $k \in \mathbf{N}$ we have

$$\alpha_G \max\left\{\operatorname{dist}_2(z^k, \mathcal{S}^\star), \frac{\operatorname{dist}_P(z^k, \mathcal{S}^\star)}{\lambda_{\max}(P)^{\frac{1}{2}}}\right\} \le \lambda_{\max}(P)^{\frac{1}{2}} \operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z^k) \cap \operatorname{range}(P)). \tag{19}$$

Fix the integer $\rho = \lceil e\nu^2 \rceil$ where $\nu = \gamma\lambda_{\max}(P)/\alpha_G$. The strategy to prove this result is simple: we show that after ρ consecutive iterations, the distance from 0 to $\mathcal{F}(z^k) \cap \operatorname{range}(P)$ contracts by a

constant factor, and, then, we relate this back to the distance from the iterates to the solution set. Consider two cases.

Case 1. First suppose that $k \geq \rho$ and let $n \in \mathbf{N}$ such that $n\rho \leq k \leq (n+1)\rho$. Hence,

$$\begin{aligned}
\operatorname{dist}^2_{P^\dagger}(0, \mathcal{F}(z^k) \cap \operatorname{range}(P)) &\leq \frac{\gamma^2}{\rho} \operatorname{dist}^2_P(z^{k-\rho}, \mathcal{S}^\star) && \text{(Assumption 4 } (i)\text{)} \\
&\leq e^{-1}(\lambda_{\max}(P)^{-1}\alpha_G)^2 \operatorname{dist}^2_P(z^{k-\rho}, \mathcal{S}^\star) && (\nu, \rho \text{ definition}) \\
&\leq e^{-1} \operatorname{dist}^2_{P^\dagger}(0, \mathcal{F}(z^{k-\rho}) \cap \operatorname{range}(P)) \,. && \text{(From (19))}
\end{aligned}$$

By recursively applying the same argument n times, we obtain

$$\begin{aligned}
\operatorname{dist}^2_{P^\dagger}(0, \mathcal{F}(z^k) \cap \operatorname{range}(P)) &\leq e^{-n} \operatorname{dist}^2_{P^\dagger}(0, \mathcal{F}(z^{k-n\rho}) \cap \operatorname{range}(P)) \\
&\leq e^{-n}\gamma^2 \operatorname{dist}^2_P(z^0, \mathcal{S}^\star) && \text{(Assumption 4 } (i)\text{)} \\
&\leq e^{1-\frac{k}{\rho}}\gamma^2 \operatorname{dist}^2_P(z^0, \mathcal{S}^\star) && (n \geq k/\rho - 1) \\
&\leq e^{1-\frac{k}{\rho}}\gamma^2 \lambda_{\max}(P) \operatorname{dist}^2_2(z^0, \mathcal{S}^\star) \,. && \text{(Lemma B.2)}
\end{aligned}$$

Taking a square root at both sides of the above inequality, we obtain

$$\operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z^k) \cap \operatorname{range}(P)) \leq \sqrt{e} \exp\left(-\frac{k}{2\rho}\right) \gamma \, \lambda_{\max}(P)^{\frac{1}{2}} \min\left\{ \frac{\operatorname{dist}_P(z^0, \mathcal{S}^\star)}{\lambda_{\max}(P)^{\frac{1}{2}}}, \operatorname{dist}_2(z^0, \mathcal{S}^\star) \right\} . \tag{20}$$

The result follows by combining (19) and (20).

Case 2. Suppose $k < \rho$, applying the same rationale as before we derive

$$\operatorname{dist}^2_{P^\dagger}(0, \mathcal{F}(z^k) \cap \operatorname{range}(P)) \leq \frac{\gamma^2}{k} \operatorname{dist}^2_P(z^0, \mathcal{S}^\star) \leq \frac{\gamma^2}{k} \lambda_{\max}(P) \operatorname{dist}^2_2(z^0, \mathcal{S}^\star).$$

Combining these inequalities with the fact that $\frac{1}{k} \leq 1 \leq e \exp(-1) \leq e \exp(-k/\rho)$ yields that (20) holds. Once more invoking (19) yields the stated bound, which completes the proof. □

B.2 Missing proofs from Section 3.3

In this section, we prove that the PPM, the ADMM, the PDHG method, and the EGM satisfy the assumptions of the meta-algorithm introduced in Section 3.2. We begin by weakening Assumption 3. This is required for the analysis of ADMM, which involves a positive semidefinite (PSD) matrix P instead of a positive definite (PD) matrix. Recall that $z^k = (x^k, y^k)$ and $\tilde{z}^k = (\tilde{x}^k, \tilde{y}^k)$ denote the main and auxiliary iterates of the meta-algorithm (8).

Assumption 3* (Weak Asymptotic Identification Conditions)**.** There exists a positive semidefinite matrix $P \in \mathcal{S}^{n+m}_+$ and a convex set $\mathcal{Z} \subseteq \mathbf{R}^{n+m}$ with $\mathbf{R}^n \times \mathbf{R}^m_+ \subseteq \mathcal{Z}$ such that the following hold.

(i) (**Convergence**) For any initial iterate $z^0 \in \mathbf{R}^{n+m}$ there exists $z^\star \in \operatorname{range}(P)$ such that

$$\max\{\|z^k - z^\star\|_P, \|\tilde{z}^k - z^\star\|_P\} \to 0 \,.$$

(ii) (**Dual update**) There exists $\eta > 0$ (stepsize) such that the dual update has the form

$$y^{k+1} = \operatorname{proj}_{\mathbf{R}^m_+}(y^k + \eta G(\tilde{x}^{k+1})) \,.$$

Further, $\tilde{z}^k \in \mathcal{Z}$.

(*iii*) (**Primal Lipschitzness**) For any radius $t > 0$ and index $j \in N$ there is a constant $L_{tj}^x \geq 0$ such that any point $z \in \overline{\mathbf{B}}_t^P(z^\star) \cap \mathcal{Z}$ satisfies

$$g_j(x) - g_j(x^\star) \leq L_{tj}^x \|z - z^\star\|_P \quad \text{and} \quad |g_j(x) - g_j(\widehat{x})| \leq L_{tj}^x \|z - \widehat{z}\|_P$$

for all $\widehat{z} \in \overline{\mathbf{B}}_t^P(z^\star) \cap \mathcal{Z}$ such that $\widehat{z}_{P^\perp} = z_{P^\perp}$.

(*iv*) (**Dual Lipschitzness**) There exists $p \in \mathbf{N} \cup \{0\}$ satisfying that for any radius $t > 0$ and index $j \in N \cup B_a$ there is a constant $L_{tj}^y \geq 0$ such that for any initial iterates $z^0, \widehat{z}^0 \in \mathbf{B}_t^P(z^\star) \cap \mathcal{Z}$ we have

$$y_j^p - y_j^\star \leq L_{tj}^y \|z^0 - z^\star\|_P \quad \text{if } j \in N, \qquad \text{and} \qquad |y_j^p - \widehat{y}_j^p| \leq L_{tj}^y \|z^0 - \widehat{z}^0\|_P \quad \text{if } j \in B_a.$$

Recall that z^p and $\widehat{z}^p$ are the p-th iterate of update (8) when initialized at z^0 and $\widehat{z}^0$, respectively.

When P is PD, Assumption 3 implies this weaker version. When P is PSD, the sequence generated by update (8) might not converge, which makes the analysis more nuanced. Indeed, our assumptions ensure that it converges in the P-seminorm, which is equivalent to having the sequence of iterates projected onto range(P) converges. Further, the point $z^\star$ might not belong to $\mathcal{S}^\star$. Moreover, the 'simpler' P-Lipschitz condition in Assumption 3 does not hold for ADMM. Nevertheless, as we will see, ADMM satisfies the Lipschitz conditions of Assumption 3* with $p = 1$. Also, we will see PPM, PDHG, and EGM satisfy Assumption 3, which implies Assumption 3* with $p = 0$.

The main ingredient in our argument is the next proposition, which identifies the generalized resolvent as a firmly nonexpansive mapping. This property immediately yields convergence of the corresponding fixed-point iteration, and in turn establishes the assumption in most cases. In its proof and throughout the rest of the appendix, we use the following notation for the decomposition of vectors onto the range of a matrix $P \in \mathcal{S}^d$ and its orthogonal complement. For any $u \in \mathbf{R}^d$, let

$$u = u_P + u_{P^\perp} \quad \text{where} \quad u_P \in \text{range}(P),\ u_{P^\perp} \in \text{range}(P)^\perp = \ker(P)\,. \tag{21}$$

The following is a well-known result; we include its proof for the reader's convenience.

Proposition B.4. *Let $T\colon \mathbf{R}^d \to \mathbf{R}^d$ be a mapping, $M\colon \mathbf{R}^d \rightrightarrows \mathbf{R}^d$ be a monotone set-valued operator,*[5] *and $P \in \mathcal{S}_+^d$ be a positive semidefinite matrix such that*

$$P(I - T)(u) \in M(T(u)) \quad \textit{for all } u \in \mathbf{R}^d\,.$$

Then, T is firmly non-expansive in the P seminorm, in the sense that

$$\|T(u) - T(v)\|_P^2 + \|(I - T)(u) - (I - T)(v)\|_P^2 \leq \|u - v\|_P^2 \quad \textit{for all } u, v \in \mathbf{R}^d\,.$$

In particular, if $\mathcal{S}_P^\star := \{u \in \mathbf{R}^d : T(u)_P = u_P\} \neq \emptyset$, for any $u \in \mathbf{R}^d$ there exists $u^\star \in \mathcal{S}_P^\star$ such that $\|T^k(u) - u^\star\|_P \to 0$ as $k \to \infty$.

Proof. Let $u, v \in \mathbf{R}^d$. Since $P(I-T)(u) \in M(T(u))$ and $P(I-T)(v) \in M(T(v))$, by the monotonicity of M we obtain

$$\langle T(u) - T(v), P(I - T)(u) - P(I - T)(v)\rangle \geq 0\,. \tag{22}$$

[5] An operator M is monotone if $\langle x - y, u - v\rangle \geq 0$ for all $x, y \in \mathbf{R}^d$, $u \in M(x)$ and $v \in M(y)$.

Therefore, expanding the square and lower-bounding the crossterm yields

$$\begin{aligned}\|u-v\|_P^2 &= \|u - T(u) + T(u) - T(v) + T(v) - v\|_P^2 \\ &= \|(I-T)(u) - (I-T)(v)\|_P^2 + \|T(u) - T(v)\|_P^2 \\ &\quad + 2\langle T(u) - T(v), P(I-T)(u) - P(I-T)(v)\rangle \\ &\geq \|(I-T)(u) - (I-T)(v)\|_P^2 + \|T(u) - T(v)\|_P^2 \qquad \text{(Using (22))}\end{aligned}$$

Then, $(T^k(u)_P)_k$ is a Krasnosel'skiĭ-Mann iteration over range(P) whence there exists a fixed point $u^\star \in \mathcal{S}_P^\star$ such that $\|T^k(u)_P - u^\star\|_P \to 0$ [54, Theorem 1]. □

Armed with Proposition B.4, we are ready to prove Propositions 3.2, 3.3, 3.4, and 3.5.

B.2.1 Proof of Proposition 3.2

By construction, the PPM update satisfies $P(z^k - z^{k+1}) \in \mathcal{F}(z^{k+1})$ with $P = \eta^{-1}I$. Then, Assumptions 3 (i), 4 (ii), and 5 (ii) follow from Proposition B.4. Moreover, [41, Theorem 1] shows

$$\|z^{k+1} - z^k\|_P \leq \operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z^k)) \leq \frac{1}{\sqrt{k}} \operatorname{dist}(z^0, \mathcal{S}^\star)\,.$$

Therefore, Assumptions 4 (i) and 5 (i) are satisfied with $\gamma = 1$. Moreover, note that the updates have the form

$$\begin{aligned}x^{k+1}, y^{k+1} &= \arg\min_{x\in\mathbf{R}^n} \max_{y\in\mathbf{R}_+^m} f(x) + \langle y, G(x)\rangle + \frac{1}{2\eta}\left(\|x - x^k\|^2 - \|y - y^k\|^2\right) \\ &= \arg\min_{x\in\mathbf{R}^n}\left\{f(x) + \frac{1}{2\eta}\|x - x^k\|^2 + \min_{y\in\mathbf{R}_+^m}\left\{-\langle y, G(x)\rangle + \frac{1}{2\eta}\|y - y^k\|^2\right\}\right\}\end{aligned}$$

Using first order optimality conditions we derive that the solution of the inner problem is $y(x) = \operatorname{proj}_{\mathbf{R}_+^m}(y^k + \eta G(x))$. Thus,

$$\begin{aligned}x^{k+1} &= \operatorname*{argmin}_{x\in\mathbf{R}^n}\Big\{f(x) + \frac{1}{2\eta}\|x - x^k\|^2 - \langle \operatorname{proj}_{\mathbf{R}_+^m}(y^k + \eta G(x)), G(x)\rangle \\ &\qquad + \frac{1}{2\eta}\|\operatorname{proj}_{\mathbf{R}_+^m}(y^k + \eta G(x)) - y^k\|^2\Big\} \\ y^{k+1} &= \operatorname{proj}_{\mathbf{R}_+^m}\left(y^k + \eta G(x^{k+1})\right)\end{aligned}$$

Then, PPM satisfies Assumption 3 (ii). To check Assumption 3 (iii) note that, by Assumption 1 (i), the functions g_j are locally Lipschitz [52, Theorem 10.4]. Then, since $P = \eta^{-1}I$, Assumption 3 (iii) holds with $L_{tj}^y = \eta$ and L_{tj}^x equal to η times a local Lipschitz constant of g_j over $\mathbf{B}_t(z^\star)$.

B.2.2 Proof of Proposition 3.3

From [42], the PDHG update satisfies $P(z^k - z^{k+1}) \in \mathcal{F}(z^{k+1})$ for P. A Schur complement argument reveals that $P \succ 0$ if, and only if, $\eta < \|A\|_{\text{op}}^{-1}$. Then, from Proposition B.4, Assumption 3 (i), 4 (ii), and 5 (ii) are satisfied. Also, [41, Theorem 1] shows that

$$\|z^{k+1} - z^k\|_P \leq \operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z^k)) \leq k^{-\frac{1}{2}} \operatorname{dist}(z^0, \mathcal{S}^\star)\,.$$

Note that $\|z^k - \tilde{z}^k\|_P \leq \|z^k - z^{k-1}\|_P$, thus, Assumptions 4 (i) and 5 (i) are satisfied with $\gamma = 1$. Moreover, the PDHG dual update trivially satisfies Assumption 3* (ii). Finally, by Lemma B.2 Assumption 3 (iii) holds with $L_{tj}^x \leq \|A\|_{\text{op}}\,\lambda_{\min}(P)^{-\frac{1}{2}}$ and $L_{tj}^y \leq \lambda_{\min}(P)^{-\frac{1}{2}}$.

B.2.3 Proof of Proposition 3.4

Our argument is based on the auxiliary primal problem

$$\min_{x\in\mathbf{R}^n, u\in\mathbf{R}^m} f(x) + \iota_{\mathbf{R}^m_+}(u) \quad \text{s.t. } Ax + u = b\,. \tag{23}$$

This problem is a recast of (6) into the standard form of ADMM [7, Section 3.1]. Explicitly, we add the auxiliary variable $u \geq 0$, which allows us to write the inequality constraint $Ax \leq b$ as an equality constraint. We focus on it because we can leverage the results from [42] to prove Assumptions 3* (i), 4, and 5 for the ADMM iterates for this problem. Subsequently, we can translate the results obtained for the auxiliary problem to the original problem. Thus, the proof follows from the next three steps.

Step 1. We prove that Assumptions 3* (i), 4 and 5 hold for the *auxiliary* problem (23). That is, they hold for the iterates $w^k = \tilde{w}^k = (u^k, x^k, y^k)$, where we take (u, x) as primal variables.

