Global convergence and acceleration of fixed point iterations of union upper semicontinuous operators: proximal algorithms, alternating and averaged nonconvex projections, and linear complementarity problems

Jan Harold Alcantara∗ Ching-pei Lee†

July 19, 2022

Abstract

We propose a unified framework to analyze fixed point iterations of a set-valued operator that is the union of a finite number of upper semicontinuous maps, each with a nonempty closed domain and compact values. We discuss global convergence, local linear convergence under a calmness condition, and component identification, and further propose acceleration strategies that drastically improve the convergence speed. Our framework is applied to analyze a class of proximal algorithms for minimizing the sum of a piecewise smooth function and the difference between pointwise minimum of finitely many weakly convex functions and a piecewise smooth convex function. When realized on two-set feasibility problems, this algorithm class recovers alternating projections and averaged projections as special cases, and our framework thus equips these classical methods with global convergence and possibilities for acceleration on a broad class of nonconvex feasibility problems. By specializing the framework to a nonconvex feasibility problem reformulation of the linear complementarity problem, we show global convergence to a solution from any initial point, with a local linear rate, of the alternating projection as well as the averaged projection methods, which is difficult to obtain on nonconvex problems. Numerical results further exemplify that the proposed acceleration algorithms significantly improve upon their non-accelerated counterparts in efficiency.

Keywords. fixed point algorithm; proximal methods; alternating projections; averaged projections; linear complementarity problem; Nonconvex optimization; Global convergence

1 Introduction

Consider the fixed point problem

\[ \text{find } w \in T(w), \quad (1.1) \]

where \( T : \mathbb{E} \rightrightarrows \mathbb{E} \) is a set-valued operator on a Euclidean space \( \mathbb{E} \), with its solution set denoted as

\[ \text{Fix}(T) := \{ w \in \mathbb{E} : w \in T(w) \}. \]

A standard fixed point algorithm for solving (1.1) is given by the iterations

\[ w^{k+1} \in T(w^k), \quad k \geq 0, \quad (\text{FPA}) \]

for some initial point \( w^0 \in \mathbb{E} \). When \( T \) is a (single-valued) continuous operator on \( \mathbb{E} \), the limit of any convergent sequence generated by (FPA) is necessarily a fixed point of \( T \). However, this need not be the case for general set-valued operators \( T \).

∗janharold@stat.sinica.edu.tw Academia Sinica.
†leechingpei@gmail.com Academia Sinica.
In this paper, we investigate the global convergence of the fixed point iterations \([FPA]\) to a point in \(\text{Fix}(T)\) when \(T\) is a union upper semicontinuous map. That is, \(T\) can be expressed as

\[
T(w) = \bigcup_{i \in \mathcal{I}, w \in D_i} T_i(w),
\]

(1.2)

and the following assumptions hold.

**Assumption 1.1.** The index set \(\mathcal{I}\) is finite. For each \(i \in \mathcal{I}, D_i\) is nonempty and closed, \(T_i : D_i \rightarrow \mathbb{E}\) is an upper semicontinuous set-valued map on \(D_i\) (see Definition B.1), and \(T_i(w)\) is nonempty and compact for any \(w \in D_i\). Moreover, \(\mathbb{E} = \bigcup_{i \in \mathcal{I}} D_i\).

We sometimes call these \(T_i\) individual or component operators of \(T\).

The main motivation of this work stems from classical projection algorithms for a nonconvex two-set feasibility formulation of the linear complementarity problem (LCP). The LCP involves finding a point \(x \in \mathbb{R}^n\) that satisfies

\[
x \geq 0, \quad Mx - b \geq 0, \quad \text{and} \quad \langle x, Mx - b \rangle = 0,
\]

where \(M \in \mathbb{R}^{n \times n}\) and \(b \in \mathbb{R}^n\). Introducing a variable \(y := Mx - b\) and letting \(w := (x, y)\), the LCP (1.3) is equivalent to the following nonconvex feasibility problem:

\[
\text{find } w \in S_1 \cap S_2,
\]

(FP)

with

\[
S_1 := \{w \in \mathbb{R}^{2n} : Aw = b\} \quad \text{with} \quad A := [M \ - \mathbb{I}] \in \mathbb{R}^{n \times 2n},
S_2 := \{w \in \mathbb{R}^{2n} : w_j \geq 0, \ w_{n+j} \geq 0, \ w_jw_{n+j} = 0, \ \forall j \in [n]\}.
\]

(1.4)

Well-known projection algorithms for solving (FP) such as the method of alternating projections (MAP) and the method of averaged projections (MAveP), are expressible in the form (FPA) satisfying Assumption 1.1 when each \(S_i\) \((i = 1, 2)\) is a finite union of nonempty closed and convex sets; see Example 2.1. For this type of nonconvex feasibility problems, only local convergence results for MAP and MAveP are currently known [Dao and Tam 2019, Drusvyatskiy and Lewis 2019, Lewis et al. 2009].

Motivated to obtain global convergence results for MAP and MAveP on the nonconvex problem (FP), we go through layered generalization of such algorithms, first to a proximal difference-of-min-weakly-convex optimization framework that generalizes the proximal difference-of-convex algorithm of [Wen et al. 2018], and then to the fixed point iterations (FPA) of union upper semicontinuous operators. We develop a Lyapunov-type analysis for (FPA) to prove its global convergence to a fixed point. Moreover, we also obtain fruitful results in our theory development, including sufficient conditions for local linear convergence of (FPA) and guarantees for a novel component identification result of \(T_i\) that is in spirit relevant to manifold identification of [Hare and Lewis 2004]. Through the identification of the component, we further propose two novel strategies for acceleration, one through same-component extrapolations with monotone objective decrease that is different from the widely-accepted nonmonotone Nesterov acceleration, and the other through solving a component-restricted fixed point problem.

These results then apply to all the problems from which we generalize to (FPA). In particular, for the motivating example of nonconvex feasibility LCP problem, we use our framework to show that the classical MAP and MAveP are able to generate iterates that globally converge to an optimal solution when the matrix \(M\) is a \(P\)-matrix (see Definition 5.2 or Cottle et al. 1992). This is a rare result for projection methods on nonconvex sets, for which usually only local convergence can be proven in the literature [Dao and Tam 2019, Drusvyatskiy and Lewis 2019, Lewis et al. 2009]. In addition, global convergence to an optimal solution is also obtained for MAP/MAveP accelerated via same-component extrapolation and component restriction described above.