Step 2. Leveraging Step 1, we prove that Assumptions 3* (i), 4 and 5 also hold for the *original* problem (6). In this case, they hold for the iterates $z^k = \tilde{z}^k = (x^k, y^k)$.

Step 3. Finally, we show that items (ii), (iii), and (iv) of Assumption 3* hold for the original problem (23) with iterates $z^k = \tilde{z}^k = (x^k, y^k)$.

To establish these steps, we will use the Lagrangian of the original and auxiliary problems, that is,

$$\mathcal{L}(x, y) = f(x) + \langle y, Ax - b\rangle - \iota_{\mathbf{R}^m_+}(y), \quad \text{and} \quad \widehat{\mathcal{L}}(u, x, y) = f(x) + \iota_{\mathbf{R}^m_+(u)} + \langle y, Ax + u - b\rangle,$$

respectively. As well as their corresponding minimax subdifferential

$$\mathcal{F}(x, y) = \begin{bmatrix} \partial f(x) + A^\top y \\ b - Ax + N_{\mathbf{R}^m_+}(y) \end{bmatrix}, \quad \text{and} \quad \widehat{\mathcal{F}}(u, x, y) = \begin{bmatrix} N_{\mathbf{R}^m_+}(u) + y \\ \partial f(x) + A^\top y \\ b - Ax - u \end{bmatrix}. \tag{24}$$

Step 1. To analyze the iterates $w^k = \widehat{w}^k = (u^k, x^k, y^k)$ for the *auxiliary* problem (23), define the matrix $\widehat{P} \in \mathcal{S}^{m+n+m}$ such that

$$\widehat{P} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \eta A^\top A & A^\top \\ 0 & A & \frac{1}{\eta} I \end{bmatrix}.$$

A Schur-complement argument shows that this matrix is PSD regardless of η. Invovking [42, Corollary 3], we obtain that the ADMM update satisfies

$$\widehat{P}(w^k - w^{k+1}) \in \widehat{\mathcal{F}}(w^{k+1})\,. \tag{25}$$

Then, Assumption 5 (ii) holds follows from Proposition B.4. Further, there exists $w^\star = (u^\star, z^\star) \in \operatorname{range}(\widehat{P})$ such that $\|w^k - w^\star\|_{\widehat{P}} \to 0$, whence Assumption 3* (i) is also satisfied. Use $\widehat{\mathcal{S}}^\star$ to denote the set of primal dual solutions of the auxiliary problem (23), i.e., $\widehat{\mathcal{S}}^\star = \widehat{\mathcal{F}}^{-1}(0)$. Using [41, Theorem 1] we get

$$\|w^k - w^{k+1}\|_{\widehat{P}} \leq \operatorname{dist}_{\widehat{P}^\dagger}(0, \mathcal{F}(w^k)) \leq k^{-\frac{1}{2}} \operatorname{dist}(w^0, \widehat{\mathcal{S}}^\star)\,.$$

Therefore, Assumption 4 (i) and 5 (i) are satisfied with $\gamma = 1$.

Step 2. To analyze the iterates $z^k = \tilde{z}^k = (x^k, y^k)$, we now consider the matrix P. By the definition of $\widehat{P}$, for any $w = (u, z)$ and $\bar{w} = (\bar{u}, \bar{z})$ in $\mathbb{R}^{m+n+m}$ we have

(a) $\|w - \bar{w}\|_{\widehat{P}} = \|z - \bar{z}\|_P$, and

(*b*) $w \in \operatorname{range}(\widehat{P})$ if, and only if, $z \in \operatorname{range}(P)$.

Therefore, we obtain $\|z^k - z^\star\|_P \to 0$ from (*a*), and $z^\star \in \operatorname{range}(P)$ from (*b*). Thus, leveraging Step 1 we conclude that Assumption 3* (*i*) holds. Furthermore, applying (*a*) together with Step 1 we derive that for any $k \in \mathbf{N}$,

$$\|z^{k+1} - z^\star\|_P = \|w^{k+1} - w^\star\|_{\widehat{P}} \geq \|w^k - w^\star\|_{\widehat{P}} = \|z^k - z^\star\|_P \,.$$

Taking limits shows that item (*ii*) of Assumption 5 holds.

To show Assumption 5 (*i*), we claim that $(x, y) \in \mathcal{S}^\star$ if and only if $(u, x, y) \in \widehat{\mathcal{S}}^\star$ for $u = Ax - b \in \mathbf{R}^m_+$. Indeed, if $(x, y) \in \mathcal{S}^\star$ and $u = Ax - b$, then immediately $0 = b - Ax - u$ and $0 \in \partial f(x) + A^\top y$, by (24). Moreover, by feasibility and complementary slackness we have $y \geq 0$, $u \geq 0$, and $\langle y, u\rangle = 0$. It follows that $-y \in N_{\mathbf{R}^m_+}(u)$, or equivalently, $0 \in N_{\mathbf{R}^m_+}(u) + y$. Thus, according to (24), $(u, x, y) \in \widehat{\mathcal{S}}^\star$. On the other hand, if $(u, x, y) \in \mathcal{S}^\star$, then $u = b - Ax$, $0 \in \partial f(x) + A^\top y$, and $-y \in \mathbf{N}_{\mathbf{R}^m_+}(u)$, by (24). Hence, $y \in \mathbf{R}^m_+$ and $y_j > 0$ implies $u_j = 0$ for any $j \in [m]$, whence $-u \in N_{\mathbf{R}^m_+}(y)$. Thus, accordin to (24), $(x, y) \in \mathcal{S}^\star$, establishing the claim. Invoking (*a*) we obtain

$$\operatorname{dist}_{\widehat{P}}(w^0, \widehat{\mathcal{S}}^\star) = \operatorname{dist}_P(z^0, \mathcal{S}^\star) \tag{26}$$

and $\|w^k - w^\star\|_{\widehat{P}} = \|z^k - z^\star\|_P$. Hence, Step 1 implies that Assumption 5 (*i*) holds with $\gamma = 1$.

Next we turn to Assumption 4. Combining Step 1 and (26) we derive that it suffices to show

$$\operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z^k) \cap \operatorname{range}(P)) \leq \operatorname{dist}_{\widehat{P}^\dagger}(0, \widehat{\mathcal{F}}(w^k) \cap \operatorname{range}(\widehat{P})) \tag{27}$$

to establish item (*i*) of Assumption 4.

Note that $\widehat{P} = \widehat{M}\widehat{M}^\top$ and $P = MM^\top$, where

$$\widehat{M}^\top = \begin{bmatrix} 0 & \sqrt{\eta}A^\top & \sqrt{\eta}^{-1}I \end{bmatrix} \quad \text{and} \quad M^\top = \begin{bmatrix} \sqrt{\eta}A^\top & \sqrt{\eta}^{-1}I \end{bmatrix} . \tag{28}$$

The columns of $\widehat{M}$ and M are linearly independent. Applying Lemma B.5 (*i*) we derive

$$\begin{aligned} \operatorname{range}(\widehat{P}) &= \operatorname{range}(\widehat{M}) = \{(0, \eta A^\top v, v) : v \in \mathbf{R}^m\}, \quad \text{and} \\ \operatorname{range}(P) &= \operatorname{range}(M) = \{(\eta A^\top, v) : v \in \mathbf{R}^m\} \,. \end{aligned} \tag{29}$$

Let $\widehat{v} \in \widehat{\mathcal{F}}(w^k) \cap \operatorname{range}(\widehat{P})$ such that $\|\widehat{v}\|_{P^\dagger} = \operatorname{dist}_{P^\dagger}(0, \widehat{\mathcal{F}}(w^k) \cap \operatorname{range}(\widehat{P}))$. Since $\widehat{v} \in \operatorname{range}(\widehat{P})$, from (29) there exists $v \in \mathbf{R}^m$ such that $\widehat{v} = \widehat{M}v$. Moreover, since $\widehat{v} \in \widehat{\mathcal{F}}(w^k)$, the y-coordinates of $\widehat{M}v$—recall (28)—and of (25) are equal: we obtain $\sqrt{\eta}^{-1}v = b - Ax^k - u^k$. Then,

$$\begin{aligned} \operatorname{dist}_{P^\dagger}(0, \widehat{\mathcal{F}}(w^k) \cap \operatorname{range}(\widehat{P})) &= \|\widehat{M}v\|_{P^\dagger} && \text{(Definition of } \widehat{v} \text{ and } v) \\ &= \|v\|_2 && \text{(Lemma B.5 } (ii)) \\ &= \sqrt{\eta}\|b - Ax^k - u^k\|_2 \,. \end{aligned}$$

Also, invoking (25) we obtain $\widehat{\mathcal{F}}(w^k) \cap \operatorname{range}(\widehat{P}) \neq \emptyset$, and so by (24) and (29) we get $-y^k \in N_{\mathbf{R}^m_+}(u^k)$. Hence, $y^k \in \mathbf{R}^m_+$ and $y^k_j > 0$ implies $u^k_j = 0$ for any $j \in [m]$, whence $-u^k \in N_{\mathbf{R}^m_+}(y^k)$. Using an analogous reasoning for P and $\mathcal{F}$ we derive

$$\begin{aligned} \operatorname{dist}_{P^\dagger}(0, \mathcal{F}(z^k) \cap \operatorname{range}(P)) &= \sqrt{\eta}\, \inf\{\|b - Ax^k + \zeta\|_2 : \zeta \in N_{\mathbf{R}^m_+}(y^k)\} \\ &\leq \sqrt{\eta}\, \|b - Ax^k - u^k\|_2 && (-u^k \in N_{\mathbf{R}^m_+}(y^k)) \\ &= \operatorname{dist}_{\widehat{P}^\dagger}(0, \widehat{\mathcal{F}}(w^k) \cap \operatorname{range}(\widehat{P})) \,. \end{aligned}$$

Thus, (27) is holds true and so Assumption 4 (*i*) holds with $\gamma = 1$.

Step 3. Next we establish item (*ii*) of Assumption 3*. Equivalentely, we wish to show that the dual update y^{k+1} can be interpreted as a projected dual ascent update on the *original problem* (2),

i.e. $y^{k+1} = \mathrm{proj}_{\mathbf{R}^m_+}(y^k + \eta(Ax^k - b))$. Note that

$$\ker(P) = \{(x, y) \in \mathbf{R}^{n+m} : y + \eta Ax = 0\}\,. \tag{30}$$

Hence, recalling (21) we get

$$u^{k+1} = \mathrm{proj}_{\mathbf{R}^m_+}\left(-\eta^{-1}y^k - Ax^k + b\right) = \mathrm{proj}_{\mathbf{R}^m_+}\left(-\eta^{-1}y^k_P - Ax^k_P + b\right)\,. \tag{31}$$

Fix $j \in [m]$. Then, we have two cases

(a) Suppose $0 > (y^k + \eta(Ax^k - b))_j$. Then, we derive from (31) that

$$u^{k+1}_j = \mathrm{proj}_{\mathbf{R}_+}(-\eta^{-1}y^k - (Ax^k - b))_j = (-\eta^{-1}y^k - (Ax^k - b))_j\,.$$

Thus, $y^{k+1}_j = (y^k + \eta^{-1}(Ax^k - \eta y^k - Ax^k))_j = 0\,.$

(b) Suppose that $0 \le (y^k + \eta(Ax^k - b))_j$. Then, we obtain invoking from (31) that

$$u_j = \mathrm{proj}_{\mathbf{R}^+_m}(-\eta^{-1}y^k - (Ax^k - b))_j = 0\,.$$

Therefore, $y^{k+1}_j = y^k_j + \eta(Ax^k - b)_j\,.$

Thus, Assumption 3* (ii) holds.

Next, we will show that items (iii) and (iv) of Assumption 3* hold. To establish Assumption 3* (iii), consider $\mathcal{Z} = \mathbf{R}^n \times \mathbf{R}^m_+$ and let $z = (x, y)$ and $\widehat{z} = (\widehat{x}, \widehat{y})$ in $\mathcal{Z}$. Since $y_P + y_{P^\perp} \ge 0$, we obtain from (30) that $y_P \ge \eta A x_{P^\perp}$. Then, for any $j \in N$ (recall the definition (11)) we have

$$\begin{aligned}
|A_j x - A_j x^\star| &= |A_j x_P + A_j x_{P^\perp} - A_j x^\star| \\
&\le |A_j x_P + \eta^{-1}(y_P)_j - (A_j x^\star + \eta^{-1} y^\star_j)| && (y^\star_j = 0 \text{ since } j \in N) \\
&\le \|A_j\|_2 \|x_P - x^\star\|_2 + \eta^{-1}\|y_P - y^\star\|_2 && (\text{Triangle Inequality}) \\
&\le (\|A_j\|_2 + \eta^{-1})\|z_P - z^\star\|_2 \\
&\le (\|A_j\|_2 + \eta^{-1})\,\lambda^+_{\min}(P)\|z_P - z^\star\|_P\,. && (\text{Proposition B.2})
\end{aligned}$$

On the other hand, if $\widehat{z}_{P^\perp} = z_{P^\perp}$ then, from Proposition B.2 we obtain

$$A_j x - A_j \widehat{x} = A_j(x_P - \widehat{x}_P) \le \|A_j\|_2 \|x_P - \widehat{x}_P\|_2 \le \|A_j\|_2 \|z_P - \widehat{z}_P\|_2 \le \|A_j\|_2\, \lambda^+_{\min}(P)\|z - \widehat{z}\|_P$$

Thus, Assumption 3* (iii) holds with $L^x_{tj} \le (\eta^{-1} + \|A_j\|_2)\,\lambda^+_{\min}(P)^{-\frac{1}{2}}$.

Finally, we show that Assumption 3* (iv) holds with $p = 1$. Consider any initial iterates $z^0 = (x^0, y^0)$, $\widehat{z}^0 = (\widehat{x}^0, \widehat{y}^0)$, and index $j \in [m]$. Then,

$$\begin{aligned}
|y^1_j - \widehat{y}^1_j| &= |\mathrm{proj}_{\mathbf{R}_+}(y^0_j + \eta(A_j x^0 - b_j)) - \mathrm{proj}_{\mathbf{R}_+}(\widehat{y}^0_j + \eta(A_j \widehat{x}^0 - b_j))| && (\text{Assumption 3* } (ii)) \\
&= |\mathrm{proj}_{\mathbf{R}_+}((y^0_P)_j + \eta(A_j x^0_P - b_j)) - \mathrm{proj}_{\mathbf{R}_+}((\widehat{y}^0_P)_j + \eta(A_j \widehat{x}^0_P - b))| && (\text{From (30)}) \\
&\le |(y^0_P)_j + \eta(A_j x^0_P - b_j) - ((\widehat{y}^0_P)_j + \eta(A_j \widehat{x}^0 - b_j))| && (\text{Non-expansiveness}) \\
&\le \|y^0_P - \widehat{y}^0_P\|_2 + \eta\|A_j\|_{\mathrm{op}}\|x^0_P - \widehat{x}^0_P\|_2 && (\text{Triangle Inequality}) \\
&\le (1 + \eta\|A_j\|_{\mathrm{op}})\|z^0_P - \widehat{z}^0_P\|_2 \\
&\le (1 + \eta\|A_j\|_{\mathrm{op}})\,\lambda^+_{\min}(P)^{-\frac{1}{2}}\|z^0 - \widehat{z}^0\|_P && (\text{Proposition B.2}).
\end{aligned}$$

So Assumption 3* (iv) holds with $p = 1$ and $L^y_{tj} \le (1 + \eta\|A_j\|_{\mathrm{op}})\,\lambda^+_{\min}(P)^{-\frac{1}{2}}$.

B.2.4 Proof of Proposition 3.5

Invoking [9, Appendix I, Theorem 9], we derive that for any $k \in \mathbf{N}$ the iterates of EGM satisfy

$$\max(\eta \operatorname{dist}_2(0, \partial \mathcal{F}(z^k)), 2^{-1}\|z^k - z^{k+1}\|_2, \|z^k - \tilde{z}^k\|_2) \leq \frac{3}{\sqrt{1-(\eta L)^2}} \frac{\operatorname{dist}_2(z^0, \mathcal{S}^\star)}{\sqrt{k}}\,. \tag{32}$$

Therefore, EGM satisfies Assumptions 4 (i) and 5 (i) with $P = I$ and $\gamma = 3(1-(\eta L)^2)^{-\frac{1}{2}}$. To verify Asusmption 3 (i), first note that the iterates of EGM are bounded [9, Appendix G, Corollary 1]. Then, its set of accumulation points, denoted as $\mathbb{A}$, is nonempty. We will see that $\mathbb{A}$ is a singleton. By Assumption 1 (i), $\mathcal{F}$ has closed graph [52, Theorem 24.4]. Then, using (32), we derive $0 \in \mathcal{F}(z)$ for all $z \in \mathbb{A}$ and so $\mathbb{A} \subseteq \mathcal{S}^\star$. Further, EGM is Fèjer monotone [9, Appendix G, Lemma 1], that is,

$$\|z^{k+1} - z\|_2 \leq \|z^k - z\|_2 \quad \text{for all} \quad z \in \mathcal{S}^\star\,. \tag{33}$$

Next, we show that there is only one accumulation point. Suppose seeking contradiction that there exist $z, z' \in \mathbb{A}$ such that $z \neq z'$, and define $\Delta = \|z - z'\|_2/2$. Since $z \in \mathbb{A}$ there exists $k_0 \in \mathbf{N}$ such that $\|z^{k_0} - z\|_2 \leq \Delta$. Furthermore, from (33) we obtain $\|z^k - \tilde{z}^1\|_2 \leq \|z^{k_0} - z\|_2 \leq \Delta$ for all $k \geq k_0$. Then, by the triangular inequality, for all $k \geq k_0$,

$$\|z' - z^k\|_2 \geq \|z' - z\|_2 - \|z - z^k\|_2 \geq 2\Delta - \Delta = \Delta\,.$$

It follows that z' is not an accumulation point, yielding a contradiction. Thus, $\mathbb{A}$ is a singleton and EGM satisfies Assumptions 3 (i) and 4 (ii). Further, EGM satisfies Assumption 5 (ii) thanks to (33). By definition, EGM satisfies Assumption 3 (ii). To check Assumption 3 (iii) note that, by Assumption 1 (i), the functions g_j are locally Lipschitz [52, Theorem 10.4]. Then, since $P = I$, Assumption 3 (iii) holds with $L^y_{tj} = 1$ and L^x_{tj} as a local Lipschitz constant of g_j over $\mathbf{B}_t(z^\star)$.

B.2.5 An auxiliary result

We prove an auxiliary result used in the proof of Proposition 3.4.

Lemma B.5. *Let $S \in \mathcal{S}^q_+$ and $M \in \mathbf{R}^{q\times k}$ with linearly independent columns such that $S = MM^\top$. Then, the following two hold true.*

(i) *The ranges of these two matrices are the same* $\operatorname{range}(S) = \operatorname{range}(M)$.

(ii) *For any $v \in \mathbf{R}^k$ and $u = Mv$ we have $\|u\|_{S^\dagger} = \|v\|_2$.*

Proof. Since M has linearly independent columns, $M^\top$ is surjective. Hence, $\operatorname{range}(S) = \operatorname{range}(M)$, whence there exists $v \in \mathbf{R}^k$ such that $u = Mv$. Since M has linearly independent columns, $M^\dagger = (M^\top M)^{-1}M^\top$ notice that

$$S^\dagger = (MM^\top)^\dagger = (M^\dagger)^\top M^\dagger = ((M^\top M)^{-1}M^\top)^\top (M^\top M)^{-1}M^\top = M(M^\top M)^{-\top}(M^\top M)^{-1}M^\top\,.$$

Therefore, for any $v \in \mathbf{R}^k$ and $u = Mv$ we have

$$\|u\|^2_{S^\dagger} = \langle u, P^\dagger u\rangle_2 = \langle Mv, M(M^\top M)^{-\top}(M^\top M)^{-1}M^\top Mv\rangle_2 = \langle v, v\rangle_2 = \|v\|_2^2\,.$$

□

C Missing proofs from Section 4.1

In this section, we prove our identification results, that is, Theorem 4.1, 4.2, and Theorem 4.3. To establish these results we derive an algebraic charaterization of the radious of active-set stability.

C.1 Radius of active-set stability

In this section, we show that the geometric definition of the radius of active-set stability introduced in (12) can be characterized algebraically through the Lipschitz constants of the constraints and the primal-dual solution to which the algorithm converges. To this end, we define the 'optimal' Lipschitz moduli for a P-ball of radius t as stated in Assumption 3*. Formally, for all $j \in [m]$ we define the *Lipschitz modulus* functions $\overline{L_j^x} : (0, +\infty) \to \mathbf{R}$ and $\overline{L_j^y} : (0, +\infty) \to \mathbf{R}$ given by

$$\overline{L_j^x}(t) = \sup_{z=(x,y)\in\mathcal{Z}_t} \frac{(g_j(x) - g_j(x^\star))_+}{\|z - z^\star\|_P} \quad \text{and} \quad \overline{L_j^y}(t) = \sup_{z=(x,y)\in\mathcal{Z}_t} \frac{(y_j^\star - y_j^p)_+}{\|z - z^\star\|_P}, \tag{34}$$

where $(\cdot)_+$ extracts the nonnegative part of its input and $\mathcal{Z}_t = (\overline{\mathbf{B}}_t^P(z^\star) \cap \mathcal{Z}) \backslash (\{z^\star\} + \ker(P))$. Recall $\mathcal{Z}$ is defined as part of Assumption 3* (*iii*).

Now we extend the definition radius defined in (12) to account for the set $\mathcal{Z}$ introduced in Assumption 3*. Define

$$\delta_G := \sup \left\{ t \in (0, \infty) \;\middle|\; \text{For all } (x, y) \in \mathbf{B}_t^P(z^\star) \cap \mathcal{Z} \text{ we have } G_N(x) < 0, \text{ and } y_{B_a}^p > 0 \right\},$$

We consider an alternative characterization of this radius, dubbed δ_A, defined implicitly via

$$\delta_A = \min \left\{ \min_{j\in N} \frac{-g_j(x^\star)}{\overline{L_j^x}(\delta_A)}, \min_{j\in B_a} \frac{y_j^\star}{\overline{L_j^y}(\delta_A)} \right\}, \tag{35}$$

where $\overline{L_j^x}(\cdot)$ and $\overline{L_j^y}(\cdot)$ are the optimal Lipschitz modulus functions defined in (34). This gives a well-defined notion of radius. We write $\overline{L_{\delta j}^x}$ and $\overline{L_{\delta j}^y}$ as shorthands for $\overline{L_j^x}(\delta_A)$ and $\overline{L_j^y}(\delta_A)$.

Lemma C.1. *Suppose that Assumption 3* holds and that δ_G is finite. Then, the function $\Delta\colon (0, +\infty) \to \mathbf{R} \cup \{+\infty\}$ given by*

$$\Delta(t) = \min \left\{ \min_{j\in N} \frac{-g_j(x^\star)}{\overline{L_j^x}(t)}, \min_{j\in B_a} \frac{y_j^\star}{\overline{L_j^y}(t)} \right\}$$

has a unique fixed point. Thus, δ_A is well-defined and finite.

In turn, δ_A and δ_G are closely related. The following is the main result of this section.

Proposition C.2. *Suppose that Assumption 3* holds and that δ_G is finite. Then,*

$$\delta_A \le \delta_G.$$

Moreover, equality holds if Assumption 3 holds with $\mathcal{Z} = \mathbf{R}^{n+m}$ and $p = 0$.*

We remark that all algorithms satisfy Assumption 3* with $\mathcal{Z} = \mathbf{R}^{n+m}$ and $p = 0$, except for ADMM, which instead requires $\mathcal{Z} = \mathbf{R}^n \times \mathbf{R}_+^m$ and $p = 1$. The rest of this section is devoted to establishing Proposition C.2. We start with a proof of Lemma C.1, since it contains some of the necessary ingredients for establishing Proposition C.2.

Proof of Lemma C.1. We need to show that there exists $t^\star \in (0, \infty)$ such that $\Delta(t^\star) = t^\star$. To prove this, first we show that $t \mapsto \Delta(t)$ is continuous and *non-increasing* in its domain, which we prove is an unbounded open interval. Clearly, $t \mapsto t$ is also continuous but *increasing* without bounds. Then, we show that there exists $t \in \mathbf{R}$ such that $\Delta(t) \ge t$. Thus, by continuity and monotonicity, Δ and the identity function must intersect at some point greater than or equal to t. To formalize this, we divide the proof into four steps. The first three steps derive properties of $\overline{L_j^x}$, $\overline{L_j^y}$, and Δ, and the final step uses these properties to derive the conclusion.

Step 1: Continuity of $\overline{L_j^x}$ and $\overline{L_j^y}$. We will see that, under Assumption 3*, $\overline{L_j^x}$ is finite and locally Lipschitz-continuous if $j \in N$. The same applies to $\overline{L_j^y}$ if $j \in N \cup B_a$.

Finiteness follows directly from Assumption 3* (iii) and (iv). To establish the continuity of $\overline{L_j^x}$ for any fixed $j \in N$, define the auxiliary function $h\colon \mathbb{R}^{n+m}\backslash(\{z^\star\} + \ker(P)) \to \mathbf{R}$ given by

$$h(z) = \frac{g_j(x) - g_j(x^\star)}{\|z - z^\star\|_P} \quad \text{for any} \quad z = (x, y) \in \mathbb{R}^{n+m}\backslash(\{z^\star\} + \ker(P)) \,.$$

Also, for any $s_0 < s_1 \in (0, \infty)$, define the '$P$-shell'

$$\mathsf{S}(s_0, s_1) = \{z \in \mathcal{Z} : s_0 \le \|z - z^\star\|_P \le s_1\} \,.$$

Fix $t_0 \in (0, \infty)$. First, we will show that h is locally Lipschitz continuous uniformly along affine sets parallel to range(P). Define $s_0 \in (0, t_0)$ and $s_1 = 2t_0 - s_0$; note that $t_0 - s_0 = s_1 - t_0 > 0$. For any $z, z' \in \mathsf{S}(s_0, s_1)$ such that $z_{P^\perp} = z'_{P^\perp}$, define $\omega = \|z - z^\star\|_P$ and $\zeta = g_j(x) - g_j(x^\star)$ and analogously ω' and ζ'. By the triangle inequality and Assumption 3* (iii),

$$\begin{aligned} &s_0 \le \omega \le s_1, \qquad s_0 \le \omega' \le s_1, \qquad |\omega - \omega'| \le \|z - z'\|_P, \\ &\max\{|\zeta|, |\zeta'|\} \le L^x_{s_1 j} s_1, \quad \text{and} \quad |\zeta - \zeta'| \le L^x_{s_1 j} \|z - z'\| \,. \end{aligned} \tag{36}$$

Therefore,

$$\begin{aligned} |h(z) - h(z')| &= (\omega\omega')^{-1} |\omega'\zeta - \omega\zeta'| && \text{(Definitions of } h, \omega, \text{ and } \zeta) \\ &\le (\omega\omega')^{-1} (\omega' |\zeta - \zeta'| + \zeta' |\omega - \omega'|) && \text{(Triangle inequality)} \\ &\le 2 L^x_{s_1 j} s_1 s_0^{-2} \|z - z'\|_P \,. && \text{(From (36))} \end{aligned} \tag{37}$$

Hence h is uniformly Lipschitz continuous over $\mathsf{S}(s_0, s_1)$. Next, we establish the continuity of $\overline{L_j^x}(\cdot)$. Let $t \in (0, \infty)$ such that $|t - t_0| \le |s_0 - t_0|$. Denote $\bar{t} = \max\{t, t_0\}$, $\underline{t} = \min\{t, t_0\}$, and for any $\varepsilon > 0$ define $z^\varepsilon \in \mathsf{S}(\underline{t}, \bar{t})$ such that $h(z^\varepsilon) + \varepsilon \ge \sup\{h(z) : z \in \mathcal{S}(\underline{t}, \bar{t})\}$. Then,

$$\overline{L_j^x}(\bar{t}) = \max\left\{ \overline{L_j^x}(\underline{t}), \sup_{z \in \mathcal{S}(\underline{t}, \bar{t})} h(z) \right\} \le \max\left\{ \overline{L_j^x}(\underline{t}), h(z^\varepsilon) \right\} + \varepsilon \,, \tag{38}$$

where the first equality follows by definition. Let $u^\varepsilon \in \overline{\mathbf{B}}^P_{\underline{t}}(z^\star)$ such that $u^\varepsilon = z^\star + \underline{t} \|z^\varepsilon - z^\star\|_P^{-1} (z^\varepsilon_P - z^\star) + z^\varepsilon_{P^\perp}$. Since $z^\star \in \text{range}(P)$,

$$\|z^\varepsilon - u^\varepsilon\|_P = (1 - \underline{t} \|z^\varepsilon - z^\star\|_P^{-1}) \|z^\varepsilon - z^\star\|_P = \|z^\varepsilon - z^\star\|_P - \underline{t} \le \bar{t} - \underline{t} \,.$$

By construction, $u^\varepsilon_{P^\perp} = z^\varepsilon_{P^\perp}$. Then, from (37), i.e. the Lipschitz continuity of h, we obtain

$$|h(z^\varepsilon) - h(u^\varepsilon)| \le 2 s_1 L^x_{tj} s_0^{-2} \|z^\varepsilon - u^\varepsilon\|_P \le 2 L^x_{tj} s_1 s_0^{-2} (\bar{t} - \underline{t}) \,. \tag{39}$$

Also, since $u^\varepsilon \in \overline{\mathbf{B}}^P_{\underline{t}}(z^\star)$, from $\overline{L_j^x}$ definition we have $h(u^\varepsilon) \le \overline{L_j^x}(\underline{t})$. It follows that

$$\begin{aligned} |\overline{L_j^x}(\bar{t}) - \overline{L_j^x}(\underline{t})| &= \overline{L_j^x}(\bar{t}) - \overline{L_j^x}(\underline{t}) && \text{(Definition of } \overline{L_j^x}, \bar{t}, \text{ and } \underline{t}) \\ &\le \max\left\{ \overline{L_j^x}(\underline{t}), h(z^\varepsilon) \right\} - \overline{L_j^x}(\underline{t}) + \varepsilon && \text{(From (38))} \\ &\le \max\{0, h(z^\varepsilon) - h(u^\varepsilon)\} + \varepsilon \\ &\le 2 L^x_{s_1 j} s_1 s_0^{-2} (\bar{t} - \underline{t}) + \varepsilon \,. && \text{(From (39))} \end{aligned}$$

Since $\varepsilon > 0$ arbitrary, $|\overline{L_j^x}(t) - \overline{L_j^x}(t_0)| \le 2 L^x_{s_1 j} s_1 s_0^{-2} |t - t_0|$. Moreover, since $t_0 \in (0, \infty)$, $\xi \in (0, t_0)$, and $t \in [t_0 - \xi, t_0 + \xi]$ are aribitrary, $\overline{L_j^x}$ is locally Lipschitz continuous. The proof for the functions $\overline{L_j^y}$ with $j \in N \cup B_a$ is completely analogous.