Our framework and analysis is versatile such that a broad range of problems and algorithms are covered and new algorithms are also derived. We especially mention the following prominent examples.
1. **Proximal difference-of-min-convex algorithms.** Our framework covers a family of proximal-subgradient-like algorithm for a class of nonsmooth and nonconvex optimization problems of the form

\[
\min_{w \in \mathbb{R}} f(w) + g(w) - h(w),
\]

where \( f \) is piecewise smooth, \( g \) is the pointwise minimum of a finite number of weakly convex functions \( g_j \) (meaning that each \( g_j \) satisfies that \( g_j + \rho_j \| \cdot \|_2^2 \) is convex for some \( \rho_j \geq 0 \)), and \( h \) is convex and piecewise smooth. Problem (OP) generalizes the difference-of-convex problem (such that \( f \) is convex and smooth and both \( g \) and \( h \) are convex) of [Dao and Tam 2019], and our flexible framework allows for more possibilities in algorithmic design, and establishes global convergence as well as local linear convergence. See Section 3.

2. **MAP and MAveP for the two-set feasibility problem (FP) with each nonconvex set being a union of finitely many convex set.** When each \( S_i \) (\( i = 1, 2 \)) is a finite union of nonempty closed and convex sets, Example 2.1 shows how MAP and MAveP fit into (FPA) with Assumption 1.1. Such sets are referred to as union convex sets in the recent work of [Dao and Tam 2019].

3. **Globally convergent projection methods for LCP.** MAP and MAveP for the nonconvex feasibility formulation of LCP introduced above in (1.4) and (FP) fit in our general framework of (FPA) since \( S_1 \) is convex and \( S_2 \) is union-convex in (1.4). Other reformulations of the LCP as an optimization problem (OP) and corresponding new proximal algorithms are discussed in Section 5.2.

4. **Projection methods for the sparse affine feasibility problem (SAFP) and sparsity-constrained optimization problems.** The former is another example of (FP) with union convex sets, where \( S_1 \) is a linear-equality constraint set, while \( S_2 \) is an \( \ell_0 \)-norm constraint set that can be written as the union of finitely many subspaces. This setting can be naturally extended to proximal-gradient-type algorithms for sparsity-constrained optimization problems of the form (OP) with \( g \) being the indicator function on the sparsity set \( S_2 \) and \( h \equiv 0 \), and thus still falls within our framework. The details are covered in Sections 3.1 and 5.1.

In comparison to existing works, our Assumption 1.1 focuses on a special class of fixed point algorithms with specific structures that are still general enough to cover many widely-used methods in various contexts such as those enumerated above, but can allow for more possibilities of acceleration than general fixed point algorithms. For (OP), our setting and analysis generalize that of [Dao and Tam 2019], which assumed that \( f \) is convex and smooth, \( g \) is a min-convex function (minimum of finitely many convex functions), and \( h \equiv 0 \). In contrast, our \( f \) could be nonconvex and nonsmooth, our \( g \) is the minimum taken over weakly convex functions, and we have an additional \( h \) term. In comparison with the proximal difference-of-convex algorithm of [Wen et al. 2018] that requires \( f \) to be convex and smooth and \( g \) to be convex, neither condition on \( f \) needs to hold in our framework and our \( g \) can be nonconvex, while we still manage to obtain global and local convergence guarantees through a more involved analysis. Our acceleration schemes with strict function decrease are also different from their nonmonotone extrapolation. By reformulations of (FP) involving union convex sets to the form (OP), we provide sufficient conditions for global convergence of MAP and MAveP to candidate solutions. For the special case of SAFP, we provide conditions weaker than those in existing literature [Beck and Teboulle 2011, Hesse et al. 2014] for guaranteeing global subsequential convergence to a stationary point, theoretically sound results on sparsity pattern identification within finite iterations, and acceleration strategies. On the other hand, we also prove global convergence to an optimal solution of the LCP feasibility reformulation of multiple projection algorithms, including not only the classical MAP and MAveP but also algorithms new to the LCP literature that we derive from our general framework. We note that for nonconvex feasibility problems, global convergence guarantees for MAP are still an open problem (for example, the seminal work [Lewis et al. 2009] provides only local convergence), and our result has shown that for our feasibility reformulation of such LCPs, iterates of MAP converge globally to the solution, without resorting to the popular Kurdyka-Lojasiewicz condition widely used in analyzing nonconvex optimization algorithms. We hope our techniques can inspire future study of global convergence.
to an optimal solution in nonconvex optimization using a different path. Accelerated algorithms for both
SAFP and LCP are also derived from our general framework, and numerical experiments demonstrate the
applicability of our proposed algorithms, as well as the significant improvement achieved by the proposed
acceleration techniques.

1.1 Organization of the paper

This paper is organized as follows. In Section 2, we propose an acceleration scheme for (FPA) and prove
the global convergence of both (FPA) and the accelerated version to a solution of (1.1) using the notion of
Lyapunov functions. The developed general theory is further applied in Section 3 to show global convergence
of proximal algorithms for (OP). We also establish relations of the fixed point set to the solutions and
critical points of (OP). Section 4 presents realizations of such proximal algorithms on the two-set feasibility
problem (FP) where the involved sets are union convex, and the special cases of SAFP and the feasibility
reformulation of LCP are tackled in Section 5. We then discuss related works in Section 6. Experiments in
Section 7 further provide numerical support for our theory and demonstrate the effectivness of the proposed
acceleration approach. Finally, concluding remarks are given in Section 8.

1.2 Notations

Throughout this paper, \( \mathbb{E} \) is a Euclidean space endowed with the inner product \( \langle \cdot, \cdot \rangle \) and the induced norm \( \| \cdot \| \). The distance function on a nonempty and closed set \( S \subseteq \mathbb{E} \) is denoted and given by \( \text{dist}(w, S) := \min_{z \in S} \| w - z \| \). We denote by \( B(S, \epsilon) := \{ z \in \mathbb{E} : \text{dist}(w, S) < \epsilon \} \) the open ball around \( S \) with radius \( \epsilon > 0 \). The projection operator onto \( S \), \( P_S : \mathbb{E} \rightharpoonup S \), is defined by \( P_S(w) := \arg \min_{z \in S} \| w - z \| \). The indicator function of \( S \) is defined by

\[
\delta_S(w) = \begin{cases} 
0 & \text{if } w \in S, \\
+\infty & \text{otherwise},
\end{cases}
\]

and \( \text{conv}(S) \) is the convex hull of \( S \). When \( T : \mathbb{E} \rightharpoonup \mathbb{E} \) is single-valued at \( w \in \mathbb{E} \), say \( T(w) = \{ z \} \), we slightly abuse the notation and write \( T(w) = z \). The identity operator on \( \mathbb{E} \) will be denoted by \( \text{Id} \), while the identity matrix in \( \mathbb{R}^n \) is denoted by \( \mathbb{I} \). For a finite collection of sets \( \mathcal{D} := \{ D_i \subseteq \mathbb{E} : i \in \mathcal{I} \} \), we define the set-valued function \( \phi_\mathcal{D} : \bigcup_{i \in \mathcal{I}} D_i \rightharpoonup 2^\mathbb{I} \) by

\[
\phi_\mathcal{D}(w) := \{ i : D_i \in D, w \in D_i \}. \tag{1.5}
\]

Given \( g : \mathbb{E} \to \mathbb{R} \cup \{ +\infty \} \), we denote by \( \text{dom}(g) = \{ w \in \mathbb{E} : g(w) < +\infty \} \) its domain. Further given \( \lambda > 0 \), the proximal mapping of \( g \), denoted by \( \text{prox}_\lambda g : \mathbb{E} \rightharpoonup \mathbb{E} \), is defined by

\[
\text{prox}_\lambda g(w) = \arg \min_{z \in \mathbb{E}} \left\{ g(z) + \frac{1}{2\lambda} \| z - w \|^2 \right\}, \tag{1.6}
\]

and the Moreau envelope of \( g \) is defined by

\[
M_\lambda g(w) = \min_{z \in \mathbb{E}} \left\{ g(z) + \frac{1}{2\lambda} \| z - w \|^2 \right\}. \tag{1.7}
\]

For \( S \subseteq \mathbb{E} \), we define \( \text{prox}_\lambda g(S) := \bigcup_{w \in S} \text{prox}_\lambda g(w) \). If there exists a finite family of functions \( \{ g_j : j \in J \} \), where \( g_j : \mathbb{E} \to \mathbb{R} \cup \{ +\infty \} \) for all \( j \in J \), such that for any \( w \in \mathbb{E} \), we have \( g(w) \in \{ g_j(w) : j \in J' \} \) for some \( J' \subseteq J \), we denote

\[
D_j(g) := \{ w \in \text{dom}(g) : g(w) = g_j(w) \}. \tag{1.8}
\]

2 A general framework

This section develops global convergence of (FPA) to solutions of the fixed point problem (1.1) and proposes
a general acceleration strategy. Proofs of results in this section are put in Appendix B.
As we will see in Section 3, there are many algorithms that take the form (FPA) satisfying Assumption 1.1. As our motivating example, we look at two classical projection algorithms for the feasibility problem (FP) where each constraint set is a union convex set.

**Example 2.1** (Projection methods for feasibility problems). Let \( S_1, S_2 \subseteq \mathbb{E} \). Two well-known algorithms for solving the corresponding feasibility problem (FP) are the method of alternating projections (MAP) and the method of averaged projections (MAveP), which take the form (FPA) respectively with \( T = P_{S_2} \circ P_{S_1} \) and \( T = \frac{1}{2}(P_{S_1} + P_{S_2}) \) [see, for example, Lewis et al. 2000, Dao and Tam 2019]. Each of these algorithms is aimed at finding fixed points of its defining operator \( T \), which are candidate solutions to the feasibility problem as \( S_1 \cap S_2 \subseteq \text{Fix}(T) \) for each operator \( T \) defined above.

When \( S_1 \) and \( S_2 \) are union convex sets, the above operators conform with Assumption 1.1. To see this, let \( S_1 = \bigcup_{j \in J_1} R_j^{(1)} \) and \( S_2 = \bigcup_{l \in J_2} R_l^{(2)} \) where \( J_1, J_2 \) are finite index sets and \( R_j^{(1)}, R_l^{(2)} \) are closed convex sets for all \( j \in J_1 \) and \( l \in J_2 \). Then

\[
P_{S_i}(w) = \bigcup_{j: w \in C_j^{(i)}} P_{R_j^{(i)}}(w), \quad i = 1, 2,
\]

where

\[
C_j^{(i)} := \{ w \in \mathbb{E} : \text{dist}(w, S_i) = \text{dist}(w, R_j^{(i)}) \}, \quad j \in J_i.
\]

By the convexity of \( R_j^{(i)} \) and the continuity of the distance function, \( C_j^{(i)} \) is closed, so each

\[
D_{j,l} := \{ w \in \mathbb{E} : w \in C_j^{(1)} \text{ and } P_{R_j^{(1)}}(w) \in C_l^{(2)} \}
\]

is also closed. Therefore, the alternating projection map can be expressed as (1.2) by

\[
(P_{S_2} \circ P_{S_1})(w) = \bigcup_{j,l: w \in D_{j,l}} \left( P_{R_l^{(2)}} \circ P_{R_j^{(1)}} \right)(w),
\]

and Assumption 1.1 holds. Similarly, the averaged projections map is given by

\[
\frac{1}{2}(P_{S_1} + P_{S_2})(w) = \bigcup_{j,l: w \in C_j^{(1)} \cap C_l^{(2)}} \frac{1}{2} \left( P_{R_j^{(1)}} + P_{R_l^{(2)}} \right)(w).
\]

### 2.1 Convergence theorems

Our global convergence analysis depends on the existence of a Lyapunov function associated with the operator \( T \), which we define as follows.

**Definition 2.1.** A function \( V : \mathbb{E} \to \mathbb{R} \cup \{ \pm \infty \} \) continuous on its domain is said to be a Lyapunov function for \( T \) if \( \inf V > -\infty \) and

\[
\sup_{w^+ \in \text{Fix}(T)} V(w^+) \leq V(w) \text{ for any } w \in \mathbb{E},
\]

with equality if and only if \( w \in \text{Fix}(T) \).

We establish in the following result that existence of a Lyapunov function \( V \) guarantees that accumulation points of a sequence produced by (FPA) if they exist, are solutions of (1.1). We emphasize that both the finiteness of \( \mathcal{I} \) and the closedness of each \( D_i \) are essential in the following theorem.

**Theorem 2.1** (Global subsequential convergence). Suppose that Assumption 1.1 holds and there exists a Lyapunov function for \( T \). Let \( \{w^k\} \) be a sequence generated by (FPA) then any accumulation point of \( \{w^k\} \) is a fixed point of \( T \).
We also propose an accelerated version of the fixed point algorithm (FPA) in Algorithm 1. With the sufficient decrease condition \(2.2\) for the stepsize of the extrapolation step, we also obtain global subsequential convergence for any sequence generated by Algorithm 1.

**Algorithm 1: Accelerated fixed point algorithm.**

Let \( V \) be a Lyapunov function for \( T \).
Choose \( \sigma > 0 \) and \( w_0 \in E \). Set \( \gamma = w_0 \) and \( k = 0 \).