Step 2: Positivity of $\overline{L_j^x}$ and $\overline{L_j^y}$. The functions $\overline{L_j^x}$ and $\overline{L_j^y}$ take non-negative values and are non-decreasing by construction. Furthermore, at least one of them is strictly positive for all t sufficiently large. This is because, since δ_G is finite, there exists $z^0 = (x^0, y^0) \in \mathbf{R}^{n+m}$ and $j \in N \cup B_a$ such that $g_j(x^0) > g_j(x^\star)$ or $y_j^p < y_j^\star$. Then, for $t_{\text{dom}} := \|z^0 - z^\star\|_P$, $\overline{L_j^x}(t_0)$ or $\overline{L_j^y}(t_0)$ is strictly positive. Furthermore, by monotonicity, $\overline{L_j^x}(t)$ or $\overline{L_j^y}(t)$ is strictly positive for all $t \geq t_{\text{dom}}$.

Step 3: Properties of Δ. The function Δ is positive, non-increasing, and locally Lipschitz-continuous on its domain, which is equal to an interval of the form (t, ∞) with $t \geq 0$. To prove this, first recall that the functions $\overline{L_j^x}$ and $\overline{L_j^y}$ take non-negative real values.Moreover, $-g_j(x^\star) > 0$ for all $j \in N$ and $y_j^\star > 0$ for all $j \in B_a$. Hence, Δ is positive. Also, the functions $\overline{L_j^x}$ and $\overline{L_j^y}$ non non-decreasing, whence Δ is non-increasing. Define

$$t_{\text{sup}} := \sup\{t \in (0, \infty) : \overline{L_j^x}(t) = 0 \text{ for all } j \in N \text{ and } \overline{L_j^y}(t) = 0 \text{ for all } j \in B_a\}\,. \tag{40}$$

From Step 2 we obtain that $t_{\text{sup}} \leq t_{\text{dom}} < \infty$. Since Δ is non-increasing, $\Delta(t) < \infty$ for all $t > t_{\text{sup}}$. Furthermore, by Step 1 the functions $\overline{L_j^x}$ and $\overline{L_j^y}$ are continuous, whence if the supremum is not equal to zero in (40), then it is attained and $\Delta(t_{\text{sup}}) = \infty$. It follows that $\text{dom}(\Delta) = (t_{\text{sup}}, \infty)$. Moreover, Δ is locally Lipschitz-continuous on $\text{dom}(\Delta)$ since it is the minimum of compositions of locally Lipschitz-continuous functions, that is, $\Delta = \min\{\min_{j\in N} h_j^N \circ \overline{L_{\delta j}^x}, \min_{j\in B_a} h_j^{B_a} \circ \overline{L_{\delta j}^y}\}$ where $h_j^N(t) = -g_j(x^\star)/t$ and $h_j^{B_a}(t) = y_j^\star/t$.

Step 4: Fixed point existence and uniqueness. Let $F\colon (0, \infty) \to [-\infty, \infty)$ be such that $F(t) = t - \Delta(t)$. Note that $\Delta(t^\star) = t^\star$ if and only if $F(t^\star) = 0$. The function F is increasing unboundedly and it is continuous, because Δ is non-increasing and continuous (see Step 3). Then, by the Intermediate Value Theorem, to verify the existence of $t^\star \in (0, \infty)$ such that $F(t^\star) = 0$ it suffices to show that exists $t \in (0, \infty)$ with $F(t) \leq 0$, or equivalently, $t \leq \Delta(t)$. If $t_{\text{sup}} = 0$ (see (40)), then $t \leq \Delta(t)$ for any $t \leq \min\{1, \Delta(1)\}$, because Δ is non-increasing by Step 3. If $t_{\text{sup}} > 0$, then $\Delta(t) \to \infty$ when $t \to t_{\text{sup}}$. This is because the supremum is attained in (40), as shown in Step 3, and the functions $\overline{L_j^x}$ and $\overline{L_j^y}$ are continuous, as established in Step 1. Hence, for any $\varepsilon > 0$ sufficiently small and $t = t_{\text{sup}} + \varepsilon$ we obtain $t \leq \Delta(t)$. Thus, there exists $t^\star \in (0, \infty)$ such that $F(t^\star) = 0$. Its uniqueness follows from the strict monotonicity of F. This concludes the proof of Lemma C.1.

□

Armed with this lemma, we are now ready to establish Proposition C.2.

Proof of Proposition C.2. We start with a claim that we will be useful for future arguments.

Claim C.3. *Suppose that Assumption 3* holds and that δ_G is finite. Then, for any $\theta \in (0, 1)$, and primal-dual point $z^0 = (x^0, y^0) \in \overline{\mathbf{B}}^P_{\theta\delta_A}(z^\star) \cap \mathcal{Z}$ we have*

$$-g_j(x^0) \geq (1-\theta)\overline{L_{\delta j}^x}\delta_A \quad \textit{for all } j \in N, \quad \textit{and} \quad y_j^p \geq (1-\theta)\overline{L_{\delta j}^y}\delta_A \quad \textit{for all } j \in B_a\,.$$

In particular, any primal-dual point $z^0 = (x^0, y^0) \in \mathbf{B}^P_{\delta_A}(z^\star) \cap \mathcal{Z}$ satisfies $G_N(x^0) < 0$ and $y_B^p > 0$.

Proof of Claim C.3. By Claim C.1, $\delta_A \in (0, \infty)$ is well defined. Let $z = (x, y) \in \overline{\mathbf{B}}^P_{\theta\delta_A}(z^\star) \cap \mathcal{Z}$ and $j \in N$. From (35), we derive $-g_j(x^\star) \geq \overline{L_{\delta j}^x}\delta_A$. Then,

$$-g_j(x) = -g_j(x^\star) - (g_j(x) - g_j(x^\star)) \geq -g_j(x^\star) - \overline{L_{\delta j}^x}\|z - z^\star\|_P \geq \overline{L_{\delta j}^x}\delta_A - \overline{L_{\delta j}^x}\theta\delta_A = (1-\theta)\overline{L_{\delta j}^x}\delta_A\,. \tag{41}$$

The result for the dual bound follows analogously. Note that for any $z^0 = (x^0, y^0) \in \mathbf{B}^P_{\delta_A}(z^\star)$ there exists $\theta \in (0,1)$ such that $z^0 \in \overline{\mathbf{B}}^P_{\theta\delta_A}(z^\star)$. Then, from (41), $g_j(x) < 0$ if $\overline{L^x_{\delta j}} > 0$. Otherwise, if $\overline{L^x_{\delta j}} = 0$, $g_j(x) \leq g_j(x^\star) < 0$, because $j \in N$. □

An immediate consequence of this claim is that $\delta_A \leq \delta_G$.

Next, we show that the converse inequality $\delta_A \geq \delta_G$ also holds when $\mathcal{Z} = \mathbf{R}^{n+m}$ and $p = 0$. Let $\varepsilon > 0$ and consider the ball $\mathbf{B}^P_{\delta_A+2\varepsilon}(z^\star)$. Since δ_A is the unique fixed point of Δ, and Δ is non-increasing, then $\Delta(\delta_A + \varepsilon) < \delta_A + \varepsilon$. Thus, either (a) there exists $j \in N$ such that $\delta_A + \varepsilon > -g_j(x^\star)/\overline{L^x_j}(\delta_A + \varepsilon)$, or (b) there exists $j \in B_a$ such that $\delta_A + \varepsilon > y^\star_j/\overline{L^y_j}(\delta_A + \varepsilon)$. Suppose (a) is true. Then, from the definition of $\overline{L^x_j}$, there exists $z \in \overline{\mathbf{B}}^P_{\delta_A+\varepsilon}(z^\star)\backslash(\{z^\star\} + \ker(P))$ with

$$\frac{g_j(x) - g_j(x^\star)}{\|z - z^\star\|_P} > \frac{-g_j(x^\star)}{\delta_A + \varepsilon}\,. \tag{42}$$

If $\|z - z^\star\|_P < \delta_A + \varepsilon$, define and $z' = z^\star + (\delta_A + \varepsilon)(z - z^\star)/\|z - z^\star\|_P$. Note that $z = \theta z' + (1-\theta)z^\star$ for $\theta = \|z - z^\star\|_P/(\delta_A + \varepsilon) \in (0,1)$. Also, $\|x - x^\star\|_2 \neq 0$, otherwise the LHS of (42) would be equal to zero, contradicting the inequality. Furthermore, by the homogeneity of the seminorm,

$$\zeta := \frac{\|z - z^\star\|_P}{\|x - x^\star\|_2} = \frac{\|\theta(z' - z^\star)\|_P}{\|\theta(x' - x^\star)\|_2} = \frac{\|z' - z^\star\|_P}{\|x' - x^\star\|_2}\,. \tag{43}$$

On the other hand, Assumption 1 (i) gives us that g_j is convex. Then, considering that z is a convex combination of z' and $z^\star$, and that $\|z' - z^\star\|_P = \delta_A + \varepsilon$, we obtain

$$\begin{aligned}
\frac{g_j(x') - g_j(x^\star)}{\delta_A + \varepsilon} &= \frac{g_j(x') - g_j(x^\star)}{\zeta\|x' - x^\star\|_2} && \text{(From (43))} \\
&\geq \frac{g_j(x) - g_j(x^\star)}{\zeta\|x - x^\star\|_2} && \text{(Convexity of } g_j\text{)} \\
&= \frac{g_j(x) - g_j(x^\star)}{\|z - z^\star\|_P} && \text{(From (43))} \\
&> \frac{-g_j(x^\star)}{\delta_A + \varepsilon} && \text{(From (42))}\,.
\end{aligned}$$

Therefore $g_j(x') - g_j(x^\star) > -g_j(x^\star)$ and, consequently, $g_j(x') > 0$. If (b) is true, by an analogous argument we get that there exists $j \in B_a$ and $z = (x, y) \in \overline{\mathbf{B}}^P_{\delta_A+\varepsilon}(z^\star)$ such that $y_j < 0$. In either case, the constraints $G_N(x^0) < 0$ and $y > 0$ are not satisfied at every point of $\mathbf{B}^P_{\delta_A+2\varepsilon}(z^\star)$. Thus, $\delta_G \leq \delta_A + 2\varepsilon$ and since ε was arbitrary, we obtain $\delta_G \leq \delta_A$. This concludes the proof of Proposition C.2.

□

C.2 Proof of Theorems 4.1 and 4.2

In this section, we prove generalized versions of Theorems 4.1 and 4.2 using Assumption 3*. We transcribe the statement of these results here for completeness. To do this, first we extend the set defined in (13) to account for the set $\mathcal{Z}$ introduced with Assumption 3*. In the context of Assumption 3*, we denote

$$\mathcal{Z}_0 = \{(x, y) \in \mathcal{Z} : y_N = 0\}.$$

Note that this is equal to (13) when $\mathcal{Z} = \mathbf{R}^{n+m}$.

Theorem C.4 (Generalization of Theorem 4.1)**.** *Suppose Assumptions 1 and 3* hold. Let z^k be the kth iterate generated by update (8). Then, there exists $K \in \mathbf{N}$ such that $z^k \in \mathbf{B}^P_{\delta_A}(z^\star) \cap \mathcal{Z}_0 \subseteq \mathcal{M}$*

for all $k \geq K$.[6]

Theorem C.5 (Generalization of Theorem 4.2)**.** *Suppose Assumptions 1, 3*, 4, and 5 hold and that the* k*-th iterate* z^k *and the* k*-th intermediate iterate* $\tilde{z}^k$ *of the meta-algorithm defined in (8) are equal. Then,* $z^k \in \mathbf{B}^P_{\delta_A/2}(z^\star) \cap \mathcal{Z}_0 \subseteq \mathcal{M}$ *provided*

$$k > \left\lceil e\left(\frac{\gamma\,\lambda_{\max}(P)}{\alpha_G}\right)^2\right\rceil \left[1 + 2\ln\left(\frac{4\,\lambda_{\max}(P)^{\frac{3}{2}}\,\mathrm{dist}_2(z^0, \mathcal{S}^\star)}{\alpha_G \delta_A}\right)\right] + \left\lceil p + \max_{j\in N} \frac{L^y_{\delta j}}{\eta C_j}\right\rceil,$$

where

$$C_j = \begin{cases} \overline{L^x_{\delta j}} & \text{if } \overline{L^x_{\delta j}} > 0 \\ -2g_j(x^\star)/\delta_A & \text{otherwise.}[6] \end{cases}$$

Let us make a couple of remarks about the statement of this result. Recall that $\overline{L^x_{\delta j}} := \sup_{z \in \mathbf{B}^P_\delta(z^\star)\setminus z^\star} \frac{g_j(x) - g_j(x^\star)}{\|z - z^\star\|_P}$. If g_j is constant in $\mathbf{B}^P_\delta(z^\star)$, then $\overline{L^x_{\delta j}} = 0$, whence the bound in Theorem 4.2 becomes unrealizable. The constant C_j in Theorem C.5 auxiliates this edge case. It is natural to wonder whether the condition $\tilde{z}^k = z^k$ for all $k \in \mathbf{N}$ is essential for the above result—note that ADMM and PPM satisfy it, but PDHG and EGM do not. This condition is not essential. The same conclusion holds provided that, for all $k \in \mathbf{N}$, $\mathrm{dist}_{P^\dagger}(0, \mathcal{F}(\tilde{z}^k)) \leq \frac{\gamma\,\mathrm{dist}_P(\tilde{z}^0, \mathcal{S}^\star)}{k^{\frac{1}{2}}}$. Under this bound, linear convergence of $\tilde{z}^k$ follows by an analogous argument to the one used for z^k (see Propositions 3.1 and B.1), and this is sufficient to conclude Theorem C.5.

The strategy we use to establish active set identification is straightforward. Recall that the identifiable set is characterized by the (in)equalities $G_N(x) < 0$, $y_{B_a} > 0$, and $y_N = 0$. Since the iterates converge to $z^\star$ in the P-norm, they remain inside $\mathbf{B}^P_{\delta_A/2}(z^\star)$ after a sufficient number of steps. Then, Claim C.3 gives us two conditions that any point inside this ball satisfies.

(i) The magnitude of the nonactive constraints G_N is negative and uniformly bounded away from zero, establishing the $G_N(x) < 0$ identification inequality.

(ii) The multipliers of the active constraints y_{B_a} are positive and uniformly bounded away from zero, establishing the $y_{B_a} > 0$ identification inequality.

To prove the remaining equality $y_N = 0$, note that any point inside $\mathbf{B}^P_{\delta_A/2}(z^\star)$ also satisfies that the multipliers of the nonactive constraints y_N are close to zero, since they are close to $y^\star_N = 0$. With this and (i), the iterative application of the meta-algorithm's dual update—gradient ascent projected onto $\mathbf{R}^m_+$—ensures that $y_N = 0$ after a sufficient number of steps. Finally, active set identification can be guaranteed in a *finite number of steps* leveraging the explicit rates of convergence of the meta-algorithm and the metric subregularity condition of the problem.

We begin with a few auxiliary propositions that will be used in both proofs. The next proposition provides a sufficient condition ensuring that the iterates eventually lie in the union of manifolds $\mathcal{M}$.

Proposition C.6. *Suppose Assumption 3* holds. For any initial iterate* z^0*, if* $z^0 \in \mathbf{B}^P_{\delta_A}(z^\star)$ *and* $z^p \in \mathbf{B}^P_{\delta_A}(z^\star) \cap \mathcal{Z}_0$*, then* $z^p \in \mathcal{M}$*. In particular, if* $p = 0$*, then* $\mathbf{B}^P_{\delta_A}(z^\star) \cap \mathcal{Z}_0 \subseteq \mathcal{M}$.[6]

Proof. Since, $z^0 \in \mathbf{B}^P_{\delta_A}(z^\star)$, from Claim C.3 we have $y^p_{B_a} > 0$. Similarly, since $z^p \in \mathbf{B}^P_{\delta_A}(z^\star)$, we have $G(x^p) > 0$. Also, since $z^p \in \mathcal{Z}_0$, we have $y^p_{B_d} \geq 0$ and $y^p_N = 0$. Thus, $z^p \in \mathcal{M}$. □

The following Proposition shows that once the iterates enter $\mathbf{B}^P_{\delta_A/2}(z^\star)$, it takes a few more steps to reach the active set $\mathcal{M}$.

[6]We emphasize that δ_A in this statement is the quantity defined in (35); by Proposition C.2, it coincides with δ_G as defined in (12) whenever Assumption 3* holds with $p = 0$.

Proposition C.7. *Suppose Assumption 3* holds and that the iterates satisfy* $z^k, \tilde{z}^k \in \overline{\mathbf{B}}^P_{\delta_A/2}(z^\star) \cap \mathcal{Z}$ *for all* $k \in \mathbf{N}_0$. *Then,* $z^k \in \mathcal{M}$ *for all*

$$k \geq p + \left\lceil \max_{j \in N} \frac{L^y_{\delta j}}{\eta C_j} \right\rceil =: K, \quad \text{where} \quad C_j = \begin{cases} \overline{L^x_{\delta j}} & \text{if } \overline{L^x_{\delta j}} > 0 \\ -2g_j(x^\star)/\delta_A & \text{otherwise.}^{6} \end{cases}$$

Proof. Fix $j \in N$. First, for any $k \geq p$ we aim to prove

$$y^k_j = \operatorname{proj}_{\mathbf{R}_+} \left(y^p_j + \sum_{\ell=p}^{k-1} \eta g_j(\tilde{x}^\ell) \right). \tag{44}$$

We proceed by induction. For $k = p$, the result is trivial. Now, assume (44) holds for some $k \geq p$. To exploit the induction hypothesis we claim that for any $a \in \mathbf{R}$ and $b \leq 0$ we have

$$\operatorname{proj}_{\mathbf{R}_+}((\operatorname{proj}_{\mathbf{R}_+} a) + b) = \operatorname{proj}_{\mathbf{R}_+}(a + b). \tag{45}$$

To prove this, suppose $a \geq 0$, then $\operatorname{proj}_{\mathbf{R}_+} a = a$ whence the result follows trivially. Now suppose $a \leq 0$, then $(\operatorname{proj}_{\mathbf{R}_+} a) + b = b \leq 0$ and $a + b \leq 0$, whence $\operatorname{proj}_{\mathbf{R}_+}((\operatorname{proj}_{\mathbf{R}_+} a) + b) = 0 = \operatorname{proj}_{\mathbf{R}_+}(a + b)$. Since $\tilde{z}^k \in \overline{\mathbf{B}}^P_{\delta_A/2}(z^\star) \cap \mathcal{Z}$, invoking Claim C.3 we derive $g_j(\tilde{x}^k) \leq -\overline{L^x_{\delta j}}\delta_A/2 \leq 0$. Thus,

$$\begin{aligned} y^{k+1} &= \operatorname{proj}_{\mathbf{R}_+}(y^k_j + \eta g_j(\tilde{x}^k)) && \text{(Assumption 3* } (ii)\text{)} \\ &= \operatorname{proj}_{\mathbf{R}_+} \left(\operatorname{proj}_{\mathbf{R}_+} \left(\sum_{\ell=p}^{k-1} y^p_j + \eta g_j(\tilde{x}^\ell) \right) + \eta g_j(\tilde{x}^k) \right) && \text{(Induction hypothesis)} \\ &= \operatorname{proj}_{\mathbf{R}_+} \left(y^p_j + \sum_{\ell=p}^{k} \eta g_j(\tilde{x}^\ell) \right) && \text{(Using (45)).} \end{aligned}$$

Thus, (44) holds for any $k \geq p$. Now, let $k \geq K$. Combining $z^p \in \overline{\mathbf{B}}^P_{\delta_A/2}(z^\star) \cap \mathcal{Z}$ and Assumption 3* (iv) we obtain

$$y^p_j = y^p_j - y^\star_j \leq L^y_{\delta j} \|z^0 - z^\star\|_P \leq L^y_{\delta j}\delta_A/2.$$

Moreover, using Claim C.3 in tandem with the fact that $\tilde{z}^\ell \in \overline{\mathbf{B}}^P_{\delta_A/2}(z^\star)$ for all $\ell \in \{0, \ldots, k-1\}$, we derive $g_j(\tilde{x}^\ell) \leq -\overline{L^x_{\delta j}}\delta_A/2$ for all $\ell \in \{0, \ldots, k-1\}$. If $\overline{L^x_{\delta j}} = 0$, then $g_j(\tilde{x}^\ell) = g_j(x^\star) < 0$, whence, $g_j(\tilde{x}^\ell) \leq -C_j\delta_A/2 < 0$. Consequently,

$$y^p_j + \sum_{\ell=p}^{k-1} \eta g_j(\tilde{x}^\ell) \leq L^y_{\delta j}\delta_A/2 - (k-p)\eta C_j \delta_A/2 \leq 0.$$

The last inequality holds because $k \geq K$. Thus $y^k_j = 0$, from (44). Since $j \in N$ is arbitrary, $y^k_N = 0$. Moreover, from Assumption 3* (ii) we have $y^k_B \geq 0$. Thus, $z^{k-p} \in \mathbf{B}^P_{\delta_A}(z^\star)$ and $z^k \in \mathbf{B}^P_{\delta_A}(z^\star) \cap \mathcal{Z}_0$, whence $z^k \in \mathcal{M}$, from Proposition C.6. □

Now we have the main ingredients for the proofs of Theorems C.4 and C.5.

Proof of Theorem C.4. From Assumption 3* (i) there exists $M \in \mathbf{N}$ such that $z^k, \tilde{z}^k \in \mathbf{B}^P_{\delta_A/2}(z^\star)$ for all $k \geq M$. From Proposition C.7 it suffices to take $K \geq M + p + \lceil \max\{L^y_{\delta j}(\eta C_j)^{-1} : j \in N\} \rceil$. □

Proof of Theorem C.5. Let $\varepsilon > 0$ and $\underline{z}^\star \in \mathcal{S}^\star$ such that $\operatorname{dist}_P(z^k, \mathcal{S}^\star) \geq \|z^k - \underline{z}^\star\|_P + \varepsilon$. Denote

$\nu = \gamma\,\lambda_{\max}(P)/\alpha_G$. $\rho = \lceil e\gamma\,\lambda_{\max}(P)/\alpha_G\rceil$. Then,

$$\begin{aligned}
\|z^k - z^\star\|_P &\le \|z^k - \underline{z}^\star\|_P + \|\underline{z}^\star - z^\star\|_P && \text{(Triangle inequality)}\\
&\le 2\|z^k - \underline{z}^\star\|_P && \text{(Assumption 5 } (ii)\text{)}\\
&\le 2\,\mathrm{dist}_P(z^k, \mathcal{S}^\star) + \varepsilon && \text{(Defintion of } \underline{z}^\star\text{)}\\
&\le 2\,\lambda_{\max}(P)^{\frac{1}{2}}\sqrt{e}\nu\exp\left(-\frac{1}{2}\frac{k}{\lceil e\nu^2\rceil}\right)\mathrm{dist}_2(z^0, \mathcal{S}^\star) + \varepsilon && \text{(Proposition B.1).}
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, $\|z^k - z^\star\|_P \le 2\gamma\,\lambda_{\max}(P)^{\frac{3}{2}}\sqrt{e}/\alpha_G \cdot \exp(-k/2\lceil e(\gamma\,\lambda_{\max}(P)/\alpha_G)^2\rceil) \cdot \mathrm{dist}_2(z^0, \mathcal{S}^\star)$. Thus, $\tilde{z}^k = z^k \in \mathbf{B}^P_{\delta_A/2}(z^\star)$ whenever

$$k > \left\lceil e\left(\frac{\gamma\,\lambda_{\max}(P)}{\alpha_G}\right)^2\right\rceil\left[1 + 2\ln\left(\frac{4\gamma\,\lambda_{\max}(P)^{\frac{3}{2}}\,\mathrm{dist}_2(z^0, \mathcal{S}^\star)}{\alpha_G\delta_A}\right)\right].$$

By Proposition C.7 it suffices at most $\lceil p + \max\{L^y_{\delta j}(\eta C_j)^{-1} : j \in N\}\rceil$ more steps to identify $\mathcal{M}$. □

C.3 Proof of Theorem 4.3

In this section, we prove a generalized version of Theorem 4.3 using Assumption 3*. We transcribe the statement of this result here for completeness.

Theorem C.8 (Generalization of Theorem 4.3)**.** *Under Assumptions 1, 3*, 4, and 5, the k-th iterate satisfy $z^k \in \mathcal{M}$ whenever*

$$k > \left(\max\left\{1, \frac{1}{\alpha_L}\right\}\frac{8\,\lambda_{\max}(P)^{\frac{3}{2}}\,\mathrm{dist}_P(z^0, \mathcal{S}^\star)}{\delta_A}\right)^2 + \left\lceil p + \max_{j\in N}\frac{L^y_{\delta j}}{\eta C_j}\right\rceil,$$

where

$$C_j = \begin{cases} \overline{L^x_{\delta j}} & \text{if } \overline{L^x_{\delta j}} > 0\\ -2g_j(x^\star)/\delta_A & \text{otherwise.}^{6}\end{cases}$$

We start with some notation and auxiliary results that will help us handle taking projections despite the lack of invertibility of P. Recall that, for any point $u \in \mathbf{R}^{n+m}$, we let

$$u = u_P + u_{P^\perp} \quad \text{where} \quad u_P \in \mathrm{range}(P),\ u_{P^\perp} \in \mathrm{range}(P)^\perp = \ker(P).$$

Similarly, for any set $S \subseteq \mathbf{R}^{n+m}$, we denote

$$S_P = \{u_P : u \in S\} \quad \text{and} \quad S_{P^\perp} = \{u_{P^\perp} : u \in S\}.$$

If S is a nonempty closed and convex set, S_P is also a nonempty and convex set that might not be closed. Thus, the P-projection of a point $z \in \mathbf{R}^n$ over S, $\mathrm{argmin}\{\|u - z\|_P : u \in S\}$, is not well defined in general. To overcome this difficulty, for any nonempty, convex $S \subseteq \mathbf{R}^n$ set, and $\varepsilon > 0$ we define the P-closure-projection and the (P, ε)-almost-projection as the point and set-valued map given by

$$\overline{\mathcal{P}}^P_S(z) = \operatorname*{argmin}_{u\in\mathrm{cl}(S_P)}\|u - z\|_P \quad \text{and} \quad \mathcal{P}^P_{S,\varepsilon}(z) = S \cap \overline{\mathbf{B}}^P_\varepsilon(\overline{\mathcal{P}}^P_S(z)), \quad \text{respectively.}$$

The P-closure-projection is well defined as a point instead of a set since $\mathrm{cl}(S_P)$ is nonempty, convex, and closed, and $\|\cdot\|_P$ is a norm when restricted to $\mathrm{range}(P)$. We will use the following auxiliary result.

Lemma C.9. *Let $S \subseteq \mathbf{R}^{n+m}$ a nonempty convex set, and $z, w \in \mathbf{R}^n$. Then, for any $u \in \mathcal{P}^P_{S,\varepsilon}(z)$ and $v \in \mathcal{P}^P_{S,\varepsilon}(w)$, the following hold true.*

(*i*) *(Convergence) Any sequence* $z^k \in \mathcal{P}^P_{S,\varepsilon_k}(z)$ *with* $\varepsilon_k \to 0$ *satisfies* $\|z^k - \overline{\mathcal{P}}^P_S(z)\|_P \to 0$.

(*ii*) *(Non-expansiveness)* $\|\overline{\mathcal{P}}^P_S(z) - \overline{\mathcal{P}}^P_S(w)\|_P \le \|z-w\|_P$ *and* $\|u-v\|_P \le \|z-w\|_P + 2\varepsilon$.

(*iii*) *(Fixed-points) For any* $z \in S$ *we have* $z = \overline{\mathcal{P}}^P_S(z) \in \mathcal{P}^P_{S,\varepsilon}(z)$.

(*iv*) *(Distance to S) For any* $\varepsilon > 0$ *and* $u \in \mathcal{P}^P_{S,\varepsilon}(z)$ *we have*

$$\operatorname{dist}_P(z,S) = \|z - \overline{\mathcal{P}}^P_S(z)\|_P \le \|z-u\|_P \le \operatorname{dist}_P(z,S) + \varepsilon\,.$$

Proof. Items (*i*) and (*iii*) follow directly by definition. Item (*iv*) follows easily from the Triangle Inequality. Finally, since $\|\cdot\|_P$ is a norm in range(P) and (norm)-projections over nonempty closed convex sets are nonexpansive, and so $\|\overline{\mathcal{P}}^P_S(z) - \overline{\mathcal{P}}^P_S(w)\|_P = \|u_P - v_P\|_P \le \|x_P - y_P\|_P = \|x-y\|_P$. Therefore, by the triangle inequality, we recover (*ii*)

$$\|u-v\|_P \le \|\overline{\mathcal{P}}^P_S(z) - \overline{\mathcal{P}}^P_S(w)\|_P + \|u - \overline{\mathcal{P}}^P_S(z)\|_P + \|v - \overline{\mathcal{P}}^P_S(w)\|_P \le \|x-y\|_P + 2\varepsilon.$$

□

Now we have all the ingredients for the proof of Theorem C.8.

Proof of Theorem C.8. Denote $\xi = \delta_A/2$ and

$$K := \left(\max\left\{1, \frac{1}{\alpha_L}\right\} \frac{8\,\lambda_{\max}(P)^{\frac{3}{2}}\operatorname{dist}_P(z^0, \mathcal{S}^\star)}{\delta_A}\right)^2.$$

For any $z \in \mathbf{R}^{n+m}$ and $\varepsilon > 0$, we write $\overline{\mathcal{P}}(z) := \operatorname{argmin}\{\|u-z\|_P : u \in \operatorname{cl}((\mathcal{S}^\star_L)_P)\}$, and $\mathcal{P}_\varepsilon(z)$ to denote some fixed element of $\mathcal{P}^P_{\mathcal{S}^\star_L,\varepsilon}(z) = \mathcal{S}^\star_L \cap \overline{\mathbf{B}}^P_\varepsilon(\overline{\mathcal{P}}(z))$ Any time we write $\mathcal{P}_\varepsilon(z)$, we refer to the same element of $\mathcal{P}^P_{\mathcal{S}^\star_L,\varepsilon}(z)$. The proof consists of six steps.

1. We show that for any $k > K$, we have
$$\|z^k - z^{k+1}\|_P < \xi/4 \quad \text{and} \quad \operatorname{dist}_P(z^k, \mathcal{S}^\star_L) < \xi/4 \quad \text{for all} \quad k > K\,. \tag{46}$$
2. We prove that for any $k > K$, *if* $\overline{\mathcal{P}}(z^k) \in \overline{\mathbf{B}}^P_{\xi/2}(z^\star)$ *then* $z^k \in \mathbf{B}^P_{\xi/2}(z^\star)$.
3. We establish that projections $\overline{\mathcal{P}}(z^k)$ are indeed close to the optimal solution $z^\star$ for large enough iterations. Formally, for any $k > K$ we have that $\overline{\mathcal{P}}(z^k) \in \overline{\mathbf{B}}^P_{\xi/2}(z^\star)$.
4. Combining steps 2 and 3, we conclude that any $k > K$ holds $z^k \in \mathbf{B}^P_{\xi/2}(z^\star)$.
5. We use Step 4 to show that for any $k > K$ we have $\tilde{z}^k \in \mathbf{B}^P_\xi(z^\star)$.
6. Leveraging Proposition C.7 we conclude that it takes $\lceil p + \max\{L^y_{\delta j}(\eta C_j)^{-1} : j \in N\}\rceil$ additional steps for the iterates to reach $\mathcal{M}$.

Next we execute these steps in detail.