**Step 1.** Set \( z^k = w^k + t_k p^k \), where \( p^k = w^k - w^{k-1} \) and \( t_k \geq 0 \) such that
\[
V(z^k) \leq V(w^k) - \frac{\sigma}{2} t_k^2 \|p^k\|^2.
\] (2.2)

**Step 2.** Set \( w^{k+1} \in T(z^k) \), \( k = k + 1 \), and go back to Step 1.

**Theorem 2.2** (Global subsequential convergence of Algorithm 1). Under the hypotheses of Theorem 2.1 and single-valuedness of \( T \), we obtain global subsequential convergence for any accumulation point of a sequence generated by Algorithm 1.

With additional information on the operators \( T_i \), we establish in Theorem 2.4 the global convergence of the full sequence generated by (FPA) to a point in \( \text{Fix}(T) \). To this end, the following lemma for component identification is needed.

**Lemma 2.3.** Let \( D = \{D_i : i \in I\} \) be any finite collection of closed sets in \( E \) and denote \( U := \bigcup_{i \in I} D_i \).
Then for any \( w^* \in U \), there exists \( \delta > 0 \) such that \( \phi_D(w) \subseteq \phi_D(w^*) \) for all \( w \in B(w^*, \delta) \cap U \), where \( \phi_D \) is defined by \(1.5\).

**Theorem 2.4.** Under the hypotheses of Theorem 2.1, let \( \{w^k\} \) be a sequence generated by (FPA) with an accumulation point \( w^* \), denote \( D := \{D_i : i \in I\} \), and suppose for each \( i \in \phi_D(w^*) \), \( T_i \) is calm at \( w^* \) with parameter \( \kappa_i \in [0,1] \) (see Rockafellar and Wets 1998 Chapter 9), that is,
\[
T_i(w) \subseteq T_i(w^*) + \kappa_i \|w - w^*\| B(0,1)
\] (2.3)
for all \( w \in D_i \) sufficiently close to \( w^* \). If a Lyapunov function for \( T \) exists and \( T \) is single-valued at \( w^* \), then \( w^k \to w^* \) and \( \phi_D(w^k) \subseteq \phi_D(w^*) \) for all sufficiently large \( k \). Moreover, the rate of convergence is locally linear if \( \kappa_i < 1 \) for all \( i \in I \).

**Remark 2.1.** Nonexpansive maps are calm everywhere, and projections onto convex sets are nonexpansive, and thus the individual operators of the alternating and averaged projection maps in Example 2.1 satisfy (2.3) (A single-valued operator \( T \) is nonexpansive if there exists \( \kappa \in [0,1] \) such that \( \|T(w) - T(z)\| \leq \kappa \|w - z\| \) for all \( w, z \in E \), and is a contraction if \( \kappa < 1 \).)

**Remark 2.2** (Component identification). If \( D \subseteq E \) is any closed set such that \( D \cap \text{Fix}(T) = \emptyset \) and if \( \{w^k\} \) generated by (FPA) is bounded, then \( D \) contains at most finitely many terms of \( \{w^k\} \) by Theorem 2.1. Thus, only those \( D_i \) which contains a fixed point of \( T \) can possibly contain infinitely many terms of \( \{w^k\} \). Moreover, the conclusion of Theorem 2.4 that \( \phi_D(w^k) \subseteq \phi_D(w^*) \) for all large \( k \) allows us to identify which operator \( T_i \) will yield the fixed point \( w^* \) of \( T \). In particular, this result implies we may choose an index \( i \in \phi_D(w^k) \) and reduce the fixed point problem (1.1) to finding a fixed point of \( T_i \). Provided that \( k \) is large enough, we have from Theorem 2.4 that a fixed point of \( T_i \) corresponds to a solution of the original problem (1.1).

In the next sections, we utilize our theory developed in this section to analyze algorithms for various problem classes by constructing suitable Lyapunov functions and showing boundedness of the iterates to guarantee the existence of accumulation points. Conditions for ensuring calmness of the component operators and single-valuedness of \( T \) on \( \text{Fix}(T) \) will also be provided to obtain full convergence and local rates.
3 Applications to min-$\rho$-convex optimization

In this section, we focus on \([\text{OP}]\) with \(f, g\) and \(h\) possibly nonconvex, and \(f + g - h\) bounded from below. We consider regularizers \(g\) that belong to the class of min-$\rho$-convex functions defined below. Recall that a function \(F\) is said to be $\rho$-convex if \(F(w) - \frac{\rho}{2}\|w\|^2\) is a convex function. \(F\) is said to be weakly convex if \(\rho < 0\), convex if \(\rho \geq 0\) and strongly convex if \(\rho > 0\).

**Definition 3.1** (min-$\rho$-convex function). We say that \(g : E \to \mathbb{R} \cup \{+\infty\}\) is a min-$\rho$-convex function if there exist an index set \(J\) with \(|J| < \infty\), and $\rho$-convex, proper, and lower semicontinuous functions \(g_j : E \to \mathbb{R} \cup \{+\infty\}, j \in J\), such that

\[
g(w) = \min_{j \in J} g_j(w), \quad \forall w \in E.
\]

We call \(g\) min-weakly convex if \(\rho < 0\), min-convex if \(\rho \geq 0\), and min-strongly convex if \(\rho > 0\).

A traditional setting in the study of algorithms for solving \([\text{OP}]\) is that \(f\) has a Lipschitz continuous gradient, and \(g\) and \(h\) are convex functions \([\text{Wen et al., 2018}]\). Using the theory we developed in Section 2, we can establish global convergence for algorithms for solving a wider class of problems that subsumes such a setting. More precisely, we will work under the following assumptions.

**Assumption 3.1.**

(a) The functions \(f, g\) and \(h\) are expressible as

\[
f = \min_{i \in I} f_i, \quad g = \min_{j \in J} g_j, \quad \text{and} \quad h = \max_{m \in M} h_m,
\]

where \(I, J\) and \(M\) are finite index sets.

(b) For each \(i \in I\), \(f_i\) has \(L_i\)-Lipschitz continuous gradient on \(E\) for some \(L_i > 0\). That is,

\[
\|\nabla f_i(w) - \nabla f_i(z)\| \leq L_i\|w - z\|, \quad \forall w, z \in E. \tag{3.1}
\]

(c) For each \(j \in J\), \(\text{dom}(g_j)\) is closed, and \(g_j\) is a proper and $\rho$-convex function continuous on \(\text{dom}(g_j)\).

(d) For each \(m \in M\), \(h_m\) is a continuously differentiable convex function on \(E\).

(e) For all \((i, j, m) \in I \times J \times M\), the function \(f_i + g_j - h_m\) is coercive over \(E\).

**Remark 3.1.** We now mention some consequences of the above assumptions.

(a) By **Assumption 3.1** (b), we have from the descent lemma [for example, **Beck [2017]** Lemma 5.7] that

\[
f_i(z) \leq f_i(w) + \langle \nabla f_i(w), z - w \rangle + \frac{L_i}{2}\|z - w\|^2, \quad \forall w, z \in E. \tag{3.2}
\]

(b) With **Assumption 3.1** (b)-(d), the sets \(D_i(f), D_j(g)\) and \(D_m(h)\) defined as in \([1.8]\) are closed for any \((i, j, m) \in I \times J \times M\). Hence, by **Assumption 3.1** (a), we may write \(\text{dom}(g)\) as the union of a finite number of closed sets:

\[
\text{dom}(g) = \bigcup_{(i, j, m) \in I \times J \times M} D_i(f) \cap D_j(g) \cap D_m(h) \tag{3.3}
\]

(c) From **Assumption 3.1** (c), \(g_j(z) + \frac{1}{\lambda\rho}\|z - w\|^2\) is a strongly convex function of \(z\) for any \(\lambda \in (0, \tilde{\lambda})\), where

\[
\tilde{\lambda} = \begin{cases} 
-1/\rho & \text{if } \rho < 0 \\
+\infty & \text{if } \rho \geq 0.
\end{cases} \tag{3.4}
\]
Thus, $\text{prox}_{\lambda g_j}$ defined by (1.6) is single-valued for any $\lambda \in (0, \bar{\lambda})$ and any $w \in E$. It is also not difficult to show that $\text{prox}_{\lambda g_j}$ is a Lipschitz continuous mapping, and in particular,

$$\|\text{prox}_{\lambda g_j}(w) - \text{prox}_{\lambda g_j}(z)\| \leq \frac{1}{1 + \rho \lambda} \|w - z\|, \quad \forall w, z \in E, \forall \lambda \in (0, \bar{\lambda}).$$

Hence, $\text{prox}_{\lambda g_j}$ is nonexpansive when $\rho \geq 0$. It also follows that the Moreau envelope $M_{g_j}^\lambda$ of $g_j$ is continuous.

By Assumption 3.1 (a) and (d), $h$ is convex, so we have from Beck [2017, Theorem 3.50] that

$$\text{conv}\{\nabla h_m(w) : m \in M \text{ such that } w \in D_m(h)\} = \partial h(w),$$

where $\partial h$ is the subdifferential of $h$:

$$\partial h(w) = \{v \in E : h(z) \geq h(w) + \langle v, z - w \rangle, \forall z \in E\}. \quad (3.5)$$

Finally, it is straightforward that Assumption 3.1 (a) and (3.3) imply that Assumption 3.1 (e) is equivalent to coerciveness of $f + g - h$.

As we will see in the forthcoming discussions, several algorithms for solving (OP) under Assumption 3.1 can be viewed as fixed point algorithms where the defining operators are union upper semicontinuous maps satisfying Assumption 1.1.

### 3.1 Proximal difference-of-min-convex algorithm

We now consider the fixed point iteration given by

$$w^{k+1} \in T_{\text{PDMC}}(w^k) := \text{prox}_{\lambda g} \left( w^k - \lambda f'(w^k) + \lambda h'(w^k) \right), \quad \text{(PDMC)}$$

where $\lambda \in (0, \bar{\lambda}) \cap (0, 1/L]$, $\bar{\lambda}$ is given by (3.4) and $L := \max_{i \in I} L_i$ given in Assumption 3.1 (b). We call the iterations (PDMC) the proximal difference-of-min-convex algorithm (or PDMC, for short). Here, $f' : E \Rightarrow E$ is defined by

$$f'(w) := \{\nabla f_i(w) : i \in I \text{ such that } w \in D_i(f)\}, \quad (3.6)$$

and similarly, $h' : E \Rightarrow E$ is given by

$$h'(w) := \{\nabla h_m(w) : m \in M \text{ such that } w \in D_m(h)\}. \quad (3.7)$$

Below, we cite some important properties of the proximal operator and the Moreau envelope of $g$ from Dao and Tam [2019, Proposition 5.2].

**Lemma 3.1.** Let $g = \min_{j \in J} g_j$, where $g_j$ is a proper function for all $j \in J$ and $\lambda > 0$. Then

(a) $M_{g}^\lambda(w) = \min_{j \in J} M_{g_j}^\lambda(w)$ for all $w \in E$.

(b) $\text{prox}_{\lambda g}(w) = \bigcup_{j : w \in D_j(M_{g_j}^\lambda)} \text{prox}_{\lambda g_j}(w)$, where $D_j(M_{g_j}^\lambda) := \{w \in E : M_{g_j}^\lambda(w) = M_{g_j}^\lambda(w)\}$.

In the setting of optimization problems, the objective function is the natural choice of Lyapunov function for descent algorithms. Indeed, we show this in the next theorem and use Theorem 2.1 to establish the global subsequential convergence of (PDMC).

**Theorem 3.2.** Let $\{w^k\}$ be any sequence generated by (PDMC) with $\lambda \in (0, \min\{\bar{\lambda}, 1/L\})$. Under Assumption 3.1, $\{w^k\}$ is bounded and its accumulation points belong to $\text{Fix}(T_{\text{PDMC}}^\lambda)$. 

---

8
\[ \frac{L_i}{2} \|z\|^2 - f_i(z) \geq \frac{L_i}{2} \|w\|^2 - f_i(w) + \langle L_i w - \nabla f_i(w), z - w \rangle, \quad \forall w, z \in \mathbb{E}, \forall i \in I, \]  
and since \( L \geq L_i \), we have that \( \frac{L}{2} \|w\| - f_i(w) \) is convex. Hence,

\[ \frac{L}{2} \|w\|^2 - f(w) = \max_{i \in I} \left\{ \frac{L}{2} \|w\|^2 - f_i(w) \right\} \]
is also a convex function. Then by [Beck, 2017, Theorem 3.50],

\[ \partial \left( \frac{L}{2} \|w\|^2 - f(w) \right) = \text{conv} \left\{ Lw - \nabla f_i(w) : i \in I \text{ such that } w \in D_i(f) \right\} \supseteq Lw - f'(w). \]

Thus,

\[ \frac{L}{2} \|z\|^2 - f(z) \geq \frac{L}{2} \|w\|^2 - f(w) + \langle Lw - y, z - w \rangle, \quad \forall w, z \in \mathbb{E}, \forall y \in f'(w). \]