Step 1. For the first inequality in (46), we have

$$\begin{aligned}\|z^{k+1} - z^k\|_P &\le \frac{\gamma}{k^{\frac{1}{2}}}\operatorname{dist}_P(z^0, \mathcal{S}^\star) && \text{(Assumption 5 (}i\text{))}\\ &\le \frac{\gamma}{k^{\frac{1}{2}}}\lambda_{\max}(P)^{\frac{1}{2}}\operatorname{dist}_2(z^0, \mathcal{S}^\star) && \text{(Proposition B.2)}\,.\end{aligned}$$

Therefore, $\|z^{k+1} - z^k\|_P < \xi/4$ provided that $k > K \geq (4\gamma\,\lambda_{\max}(P)^{\frac{1}{2}}\,\mathrm{dist}_P(z^0, \mathcal{S}^\star)/\xi)^2$. For the second inequality (46),

$$\begin{aligned}
\mathrm{dist}_P(z^k, \mathcal{S}_L^\star) &\leq \frac{1}{\alpha_L}\,\lambda_{\max}(P)\,\mathrm{dist}_{P^\dagger}(0, \mathcal{F}(z^k) \cap \mathrm{range}(P)) && \text{(Definition (14) \& Proposition B.3)} \\
&\leq \frac{\gamma}{\alpha_L k^{\frac{1}{2}}}\,\lambda_{\max}(P)\,\mathrm{dist}_P(z^0, \mathcal{S}^\star) && \text{(Assumption 4 } (i)\text{)} \\
&\leq \frac{\gamma}{\alpha_L k^{\frac{1}{2}}}\,\lambda_{\max}(P)^{\frac{3}{2}}\,\mathrm{dist}_2(z^0, \mathcal{S}^\star) && \text{(Proposition B.2).}
\end{aligned}$$

Therefore $\mathrm{dist}_P(z^k, \mathcal{S}_L^\star) < \xi/4$ whenever $k > K \geq (4\gamma\,\lambda_{\max}(P)^{\frac{3}{2}}\,\mathrm{dist}_P(z^0, \mathcal{S}^\star)/\alpha_L\xi)^2$.

Step 2. We make use of the following simple lemma.

Lemma C.10. *Suppose Assumption 3* holds. Then, $\mathcal{S}_L^\star \cap \mathbf{B}_{\delta_A}^P(z^\star) \subseteq \mathcal{S}^\star$.*

Proof of Lemma C.10. Let $z \in \mathcal{S}_L^\star \cap \mathbf{B}_{\delta_A}^P(z^\star)$. Note that $\mathcal{S}_L^\star \subseteq \mathcal{Z}$. Since $z \in \mathbf{B}_{\delta_A}^P(z^\star) \cap \mathcal{Z}$, from Claim C.3 we obtain $G(x)_N \leq 0$. Since $z \in \mathcal{S}_L^\star$, $G(x)_B \leq 0$, $y \geq 0$, and $f(x) - h(y) \leq 0$. Hence, $z \in \mathcal{S}^\star$. □

Let $k \geq K$ and suppose $\|\overline{\mathcal{P}}(z^k) - z^\star\|_P \leq \xi/2$. Then, there exists $\bar{\varepsilon} > 0$ such that $\mathcal{P}_\varepsilon(z^k) \in \overline{\mathbf{B}}_\xi^P(z^\star) \subseteq \mathbf{B}_{\delta_A}^P(z^\star)$ for all $\varepsilon \in (0, \bar{\varepsilon}]$. By Lemma C.10, $\mathcal{P}_\varepsilon(z^k) \in \mathcal{S}^\star$ for any $\varepsilon \in (0, \bar{\varepsilon}]$, whence

$$\begin{aligned}
\|z^k - z^\star\|_P &\leq \|z^k - \mathcal{P}_\varepsilon(z^k)\|_P + \|z^\star - \mathcal{P}_\varepsilon(z^k)\|_P && \text{(Triangle inequality)} \\
&\leq 2\|z^k - \mathcal{P}_\varepsilon(z^k)\|_P && \text{(Assumption 5 } (ii)\text{)} \\
&\leq 2\,\mathrm{dist}_P(z^k, \mathcal{S}_L^\star) + 2\varepsilon && \text{(Lemma C.9)}
\end{aligned}$$

Since $\varepsilon \in (0, \bar{\varepsilon}]$ is arbitrary, $\|z^k - z^\star\|_P \leq 2\,\mathrm{dist}_P(z^k, \mathcal{S}_L^\star)$. Then, from Step 1 we get $z^k \in \mathbf{B}_{\xi/2}^P(z^\star)$.

Step 3. We aim to prove that $\overline{\mathcal{P}}(z^k)$ belongs to $\mathbf{B}_{\xi/2}^P(z^\star)$ for any $k > K$ by contradiction. Let $k > K$ and suppose that $\|\overline{\mathcal{P}}(z^k) - z^\star\|_P > \xi/2$. By Assumption 3* (i), there exists a maximum $\ell \in \mathbf{N}$ such that $\|z^\star - z^\ell\|_P > \xi/2$. Our goal is to show that (a) $\ell > K$, and (b) z^ℓ and $z^{\ell+1}$ belong to two parallel halfspaces that are at distance $\xi/4$. These together contradict the bound on consecutive iterates derived in Step 1.

To prove that $\ell > K$, we claim that $z^\star \in \mathrm{cl}(\mathcal{S}_P^\star)$. To prove this, fix $\varepsilon > 0$. From Assumption 3* (i) and Proposition B.1 there exists $\bar{k} \in \mathbf{N}$ such that $\|z^{\bar{k}} - z^\star\|_P \leq \varepsilon/3$ and $\mathrm{dist}_P(z^{\bar{k}}, \mathcal{S}^\star) \leq \varepsilon/3$, respectively. Let $\underline{z}^\star \in \mathcal{S}^\star$ such that $\mathrm{dist}_P(z^{\bar{k}}, \mathcal{S}^\star) \geq \|z^{\bar{k}} - \underline{z}^\star\|_P + \varepsilon/3$. Then,

$$\begin{aligned}
\mathrm{dist}_P(z^\star, \mathcal{S}^\star) &\leq \|z^\star - \underline{z}^\star\|_P \\
&\leq \|z^\star - z^{\bar{k}}\|_P + \|z^{\bar{k}} - \underline{z}^\star\| \\
&\leq \|z^\star - z^{\bar{k}}\|_P + \mathrm{dist}_P(z^{\bar{k}}, \mathcal{S}^\star) + \varepsilon/3 \\
&\leq \varepsilon\,.
\end{aligned}$$

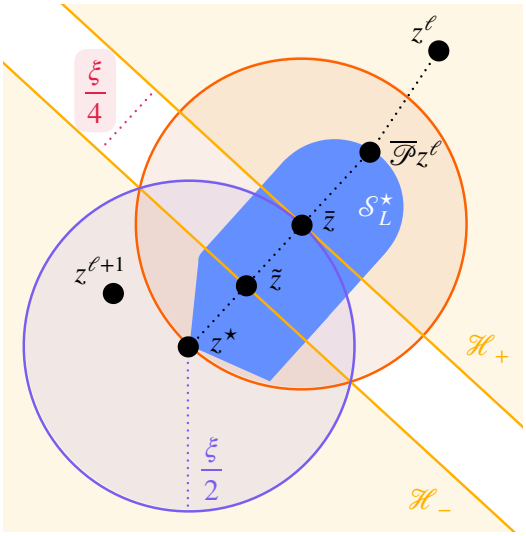


Figure 8: Visualization of Step 3. The orange ball $\overline{\mathbf{B}}_{\xi/2}^P(\bar{z})$ depicts a forbidden region for the iterates, given by star non-expansiveness (Assumption 5 (ii)). The blue solid region depicts the set $\mathcal{S}_L^\star$.

Since $\varepsilon > 0$ is arbitrary, $\operatorname{dist}_P(z^\star, \mathcal{S}^\star) = 0$, whence $z^\star \in \operatorname{cl}(\mathcal{S}^\star_P)$. Then, since $\operatorname{cl}(\mathcal{S}^\star_P) \subseteq \operatorname{cl}((\mathcal{S}^\star_L)_P)$, from Lemma C.9 (ii) we obtain

$$\|z^\star - z^k\|_P \geq \|z^\star - \overline{\mathcal{P}}(z^k)\|_P > \xi/2\,.$$

Therefore $\ell \geq k > K$, by maximality.

Now we delve into the proof of $\|z^\ell - z^{\ell+1}\|_P \geq \xi/4$. The strategy is to construct a point that is both in $\mathcal{S}^\star_L$ and in the boundary of $\mathbf{B}^P_{\xi/2}(z^\star)$ that faces z^ℓ; in particular, by Lemma C.10 it has to be also in $\mathcal{S}^\star$. From star non-expansiveness we obtain that $z^{\ell+1}$ has to be far away from that point, and from convexity we get that it has to be also far from z^ℓ—see Figure 8 for an illustration of the geometry of the argument. To formalize this, first note that, since $\|z^\star - z^\ell\|_P > \xi/2$, from Step 2 we obtain $\|z^\star - \overline{\mathcal{P}}(z^\ell)\|_P > \xi/2$. Then, there exists $\bar{\varepsilon} > 0$ such that

$$\inf_{\varepsilon \in (0, \bar{\varepsilon}]} \|z^\star - \mathcal{P}_\varepsilon(z^\ell)\|_P > \xi/2\,. \tag{47}$$

Also, since $z^\star \in \operatorname{cl}(\mathcal{S}^\star_P)$, for all $\varepsilon > 0$ there exists $z^{\star,\varepsilon} \in \mathcal{S}^\star$ such that $\|z^\star - z^{\star,\varepsilon}\|_P \leq \varepsilon$. Then, define

$$\bar{z} = z^\star + \frac{\xi}{2} \frac{\overline{\mathcal{P}}(z^\ell) - z^\star}{\|\overline{\mathcal{P}}(z^\ell) - z^\star\|_P} \quad \text{and} \quad \bar{z}^\varepsilon = z^{\star,\varepsilon} + \frac{\xi}{2} \frac{\mathcal{P}_\varepsilon(z^\ell) - z^{\star,\varepsilon}}{\|\mathcal{P}_\varepsilon(z^\ell) - z^{\star,\varepsilon}\|_P} \quad \text{for any } \varepsilon \in (0, \bar{\varepsilon}]\,. \tag{48}$$

From (47), $\bar{z}^\varepsilon$ is a convex combination of $z^{\star,\varepsilon} \in \mathcal{S}^\star \subseteq \mathcal{S}^\star_L$ and $\mathcal{P}_\varepsilon(z^{\ell+1}) \in \mathcal{S}^\star_L$, and from Assumption 1 (i), $\mathcal{S}^\star_L$ is convex, whence $\bar{z}^\varepsilon \in \mathcal{S}^\star_L$. Then, since $\bar{z}^\varepsilon \in \overline{\mathbf{B}}^P_{\xi/2}(z^\star)$, from Lemma C.10 we obtain $\bar{z}^\varepsilon \in \mathcal{S}^\star$. Denoting $\delta_A = \|\overline{\mathcal{P}}(z^\ell) - z^\star\|_P$ and $\Delta_\varepsilon = \|\mathcal{P}_\varepsilon(z^\ell) - z^\star\|_P$, by the triangle inequality we have $|\delta_A - \Delta_\varepsilon| \leq \|\overline{\mathcal{P}}(z^\ell) - \mathcal{P}_\varepsilon(z^\ell)\|_P \leq \varepsilon$ and $\delta_A - \varepsilon \leq \Delta_\varepsilon \leq \delta_A + \varepsilon$. Hence, for any $\varepsilon \in (0, \min\{\bar{\varepsilon}, 1, \delta_A/2\})$,

$$\begin{aligned}
\|\bar{z} - \bar{z}^\varepsilon\|_P &\overset{\text{(I)}}{\leq} \xi/2 \|\Delta_\varepsilon^{-1}(\mathcal{P}_\varepsilon(z^\ell) - z^{\star,\varepsilon}) - \delta_A^{-1}(\overline{\mathcal{P}}(z^\ell) - z^\star)\|_P + \|z^\star - z^{\star,\varepsilon}\|_P \\
&\overset{\text{(II)}}{\leq} \xi/2 \|\Delta_\varepsilon^{-1}(\mathcal{P}_\varepsilon(z^\ell) - z^{\star,\varepsilon}) - \delta_A^{-1}(\overline{\mathcal{P}}(z^\ell) - z^\star)\|_P + \varepsilon \\
&= \xi(2\delta_A\Delta_\varepsilon)^{-1} \|\delta_A(\mathcal{P}_\varepsilon(z^\ell) - z^{\star,\varepsilon}) - \Delta_\varepsilon(\overline{\mathcal{P}}(z^\ell) - z^\star)\|_P + \varepsilon \\
&= \xi(2\delta_A\Delta_\varepsilon)^{-1} \|\delta_A(\mathcal{P}_\varepsilon(z^\ell) - \overline{\mathcal{P}}(z^\ell)) + (\delta_A - \Delta_\varepsilon)\overline{\mathcal{P}}(z^\ell) + \delta_A(z^\star - z^{\star,\varepsilon}) + (\Delta_\varepsilon - \delta_A)z^\star\|_P \\
&\qquad + \varepsilon \\
&\overset{\text{(III)}}{\leq} \xi(2\delta_A\Delta_\varepsilon)^{-1}(\delta_A \|\mathcal{P}_\varepsilon(z^\ell) - \overline{\mathcal{P}}(z^\ell)\|_P + |\Delta_\varepsilon - \delta_A| \|\overline{\mathcal{P}}(z^\ell)\|_P + \delta_A \|z^\star - z^{\star,\varepsilon}\|_P \\
&\qquad + |\delta_A - \Delta_\varepsilon| \|z^\star\|_P) + \varepsilon \\
&\overset{\text{(IV)}}{\leq} \xi(2\delta_A\Delta_\varepsilon)^{-1}\varepsilon(2\delta_A + \|\overline{\mathcal{P}}(z^\ell)\|_P + \|z^\star\|_P) + \varepsilon \\
&\overset{\text{(V)}}{\leq} \varepsilon(1 + \xi\delta_A^{-2}(2\delta_A + \|\overline{\mathcal{P}}(z^\ell)\|_P + \|z^\star\|_P)) \\
&\to 0 \quad \text{as} \quad \varepsilon \to 0\,,
\end{aligned} \tag{49}$$

where (I) follows from the triangle inequality and (48); (II) from the definition of $z^{\star,\varepsilon}$; (III) from the Triangle Inequality and homogeneity; (IV) from the definitions of $z^{\star,\varepsilon}$, Δ_ε, and $\mathcal{P}_\varepsilon z^\ell$; and (V) from the inequalities $\Delta_\varepsilon \geq \delta_A - \varepsilon \geq \delta_A/2$. In particular, $\bar{z}_P \in \operatorname{cl}(\mathcal{S}^\star_P)$. Then, consider the halfspace

$$\mathcal{H}_+ = \{z \in \mathbf{R}^{n+m} : \langle z - \bar{z}, \bar{z} - z^\star \rangle_P \geq 0\}\,.$$

We claim that $z^\ell \in \mathcal{H}_+$. Suppose to the contrary that $z^\ell \notin \mathcal{H}_+$. Note that, by construction $\overline{\mathcal{P}}(z^\ell) - z^\star = 2\delta_A/\xi \cdot (\bar{z} - z^\star)$, whence $\overline{\mathcal{P}}(z^\ell) - \bar{z} = (\overline{\mathcal{P}}(z^\ell) - z^\star) - (\bar{z} - z^\star) = (2\delta_A\xi^{-1} - 1)(\bar{z} - z^\star)$,

and then $\langle z^\ell - \bar{z}, \overline{\mathcal{P}}(z^\ell) - \bar{z}\rangle_P < 0$ since $z^\ell \notin \mathcal{H}^+$ and $2\delta_A/\xi > 1$. It follows that

$$\begin{aligned}\|z^\ell - \overline{\mathcal{P}}(z^\ell)\|_P^2 &= \|z^\ell - \bar{z} - (\overline{\mathcal{P}}(z^\ell) - \bar{z})\|_P^2 \\ &= \|z^\ell - \bar{z}\|_P^2 - 2\langle z^\ell - \bar{z}, \overline{\mathcal{P}}(z^\ell) - \bar{z}\rangle_P + \|\overline{\mathcal{P}}(z^\ell) - \bar{z}\|_P^2 \\ &> \|z^\ell - \bar{z}\|_P^2 .\end{aligned}$$