By reversing the algebraic manipulations done to get (3.8) from (3.2), we have

\[ f(z) \leq Q^f(z, w) := f(w) + \langle y, z - w \rangle + \frac{1}{2\lambda} \|z - w\|^2, \quad \forall w, z \in \mathbb{E}, \]  
for any \( y \in f'(w) \) and \( \lambda \in (0, 1/L) \). Now, let \( w^+ \in T^\lambda_{\text{PDMC}}(w) \) for some \( w \notin \text{Fix}(T^\lambda_{\text{PDMC}}) \), say \( w^+ \in \text{prox}_{\lambda g}(w - \lambda y + \lambda v) \) for some \( y \in f'(w) \) and \( v \in h'(w) \subseteq \partial h(w) \). From (3.9) and (3.5) we have

\[ V(z) \leq Q^f(z, w) - Lh(z, w) + g(z), \quad \forall z \in \mathbb{E}, \]

where \( L_h(z, w) := h(w) + \langle v, z - w \rangle \). From that \( Q^f(z, w) - Lh(z, w) + g(z) \equiv V(w) \), standard calculations (for instance, see [Attouch et al., 2013, Section 5]) yield for \( \lambda \in (0, \bar{\lambda}) \) that

\[ \text{prox}_{\lambda g}(w - \lambda y + \lambda v) = \arg \min_{z \in \mathbb{E}} Q^f(z, w) - L_h(z, w) + g(z), \]  
(3.11)

\[ V(w) - V(w^+) \geq \frac{1 - \lambda L}{2\lambda} \|w^+ - w\|^2. \]  
(3.12)

Thus, \( V(w^+) < V(w) \) provided \( \lambda \in (0, \min\{\bar{\lambda}, 1/L\}) \). This proves that \( V \) is a Lyapunov function for (PDMC). By Theorem 2.1, it now suffices to show that \( T^\lambda_{\text{PDMC}} \) can be expressed as (1.2) satisfying Assumption 1.1. For each \( i := (i, j, m) \in I \times J \times M \), define \( T_i : D_i \to \mathbb{E} \) by

\[ T_i := \text{prox}_{\lambda g_j} \circ (Id - \lambda \nabla f_i + \lambda \nabla h_m), \quad \text{where} \]

\[ D_i := \left\{ w \in \mathbb{E} : w \in D_i(f) \cap D_m(h) \text{ and } w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w) \in D_j \left( M^\lambda_g \right) \right\}. \]  
(3.13)

By Lemma 3.1 (b), we obtain \( T^\lambda_{\text{PDMC}}(w) = \bigcup_{c \in c_i} T_i(w) \). Since \( \text{prox}_{\lambda g_j} \) is continuous on \( \mathbb{E} \) (see Remark 3.1 (c)), so is \( T_i \). By the continuity of \( M^\lambda_g \), together with Lemma 3.1 (a), \( D_j(M^\lambda_g) \) is closed. Hence, the continuity of \( \nabla f_i \) and \( \nabla h_m \) plus the closedness of \( D_i(f) \cap D_m(h) \) imply that \( D_i \) is closed. This completes the proof.

**Remark 3.2.** When \( \lambda = 1/L \), we cannot guarantee from (3.12) that \( V(w^+) \) is strictly less than \( V(w) \). But the result of Theorem 3.2 can still be valid for \( \lambda = 1/L \) if monotonicity of \( V \) is ensured by other mechanisms. One instance is provided in the following corollary, where the regularizer \( g \) is convex. We will explore some other particular examples in Section 5.2.
Corollary 3.3. If Assumption 3.1 (a), (b), (d), and (e) hold, and g is a convex function, then the conclusions of Theorem 3.2 hold for \( \lambda = 1/L \).

Proof. It suffices to show that \( V = f + g - h \) considered in Theorem 3.2 is a Lyapunov function when \( \lambda = 1/L \). Indeed, if g is convex, the objective function in (3.11) is \((1/\lambda)\)-strongly convex. Thus,

\[
\left[ Q_f^{1/L}(w, w) - L_h(w, w) + g(w) \right] - \left[ Q_f^{1/L}(w^+, w) - L_h(w^+, w) + g(w^+) \right] \geq \frac{L}{2} \|w^+ - w\|^2,
\]

which together with (3.10) gives \( V(w) - V(w^+) \geq (2\lambda)^{-1} \|w^+ - w\|^2 \). Hence, V is a Lyapunov function, as desired. \( \square \)

If g is a min-strongly convex function with parameter \( \rho \geq 1 \), then we obtain the global convergence of the full sequence generated by (PDMC) by utilizing Theorem 2.4.

Theorem 3.4. Let \( \{w^k\} \) be any sequence generated by (PDMC) with \( \lambda \in (0, 1/L) \). Suppose that Assumption 3.1 holds such that \( \text{Id} - \lambda \nabla f_i \) and \( \nabla h_m \) are nonexpansive for all \((i, m) \in I \times M\) and \( g_j \) is \( \rho \)-convex with \( \rho \geq 1 \) for all \( j \in J \). If \( w^* \) is an accumulation point of \( \{w^k\} \) and \( T^\lambda_{\text{PDMC}} \) is single-valued at \( w^* \), then \( w^k \to w^* \). Moreover, if \( \rho > 1 \), the convergence is locally linear.

Proof. By Theorem 3.2 and Theorem 2.4, it suffices to show that \( T_i \) given in (3.13) is calm at \( w^* \) with parameter \( \kappa_i \in [0, 1] \). Indeed, we have from Remark 3.1 (c) and the assumptions on \( \text{Id} - \lambda \nabla f_i \) and \( \nabla h_m \) that \( T_i \) is in fact nonexpansive if \( \rho \geq 1 \), and a contraction if \( \rho > 1 \). Thus, the proof is complete. \( \square \)

Remark 3.3. When \( f_i \) is a convex function, nonexpansiveness of \( \text{Id} - \lambda \nabla f_i \) holds as follows. From [Beck 2017 Theorem 5.8], (3.1) implies

\[
\langle \nabla f_i(w) - \nabla f_i(z), (w - z) \rangle \geq \frac{1}{L_i} \|\nabla f_i(w) - \nabla f_i(z)\|^2, \quad \forall w, z \in \mathbb{E},
\]

and thus, routine calculations give

\[
\|\text{Id} - \lambda \nabla f_i\|_2(w) - \|\text{Id} - \lambda \nabla f_i\|_2(z) \leq \|w - z\|^2 + \lambda(\lambda - 2/L)\|\nabla f_i(w) - \nabla f_i(z)\|^2. \tag{3.14}
\]

Since \( \lambda \in (0, 1/L) \), the second term on the right-hand side is nonpositive. Hence, \( \text{Id} - \lambda \nabla f_i \) is nonexpansive.