This is contradiction to the definition of $\overline{\mathcal{P}}(z^\ell)$, since $\bar{z}_P \in \mathrm{cl}(\mathcal{S}_P^\star) \subseteq \mathrm{cl}((\mathcal{S}_L^\star)_P)$. Therefore, $z^\ell \in \mathcal{H}_+$. Now, let $\tilde{z} = (\bar{z} + z^\star)/2$, and consider the following halfspace

$$\mathcal{H}_- = \{z \in \mathbf{R}^{n+m} : \langle z - \tilde{z}, \tilde{z} - z^\star\rangle_P \leq 0\} .$$

We will show that $z^{\ell+1} \in \mathcal{H}_-$. From the definition of n, $z^{\ell+1} \in \overline{\mathbf{B}}_{\xi/2}^P(z^\star)$. Furthermore,

$$\begin{aligned}\|z^{\ell+1} - \bar{z}\|_P &\geq \|z^{\ell+1} - \bar{z}^\varepsilon\|_P - \|\bar{z} - \bar{z}^\varepsilon\|_P && \text{(Triangle inequality)} \\ &\geq \|z^\star - \bar{z}^\varepsilon\|_P - \|\bar{z} - \bar{z}^\varepsilon\|_P && \text{(Assumption 5 } (ii)) \\ &\geq \|z^\star - \bar{z}\|_P - 2\|\bar{z} - \bar{z}^\varepsilon\|_P && \text{(Triangle inequality)} \\ &\geq \xi/2 - 2\|\bar{z} - \bar{z}^\varepsilon\|_P && \text{(Definition of } \bar{z}) \\ &\to \xi/2 \quad \text{as} \quad \varepsilon \to 0 && \text{(From (49))} .\end{aligned}$$

Since ε is arbitrarily small, $z^{\ell+1} \notin \mathbf{B}_{\xi/2}^P(\bar{z})$. Then, to conclude that $z^{\ell+1} \in \mathcal{H}_-$ it suffices to show that $\overline{\mathbf{B}}_{\xi/2}^P(z^\star) \cap \mathcal{H}_-^c \subseteq \mathbf{B}_{\xi/2}^P(\bar{z})$. To prove this, let $u = 2/\xi \cdot (\bar{z} - z^\star)$. Note that $\tilde{z} - z^\star = (\bar{z} - z^\star)/2 = \xi/4 \cdot u$ and $z - \tilde{z} = z - (\bar{z} + z^\star)/2 = (z - z^\star) - (\bar{z} - z^\star)/2 = \xi/2 \cdot (2/\xi \cdot (z - z^\star) - u/2)$. Hence, $\langle z, -\tilde{z}, \tilde{z} - z^\star\rangle_P \leq 0$ if and only if $\langle 2/\xi \cdot (z - z^\star) - u/2, u\rangle_P \leq 0$. It follows that

$$\begin{aligned}\mathcal{H}_- &= z^\star + \xi/2 \cdot \{z \in \mathbf{R}^n : \langle z - u/2, u\rangle_P \leq 0\} \\ \overline{\mathbf{B}}_{\xi/2}^P(z^\star) &= z^\star + \xi/2 \cdot \overline{\mathbf{B}}_1^P(0), \quad \text{and} \\ \mathbf{B}_{\xi/2}^P(\bar{z}) &= z^\star + \xi/2 \cdot \mathbf{B}_1^P(u) .\end{aligned}$$

Then, by the possibility of translating by $-z^\star$ and then scaling by $2/\xi$, assume without loss of generality $\xi/2 = 1$ and $z^\star = 0$, implying $u = \bar{z}$ in the above representation. Let $z \in \overline{\mathbf{B}}_1^P(0) \cap \mathcal{H}_-^c$. Since $z \in \mathcal{H}_-^c$, $2\langle z, \bar{z}\rangle_P > \|\bar{z}\|_P^2 = 1$. Then, $\|\bar{z} - z\|_P^2 = \|\bar{z}\|_P^2 - 2\langle \bar{z}, z\rangle_P + \|z\|_P^2 < 2 - 1 = 1$. Therefore, $z \in \mathbf{B}_1^P(\bar{z})$. Since $z \in \overline{\mathbf{B}}_1^P(0) \cap \mathcal{H}_-^c$ is arbitrary, $\overline{\mathbf{B}}_1^P(0) \cap \mathcal{H}_-^c \subseteq \mathbf{B}_1^P(\bar{z})$. Thus, $z^{\ell+1} \in \mathcal{H}_-$. Now, we claim that $\mathrm{dist}_P(\mathcal{H}_+, \mathcal{H}_-) = \xi/4$. Similarly as before, for $u = 2/\xi \cdot (\bar{z} - z^\star)$ we have

$$\mathcal{H}_+ = z^\star + \xi/2 \cdot \{z \in \mathbf{R}^n : \langle z - u, u\rangle_P \geq 0\} .$$

Hence, assume without loss of generality $\xi/2 = 1$ and $z^\star = 0$. Let $z^- = \bar{z}/2 \in \mathcal{H}_-$ and $z^+ = \bar{z} \in \mathcal{H}_+$. Clearly, $\|z^- - z^+\|_P = \|\bar{z}\|_P/2 = 1/2$. Therefore $\mathrm{dist}_P(\mathcal{H}_-, \mathcal{H}_+) \leq 1/2$. For the reverse inequality, let $z^- \in \mathcal{H}_-$ and $z^+ \in \mathcal{H}_+$. Denote

$$z^\pm = z_{\bar{z}}^\pm + z_{\bar{z}^\perp}^\pm \quad \text{where} \quad z_{\bar{z}}^\pm = \bar{z}\left\langle z^\pm, \|\bar{z}\|_P^{-1}\bar{z}\right\rangle_P \quad \text{and} \quad z_{\bar{z}^\perp}^\pm = z^\pm - z^\pm ,$$

where $\pm$ represents $+$ or $-$. Then, by orthogonality, we have

$$\begin{aligned}\|z^+ - z^-\|_P^2 &= \|z_{\bar{z}}^+ - z_{\bar{z}}^-\|_P^2 + \|z_{\bar{z}^\perp}^+ - z_{\bar{z}^\perp}^-\|_P^2 \\ &\geq \|z_{\bar{z}}^+ - z_{\bar{z}}^-\|_P^2 \\ &= \left|\langle z^+ - z^-, \|\bar{z}\|_P^{-1}\bar{z}\rangle_P\right| \|\bar{z}\|_P \\ &\geq |\langle z^+, \bar{z}\rangle_P| - |\langle z^-, \bar{z}\rangle_P| \\ &\geq \|\bar{z}\|_P^2(1 - 1/2) \\ &= 1/2 .\end{aligned}$$

Therefore, $\operatorname{dist}_P(\mathcal{H}_-, \mathcal{H}_+) \geq 1/2 = \xi/4$. It follows that $\|z^{\ell+1} - z^\ell\| \geq \xi/4$. But, from Step 1 we have $\|z^\ell - z^{\ell+1}\|_P < \xi/4$, a contradiction.

Step 4. Combining Steps 2 and 3, it holds for any $k > K$ that $z^k \in \mathbf{B}^P_{\xi/2}(z^\star)$.

Step 5. From Assumption 5 (i) we have

$$\|\tilde{z}^k - z^\star\|_P \leq \gamma \operatorname{dist}_P(z^0, \mathcal{S}^\star)/k^{\frac{1}{2}}\,.$$

Then, whenever $k > K \geq (\gamma \operatorname{dist}_P(z^0, \mathcal{S}^\star)/\xi)^2$ we obtain $\|\tilde{z}^k - z^k\|_2 < \xi/2$, whence

$$\|\tilde{z}^k - z^\star\|_P \leq \|\tilde{z}^k - z^k\|_P + \|z^k - z^\star\|_P < \xi/2 + \xi/2 = \xi\,.$$

Step 6. Using Steps 4 and 5 we have that $z^k, \tilde{z}^k \in \mathbf{B}^P_\xi(z^\star)$ for all $k > K$. Then, by Proposition C.7 the iterates need at most $\lceil p + \max\{L^y_{\delta j}(\eta C_j)^{-1} : j \in N\}\rceil$ more steps to reach $\mathcal{M}$. □

C.4 Proof of Theorem 4.5

From Proposition 4.3, the kth iterate generated by update (8) satifies $z^k \in \mathbf{B}^P_{\delta_A/2}(z^\star) \cap \mathcal{M}$ whenever

$$k > K := \left(\max\left\{1, \frac{1}{\alpha_L}\right\} \frac{8\,\lambda_{\max}(P)^{\frac{3}{2}} \operatorname{dist}_2(z^0, \mathcal{S}^\star)}{\delta_A}\right)^2 + \left\lceil \max_{j \in N} \frac{L^y_{\delta j}}{\eta \overline{L}^x_{\delta j}} \right\rceil\,.^{6}$$

Then, from Proposition B.1, for all $k > K$ we have

$$\operatorname{dist}_2(z^k, \mathcal{S}^\star) \leq \sqrt{e}\nu \exp\left(-\frac{1}{2}\frac{k}{\lceil e\nu^2\rceil}\right) \frac{\delta_A}{\lambda_{\max}(P)^{\frac{1}{2}}} \quad \text{where} \quad \nu = \frac{\gamma\,\lambda_{\max}(P)}{\alpha_G}\,.$$

Rearranging, we recover the result of Theorem 4.5.

C.5 Proof of Proposition 4.6

$(\alpha_{\mathcal{M}} \geq \alpha_G)$ Clearly $\mathcal{D}_{\delta_A} \cap \mathcal{M} \subseteq \mathcal{D}$,[6] whence

$$\alpha_G = \inf_{z \in \mathcal{D}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} \leq \inf_{z \in \mathcal{D}_{\delta_A} \cap \mathcal{M}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star_L)} = \alpha_{\mathcal{M}}\,.$$

$(\alpha_L \geq \alpha_G)$ By construction, $\mathcal{S}^\star \subseteq \mathcal{S}^\star_L$, and then $\operatorname{dist}_2(z, \mathcal{S}^\star_L) \leq \operatorname{dist}_2(z, \mathcal{S}^\star)$ for all $z \in \mathbf{R}^{n+m}$. Hence,

$$\alpha_G = \inf_{z \in \mathcal{D}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} \leq \inf_{z \in \mathcal{D}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star_L)} = \alpha_L\,.$$

$(\alpha_{\mathcal{M}} \geq \alpha_L)$ Let $z \in \mathbf{B}^P_{\delta_A/2}(z^\star)$. Let $\underline{z}^\star \in \mathcal{S}^\star_L$ such that $\|z - \underline{z}^\star\|_2 = \operatorname{dist}_2(z, \mathcal{S}^\star_L)$. Suppose to the contrary that $\underline{z}^\star \notin \mathbf{B}^P_{\delta_A}(z^\star)$. In particular, $\underline{z}^\star \notin \overline{\mathbf{B}}^P_\zeta(z^\star)$ for $\zeta = \|z - z^\star\|_P$. Let $\zeta_P = \lambda_{\min}(P)^{-\frac{1}{2}}\zeta/2$. From Proposition B.2 and since $\kappa(P) \leq 4$,

$$\|\underline{z}^\star - z^\star\|_2 \geq \lambda_{\max}(P)^{-\frac{1}{2}}\|\underline{z}^\star - z^\star\|_P > \lambda_{\max}(P)^{-\frac{1}{2}}\zeta = \kappa(P)^{-\frac{1}{2}}\,\lambda_{\min}(P)^{-\frac{1}{2}}\zeta \geq \lambda_{\min}(P)^{-\frac{1}{2}}\zeta/2 = \zeta_P\,.$$

Thus, $\underline{z}^\star \notin \overline{\mathbf{B}}_{\zeta_P}(z^\star)$, whence $\operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)}\underline{z}^\star \neq \underline{z}^\star$. Conversely, from Proposition B.2 we obtain

$$\|z - z^\star\|_2 \leq \lambda_{\min}(P)^{-\frac{1}{2}}\|z - z^\star\|_P \leq \lambda_{\min}{}^{-\frac{1}{2}}\,\zeta/2 = \zeta_P\,.$$

Then, $z \in \overline{\mathbf{B}}_{\zeta_P}(z^\star)$, whence $\operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)} z = z$. From the firm-non-expansiveness of the projection, we obtain

$$\begin{aligned}\|\underline{z}^\star - z\|_2^2 &\geq \|\operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)}\underline{z}^\star - \operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)} z\|_2^2 + \|(\underline{z}^\star - \operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)}\underline{z}^\star) - (z - \operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)} z)\|_2^2 \\ &= \|z - \operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)}\underline{z}^\star\|_2^2 + \|\underline{z}^\star - \operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)}\underline{z}^\star\|_2^2 \\ &> \|z - \operatorname{proj}_{\overline{\mathbf{B}}_{\zeta_P}(z^\star)}\underline{z}^\star\|_2^2 .\end{aligned}$$

This contradicts that, by construction, $\underline{z}^\star = \operatorname{argmin}\{\|z - z'\|_2 : z' \in \mathcal{S}_L^\star\}$. Therefore, $\underline{z}^\star \in \mathbf{B}_{\delta_A}^P(z^\star)$. Moreover, from Lemma C.10, $\underline{z}^\star \in \mathcal{S}^\star$. Hence, $\operatorname{dist}_2(z, \mathcal{S}^\star) \leq \operatorname{dist}_2(z, \mathcal{S}_L^\star)$ for any $z \in \mathbf{B}_{\delta_A/2}^P(z^\star)$. Then, since $\mathcal{D}_{\delta_A} \subseteq \mathbf{B}_{\delta_A/2}^P(z^\star)$, we obtain

$$\alpha_L \leq \inf_{z \in \mathcal{D}_{\delta_A}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}_L^\star)} \leq \inf_{z \in \mathcal{D}_{\delta_A}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} \leq \inf_{z \in \mathcal{D}_{\delta_A} \cap \mathcal{M}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} = \alpha_{\mathcal{M}} .$$

Completing the proof of Proposition 4.6. □

C.6 Derivation of the results from Example 4.7

In this section, we prove that the problem of Example 4.7 satisfies

$$\alpha_G \leq \min\{c_1, c_2\}, \quad 0.037 \leq \alpha_L \leq 0.44, \quad \text{and} \quad \alpha_{\mathcal{M}} = 1.$$

To do this, rewrite the problem as

$$\min_{x \in \mathbf{R}^2} \langle c, x \rangle \quad \text{s.t.} \quad Ax \leq b, \quad \text{where} \quad A = -\begin{bmatrix} c_1 & c_2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b = -\begin{bmatrix} \|c\|_1 \\ 0 \\ 0 \end{bmatrix} .$$

By inspection, it is easy to see that

$$\mathcal{S}^\star = \mathcal{S}_x^\star \times \mathcal{S}_y^\star \quad \text{where} \quad \mathcal{S}_x^\star = \{x \in \mathbf{R}_+^2 : \langle c, x \rangle = \|c\|_1\} \quad \text{and} \quad \mathcal{S}_y^\star = \{(1, 0, 0)\} .$$

We now show an explicit form for the distance to zero of the saddle subdifferential. Let $z = (x, y) \in \mathbf{R}^n \times \mathbf{R}_+^m$ and $I = \{j \in [m] : y_j = 0\}$. Using Proposition C.11, we obtain

$$\begin{aligned}\operatorname{dist}_2^2(0, \mathcal{F}(z)) &= \|c + A^\top y\|_2^2 + \|\max(0, (Ax - b)_I)\|_2^2 + \|(Ax - b)_{I^c}\|_2^2 . \\ &= \|(1 - y_1)c - y_2 e_2 - y_3 e_3\|_2^2 + \left\| \left(\begin{bmatrix} \langle c, x \rangle - \|c\|_1 \\ x_1 \\ x_2 \end{bmatrix}_I \right)_- \right\|_2^2 + \left\| \begin{bmatrix} \langle c, x \rangle - \|c\|_1 \\ x_1 \\ x_2 \end{bmatrix}_{I^c} \right\|_2^2 \end{aligned} \tag{50}$$

where e_i is the i-th canonical vector.