Forward-backward and projected subgradient algorithms

We now consider the simplified case of \( h \equiv 0 \) in (OP) with Assumption 3.1. That is, we look at the optimization problem

\[
\min_{w \in \mathbb{E}} f(w) + g(w). \tag{3.15}
\]

The forward-backward (FB) algorithm for it is given by

\[
w^{k+1} \in T^\lambda_{\text{FB}}(w^k) := \text{prox}_{\lambda g} \left( w^k - \lambda f'(w^k) \right), \tag{FB}
\]

where \( \lambda \in (0, \bar{\lambda}) \cap (0, 1/L) \). As a direct consequence of Theorem 3.2 and Corollary 3.3 we have the following result.

Theorem 3.5. Suppose that Assumption 3.1 holds with \( h \equiv 0 \). Then any sequence \( \{w^k\} \) generated by (FB) with \( \lambda \in (0, \min\{\bar{\lambda}, 1/L\}) \) is bounded, and its accumulation points belong to \( \text{Fix}(T^\lambda_{\text{FB}}) \). The same conclusions hold true when a stepsize of \( \lambda = 1/L \) is used, provided that g is convex.

In the case that each \( f_i \) is convex and \( g_j \) is \( \rho \)-convex with \( \rho \geq 0 \), we obtain a stronger result.
Definition 3.2. Let \( S \) point of \( \mathbb{R} \) problem, which is potentially easier to solve than the original one. If there is a unique the convergence point \( w \) that \( g \) is a point in \( \mathbb{R} \) when \( \lambda \in \{0, 1/L\} \), the mapping \( T_\lambda = \text{prox}_{\lambda g_j} (\text{Id} - \lambda \nabla f_i) \) satisfies \([2.3]\) where \( D_i \) is given in \([3.13]\). This immediately follows by noting from Remark 3.1 (c) that \( \text{prox}_{\lambda g_j} \) is nonexpansive when \( \rho \geq 0 \), and in particular a contraction when \( \rho > 0 \). \( \square \)

Observe that the above result only requires \( \rho \geq 0 \), whereas Theorem 3.4 needs \( \rho \) to be at least 1.

Projected subgradient algorithm

Suppose that \( g \) in \([3.15]\) is the indicator function of a union convex set \( S = \bigcup_{j \in J} R_j \), where \( J \) is a finite index set and \( R_j \) is closed and convex for each \( j \in J \). Thus, \( g \) is a min-convex function that satisfies Assumption 3.1 (a) and (c). Hence, we are looking at the union-convex-set-constrained problem

\[
\min f(w) \quad \text{subject to } w \in S.
\]

(3.16)

To solve \((3.16)\), we use \([FB]\) to obtain the projected subgradient \((PS)\) algorithm\(^{4}\)

\[
w^{k+1} \in T_{PS}^\lambda(w^k) := P_S(w^k - \lambda f'(w^k)),
\]

\[(PS)\]

where \( \lambda \in (0, 1/L] \). As this is simply a special case of \([FB]\), its global subsequential convergence under Assumption 3.1 (b) and (c) can be obtained directly from Theorem 3.5.

Theorem 3.7. Let \( S = \bigcup_{j \in J} R_j \) be a union convex set with \( J \) finite. Suppose that Assumption 3.1 (a), (b), and (e) hold. Then any sequence \( \{w^k\} \) generated by the projected subgradient algorithm \([PS]\) with \( \lambda \in (0, 1/L] \) is bounded, and its accumulation points belong to \( \text{Fix}(T_{PS}^\lambda) \). The same conclusions hold true when \( \lambda = 1/L \), provided \( S \) is convex.

We now explore some further properties of the iterations \([PS]\), starting with the following one.

Theorem 3.8 (Convex sets identification). Consider the setting of Theorem 3.7. Let \( \{w^k\} \) be a sequence generated by \([PS]\) with \( \lambda \in (0, 1/L] \), and let \( w^* \) be an accumulation point of it. If \( \text{Id} - \lambda \nabla f_i \) is nonexpansive for all \( i \in I \) and \( T_{PS}^\lambda \) is single-valued at \( w^* \), then \( w^k \to w^* \) and there exists \( N \geq 0 \) such that \( \{w^k\}_{k=N}^{\infty} \subseteq \bigcup_{j:w^* \in R_j} R_j \).

Proof. We have \( w^k \to w^* \) by invoking Theorem 3.6. By Remark 2.2, \( R_j \) contains at most finitely many terms of \( \{w^k\} \) if \( R_j \cap \text{Fix}(T_{PS}^\lambda) = \emptyset \). Thus, if \( j \in J \) such that \( w^* \notin R_j \), there are only finitely many terms of \( \{w^k\} \) in \( R_j \). Otherwise, there must be an infinite subsequence that lies on \( R_j \), which must accumulate to a point in \( R_j \). As the full sequence converges to \( w^* \) and \( R_j \) is closed, this contradicts the assumption that \( w^* \notin R_j \). As \( J \) is a finite index set, the claim now follows. \( \square \)

The above property has practical consequences in the same spirit as the component identification described in Remark 2.2. In particular, the locations of the iterates can be used to identify which \( R_j \)'s contain the convergence point \( w^* \). In turn, using an identified \( R_j \), we may reduce \((3.16)\) into a convex-constrained problem, which is potentially easier to solve than the original one. If there is a unique \( R_j \) that contains \( w^* \), we know from Theorem 3.8 that all iterates will eventually be in \( R_j \). In this case, we call \( w^* \) a nondegenerate point of \( S \).

Definition 3.2. Let \( \{R_j\}_{j \in J} \) be a finite collection of closed convex sets, \( S = \bigcup_{j \in J} R_j \), and \( w \in S \). We say that \( w \) is a nondegenerate point of \( S \) if there exists a unique \( j \in J \) such that \( w \in R_j \). Otherwise, it is called a degenerate point of \( S \).

---

\(^{4}\) Clearly, \( \text{prox}_{\lambda g} = P_S \) for any \( \lambda > 0 \) if \( g = \delta_S \).
Following [Theorem 3.6], we obtain full convergence of the PS iterates when \( Id - \lambda \nabla f_i \) is nonexpansive. Using [Theorem 3.8], we can further attain a local linear rate of convergence when the restriction of each function \( Id - \lambda \nabla f_i \) to each convex set \( R_j \) is a contraction.