The following analysis holds for $P = I$ and any solution $z^\star \in \operatorname{relint}(\mathcal{S}^\star)$.

Estimation of $\alpha_{\mathcal{M}}$. Let $z = (x, y) \in \mathbf{B}_\delta(z^\star) \cap \mathcal{M}$. Note that

$$\mathcal{M} = \{(x, y) \in \mathbb{R}^{n+m} : x_1, x_2 > 0,\ y_1 > 0,\ y_2, y_3 = 0\} .$$

Then $I = \{2, 3\}$ and (50) becomes

$$\operatorname{dist}_2^2(0, \mathcal{F}(z)) = \|c\|_2^2 (1 - y_1)^2 + (\langle c, x \rangle - \|c\|_1)^2 .$$

Furthermore, since $z \in \mathbf{B}_\delta(z^\star) \cap \mathcal{M}$,

$$\operatorname{dist}_2^2(z, \mathcal{S}^\star) = \operatorname{dist}_2^2(y, \mathcal{S}_y^\star) + \operatorname{dist}_2^2(x, \mathcal{S}_x^\star) = (1 - y_1)^2 + \left(\frac{\langle c, x \rangle - \|c\|_1}{\|c\|_2} \right)^2 .$$

But $\|c\|_2 = 1$, hence

$$\alpha_{\mathcal{M}} = \inf_{z \in \mathcal{D}_\delta \cap \mathcal{M}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} = 1\,.$$

Estimation of α_G. We provide upper bound estimates for α_G in terms of τ and $\min\{c_1, c_2\}$. *Dependence on* $\min\{c_1, c_2\}$. Assume without loss of generality that $c_1 \geq c_2$. For any $\varepsilon \in (0, \tau)$ let $\bar{z} = (\bar{x}, \bar{y}) \in \mathbb{R}^{n+m}$ such that $\bar{x} = (0, (1 + \frac{c_1}{c_2}) + \varepsilon)$ and $\bar{y} = (1, 0, 0)$. Note that $\bar{z} \in \mathcal{D}$. Moreover $c + A^\top \bar{y} = 0$ and $I = \{2, 3\}$, whence (50) becomes

$$\operatorname{dist}_2^2(0, \mathcal{F}(\bar{z})) = |(A\bar{x} - b)_1|^2 = ((c_2 + c_1) + c_2\varepsilon - \|c\|_1)^2 = (c_2\varepsilon)^2$$

On the other hand, we have $\operatorname{dist}_2(\bar{z}, \mathcal{S}^\star) = \varepsilon$. Gathering up, we obtain that

$$\alpha_G = \inf_{z \in \mathcal{D}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} \leq \frac{\operatorname{dist}_2(0, \mathcal{F}(\bar{z}))}{\operatorname{dist}_2(\bar{z}, \mathcal{S}^\star)} = \min\{c_1, c_2\}\,.$$

Dependence on τ. For any $\varepsilon > 0$ let $\tau_\varepsilon = \tau - \varepsilon$ and $\bar{z}^\varepsilon = (\bar{x}, \bar{y}) \in \mathbf{R}^{n+m}$ such that $\bar{x}^\varepsilon = x^\star + \tau_\varepsilon c$ and $\bar{y} = 0$. Note that $\bar{z} \in \mathcal{D}$. Moreover, in view of (50), we have $I = \{1, 2, 3\}$. Hence,

$$\operatorname{dist}_2^2(0, \mathcal{F}(\bar{z})) = \|c\|_2^2 = 1\,.$$

Furthermore, we have

$$\operatorname{dist}_2^2(\bar{z}, \mathcal{S}^\star) = \operatorname{dist}_2^2(\bar{x}, \mathcal{S}_x^\star) + \operatorname{dist}_2^2(\bar{y}, \mathcal{S}_y^\star) = \left(\frac{\langle c, \bar{x}\rangle - \|c\|_1}{\|c\|_2}\right)^2 + 1 = 1 + \tau_\varepsilon^2\,.$$

Gathering up, we obtain that $\alpha_G = O(\tau^{-1})$:

$$\alpha_G = \inf_{z \in \mathcal{D}} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} \leq \frac{\operatorname{dist}_2(0, \mathcal{F}(\bar{z}))}{\operatorname{dist}_2(\bar{z}, \mathcal{S}^\star)} = \frac{1}{\sqrt{1 + \tau_\varepsilon^2}} \xrightarrow{\varepsilon \to 0} \frac{1}{\sqrt{1 + \tau^2}}\,.$$

Estimation of α_L. Invoking Proposition C.12 we obtain

$$\mathcal{S}_L^\star = \mathcal{S}_{z^\star, x} \times \mathcal{S}_{z^\star, y} \quad \text{where} \quad \mathcal{S}_{z^\star, x} = \{x \in \mathbf{R}^2 : \langle c, x\rangle = \|c\|_1\} \quad \text{and} \quad \mathcal{S}_{z^\star, y} = \{(1, 0, 0)\}\,.$$

Hence, following the analysis of α_G, we derive $\alpha_L \leq 1/\sqrt{1 + \tau^2}$. In particular, when $\tau = 2$ we obtain $\alpha_L \leq 0.44$ However, we can not follow our previous analysis for α_G to obtain $\alpha_L \leq \min\{c_1, c_2\}$, as the $\bar{z}$ chosen in that case satisfies

$$\operatorname{dist}_2^2(\bar{z}, \mathcal{S}_L^\star) = \left(\frac{\langle c, \bar{x}\rangle - \|c\|_1}{\|c\|_2}\right)^2 = ((c_2 + c_1) + c_2\varepsilon - \|c\|_1)^2 = (c_2\varepsilon)^2 = \operatorname{dist}_2^2(0, \mathcal{F}(\bar{z}))\,.$$

Nevertheless, in the following we will show that $\alpha_L \gg \min\{c_1, c_2\}$ when $\min\{c_1, c_2\}$ is small.

Numerical lower bound for α_L. To check that $\alpha_L \gg \min\{c_1, c_2\}$ when $\min\{c_1, c_2\}$ is small, let

$$G = \{z \in \mathbf{R}^{n+m} : Hz \leq 0\}, \quad \text{where} \quad H = \begin{bmatrix} A_1 & 0_{1\times 3} \\ 0_{2\times 2} & A^\top \\ 0_{2\times 2} & -A^\top \\ 0_{2\times 2} & I_3 \\ \frac{1}{R}c^\top & \frac{1}{R}b^\top \end{bmatrix} \quad \text{and} \quad I_3 = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

From [43, Lemma 3 (c)] and the invariance to translations of the sharpness constants, for any $\underline{z}^\star \in \mathcal{S}^\star$ we have

$$\inf_{z \in \overline{\mathbf{B}}_\tau(\underline{z}^\star)} \frac{\operatorname{dist}_2(0, \mathcal{F}(z))}{\operatorname{dist}_2(z, \mathcal{S}^\star)} \geq \inf_{z \in \mathbf{R}^{n+m}} \frac{\operatorname{dist}_2(Hz, \mathbf{R}^m_-)}{\operatorname{dist}_2(z, G)} =: \mathcal{H}\,,$$

where $\mathcal{H}$ is the *sharpness constant* (the inverse of the *Hoffman constant*) of the homogeneous linear system $Hz \leq 0$. Then, taking into account that $\mathcal{D} = \mathcal{S}^\star + \tau\mathbf{B}$, we obtain

$$\alpha_L = \inf_{z\in\mathcal{D}} \frac{\operatorname{dist}_2(0,\mathcal{F}(z))}{\operatorname{dist}_2(z,\mathcal{S}^\star)} \geq \inf_{\underline{z}^\star\in\mathcal{S}^\star} \inf_{z\in\overline{\mathbf{B}}_\tau(\underline{z}^\star)} \frac{\operatorname{dist}_2(0,\mathcal{F}(z))}{\operatorname{dist}_2(z,\mathcal{S}^\star)} \geq \mathcal{H}\,.$$

Furthermore, using the procedure proposed by Peña [48], we can numerically lower bound $\mathcal{H}$. This way, when $\min\{c_1,c_2\} = 10^{-7}$ we obtain numerically $\alpha_L \geq \mathcal{H} \geq 0.036$.

Auxiliary results. We prove two auxiliary results used in the derivation of Example 4.7. The first gives an exact form for the distance to zero of the saddle subdifferential $\mathcal{F}$ of the minimax problem (2). The second relates the set $\mathcal{S}_L^\star$ defined in (15) with the solution set of a reduced optimization problem.

Proposition C.11. *For any $z = (x,y) \in \mathbf{R}^{n+m}$, the distance to zero of the saddle subdifferential of the minimax problem (2) has the following form*

$$\operatorname{dist}_2^2(0,\mathcal{F}(z)) = \begin{cases} \operatorname{dist}_2^2(0,\partial f(x) + J_G^\top(x)y) + \|(G(x)_I)_+\|_2^2 + \|G(x)_{[m]\setminus I}\|_2^2 & \text{if } y \geq 0 \\ \infty & \text{otherwise} \end{cases}.$$

where $I = \{j \in [m] : y_j = 0\}$ and $J_G(x)_j = \partial g_j(x)$ for all $j \in [m]$.

Proof. Let $z = (x,y) \in \mathbf{R}^{n+m}$. Recall that

$$\mathcal{F}(x,y) = \begin{bmatrix} \partial_x\mathcal{L}(x,y), \\ \partial_y(-\mathcal{L})(x,y) \end{bmatrix} \quad \text{where} \quad \mathcal{L}(x,y) = f(x) + \langle y, G(x)\rangle - \iota_{\mathbf{R}_+^m}(y)\,.$$

By subdifferential calculus, we obtain

$$\partial_x\mathcal{L}(x,y) = \partial f(x) + J_G^\top(x)y \quad \text{and} \quad \partial_y(-\mathcal{L})(x,y) = \begin{cases} -G(x) + N_{\mathbf{R}_+^m}(y) & \text{if } y \geq 0 \\ \emptyset & \text{otherwise} \end{cases}.$$

Therefore, if $y_j < 0$ for some $j \in [m]$ then $\mathcal{F}(z) = \emptyset$, whence $\operatorname{dist}_2(0,\mathcal{F}(z)) = \infty$. Then, suppose $y \geq 0$ and fix $j \in [m]$. Note that $N_{\mathbf{R}_+^m(y)} = \bigtimes_{j=1}^m N_{\mathbf{R}_+(y_j)}$. Moreover,

$$\min_{t\in N_{\mathbf{R}_+}(y_j)} |G(x)_j - t|^2 = \begin{cases} \max\{0, G(x)_j\}^2 & \text{if } y_j = 0 \\ |G(x)_j|^2 & \text{otherwise} \end{cases}.$$

It follows that

$$\begin{aligned} \operatorname{dist}_2^2(0,\partial_y(-\mathcal{L})(x,y)) &= \min_{z\in -N_{\mathbf{R}_+^m}(y)} \sum_{j=1}^m |G(x)_j + z_j|^2 \\ &= \sum_{j=1}^m \min_{z_j\in -N_{\mathbf{R}_+}(y_j)} |G(x)_j + z_j|^2 \\ &= \|(G(x)_I)_+\|_2^2 + \|G(x)_{[m]\setminus I}\|_2^2\,. \end{aligned}$$

Gathering up, we obtain

$$\begin{aligned} \operatorname{dist}_2^2(0,\mathcal{F}(z)) &= \operatorname{dist}_2^2(0,\partial_x\mathcal{L}(x,y)) + \operatorname{dist}_2^2(0,\partial_y(-\mathcal{L})(x,y)) \\ &= \operatorname{dist}_2^2(0,\partial f(x) + J_G^\top(x)y) + \|(G(x)_I)_+\|_2^2 + \|G(x)_{[m]\setminus I}\|_2^2\,. \end{aligned}$$

□

Consider the simplified primal problem that only incorporates the constraints indexed by B,

$$p_B^\star = \begin{cases} \min\limits_{x \in \mathbf{R}^n} & f(x) \\ \text{s.t.} & g_j(x) \leq 0 \quad \text{for all } j \in B\,. \end{cases} \tag{51}$$

Denote $\mathcal{S}_B^\star \subseteq \mathbf{R}^{n+q}$, where $q = |B|$, as the set of primal-dual solutions to the minimax problem associated to (51)

$$\min_{x \in \mathbf{R}^n} \max_{y \in \mathbf{R}_+^q} f(x) + \langle y, G_B(x) \rangle\,.$$

The following proposition establishes that the set $\mathcal{S}_L^\star$ is equal to $\mathcal{S}_B^\star \times \{0_N\}$ up to dual coordinates reordering.

Proposition C.12. *Assume that $B = \{1, \dots, q\}$. Then, $\mathcal{S}_L^\star = \mathcal{S}_B^\star \times \{0_N\}$.*

Proof. Let $\bar{\mathcal{S}}_B \subseteq \mathbf{R}^{n+m}$ be the solution to the following system

$$f(x) - h_B(y) \leq 0, \quad G_B(x) \leq 0, \quad y_B \geq 0 \quad \text{and} \quad y_N = 0\,,$$

where $h_B : \mathbf{R}_+^m \to \mathbf{R} \cup \{+\infty\}$ is such that $h_B(y) = \min_{x \in \mathbf{R}^n} f(x) + \sum_{j \in B} y_j g_j(x)$ is, modulo y_N, the dual function of the reduced problem (51). By construction, $\bar{S}_B = \mathcal{S}_B^\star \times \{0_N\}$. Let $z = (x, y) \in \mathcal{S}_L^\star$. If $z \in \overline{\mathbf{B}}_{\delta_A/2}^P(z^\star)$, from Propositions C.10 and C.3 we obtain $z \in \mathcal{S}^\star$ and $G_N(x) < 0$, respectively. By complementary slackness, $y_N = 0$. If $z \notin \overline{\mathbf{B}}_{\delta_A/2}^P(z^\star)$, let $\theta = \delta_A/(2\|z - z^\star\|_P) \in (0, 1)$ and

$$\bar{z} = (\bar{x}, \bar{y}) := (1 - \theta) z^\star + \theta z = z^\star + \frac{\delta_A}{2} \frac{z - z^\star}{\|z - z^\star\|_P} \in \overline{\mathbf{B}}_{\delta_A}^P(z^\star)\,.$$

Since $\mathcal{S}_L^\star$ is convex, $\bar{z} \in \mathcal{S}_L^\star$. Therefore, as argued in the first case, $\bar{y}_N = 0$. Hence, since $\theta \neq 0$, $y_N = \theta^{-1}(\bar{y}_N - (1 - \theta) y_N^\star) = 0$. Furthermore,

$$h_B(y) = \min_{x \in \mathbf{R}^n} f(x) + \sum_{j \in B} y_j g_j(x) = \min_{x \in \mathbf{R}^n} f(x) + \sum_{j \in [m]} y_j g_j(x) = h(y)\,.$$

Then, $z \in \bar{\mathcal{S}}_B$, whence $\mathcal{S}_L^\star \subseteq \bar{\mathcal{S}}_B$. On the other hand, any $z = (x, y) \in \bar{\mathcal{S}}_B$ satisfies $y_N = 0$, whence $h(y) = h_B(y)$ and then $z \in \mathcal{S}_L^\star$. Therefore $\bar{\mathcal{S}}_B \subseteq \mathcal{S}_L^\star$. □