**Proposition 3.9** (Linear convergence). Consider the setting of [Theorem 3.7] and suppose \( Id - \lambda \nabla f_i \) is \( \kappa_{ij} \)-Lipschitz continuous on \( R_j \) with \( \kappa_{ij} \in [0,1) \) for all \((i,j) \in I \times J \) and \( \lambda \in (0,1/L) \). Let \( \{w^k\} \) be any sequence generated by \( \text{PS} \), and let \( w^* \) be an accumulation point of it. If \( T_{\text{PS}} \) is single-valued at \( w^* \), then \( w^k \to w^* \) with a local linear rate.

### 3.2 Fixed point sets and critical points

We have thus established the convergence of \( \text{PDMC} \) to fixed points of \( T^\lambda_{\text{PDMC}} \). In the next theorem, we show that being a fixed point is a necessary condition for optimality.

**Theorem 3.10.** Let \( w \) be a local minimum of \( \text{OP} \) with Assumption 3.1 holds, then

(a) \( w \) is a local minimum of \( f_i + g_j - h_m \) for all \((i,j,m) \in I \times J \times M \) such that \( w \in D_i(f) \cap D_j(g) \cap D_m(h) \);

(b) there exists \( \varepsilon \in (0,1/L) \), dependent on \( w \), such that \( w \in \text{Fix}(T^\lambda_{\text{PDMC}}) \) for any \( \lambda \in (0,\bar{\lambda}) \cap (0,\varepsilon] \); and

(c) if \( w \) is a global minimum, then \( T^\lambda_{\text{PDMC}} \) is single-valued at \( w \), and \( w \in \text{Fix}(T^\lambda_{\text{PDMC}}) \) for all \( \lambda \in (0,\bar{\lambda}) \cap (0,1/L) \).

One disadvantage of the local optimality condition given in Theorem 3.10 (b), however, is that a local minimum might not be a fixed point of \( T^\lambda_{\text{PDMC}} \) when \( \lambda \in (0,\bar{\lambda}) \) but \( \lambda \in (\varepsilon,1/L) \) (see Example 4.1). In search for global minima of \( \text{OP} \), on the other hand, the above theorem provides an intuition that larger but permissible values of \( \lambda \) must be avoided to avoid getting stuck at spurious local optima. Of course, from a numerical point of view, a larger stepsize is often also more desirable to obtain faster convergence of the algorithms.

A more standard necessary condition for optimality is criticality. We first need the following technical definition [see, for example, Rockafellar and Wets, 1998]. For any function \( F \), its subdifferential at \( w \) is

\[
\partial F(w) := \limsup_{\bar{w} \to w, F(\bar{w}) \to F(w)} \hat{\partial} F(\bar{w}),
\]

where \( \hat{\partial} F(\bar{w}) := \{v : v \in \mathbb{E}, h(z) \geq h(w) + \langle v, z-w \rangle + o(\|z-w\|)\} \),

which coincides with (3.5) when \( F \) is convex. With this definition, we say that \( w \) is a critical point of \( f + g - h \) if

\[
0 \in \partial f(w) + \partial g(w) - \partial h(w).
\]

Indeed, by Assumption 3.1 (b) and (d), \( f \) and \( h \) are piecewise smooth functions in the sense of Facchinei and Pang, 2003, Definition 4.5.1, which are locally Lipschitz continuous at any point [Facchinei and Pang, 2003, Lemma 4.6.1 (a)]. Consequently, we obtain from Rockafellar and Wets, 1998, Exercise 10.10 that \( \partial(f + g - h)(w) \subseteq \partial f(w) + \partial g(w) - \partial h(w) \), where equality holds if \( f \) and \( h \) are differentiable at \( w \). Hence, by Rockafellar and Wets, 1998, Theorem 10.1], a local minimum of \( f + g - h \) is a critical point. We now show that criticality is a tighter condition than being a fixed point.

**Theorem 3.11.** Suppose that Assumption 3.1 holds and \( D_i(f) \) is a regular closed set, that is, \( D_i(f) = \text{cl}(\text{int}(D_i(f))) \) for any \( i \in I \), where \( \text{cl} \) and \( \text{int} \) are respectively the closure and the interior of a set. Then any fixed point of \( T^\lambda_{\text{PDMC}} \) is a critical point of \( f + g - h \).

**Proof.** Let \( w \in \text{Fix}(T^\lambda_{\text{PDMC}}) \), say \( w \in \text{prox}_{\lambda g}(w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w)) \) for some \((i,m) \in I \times M \). Then \(-\nabla f_i(w) + \nabla h_m(w) \in \partial g(w) \) by Rockafellar and Wets, 1998, Theorem 10.1]. Since \( \nabla h_m(w) \in \partial h(w) \) by Remark 3.1 (d), it suffices to show that \( \nabla f_i(w) \in \partial f(w) \). Since \( w \in D_i(f) \) and we have from hypothesis that \( D_i(f) = \text{cl}(\text{int}(D_i(f))) \), there exists a sequence \( \{w^k\} \subseteq \text{int}(D_i(f)) \) such that \( w^k \to w^* \). Moreover, since \( f \equiv f_i \) on \( \text{int}(D_i(f)) \), \( f \) is differentiable on \( \text{int}(D_i(f)) \), and so \( \partial f(w^k) = \{\nabla f_i(w^k)\} \) for all \( k \). By the continuity of \( \nabla f_i \), we then have \( \nabla f_i(w^k) \to \nabla f_i(w) \) so that \( \nabla f_i(w) \in \partial f(w) \), as desired. \( \square \)
Remark 3.4. (a) Theorem 3.11 together with Theorem 3.2 guarantees that (PDMC) is globally subsequentially convergent to critical points of the objective function. Note that the assumption on $D_i(f)$ trivially holds when $|I| = 1$, as in the illustrative applications that we will see in Section 5.

(b) The fixed point set $\text{Fix}(T_{\text{PDMC}}^\lambda)$ may be strictly contained in the set of critical points; for instance, see Example 4.1.

We close this section with the following property of fixed point sets of $T_{\text{PS}}^\lambda$.

Proposition 3.12. If $w \in \text{Fix}(T_{\text{PS}}^\lambda)$ for some $\lambda \in (0, 1/L]$ is a nondegenerate point of $S$, then there exists $\varepsilon > 0$ such that $w \in \text{Fix}(T_{\text{PS}}^\lambda)$ for all $\lambda \in (0, \varepsilon]$.

3.3 Acceleration schemes for PDMC

We may accelerate PDMC using the scheme described in Algorithm 1. In this section, we propose two further acceleration techniques. The first one is motivated by the following example.

Example 3.1. Let $E = \mathbb{R}^2$, $S_1$ be any straight line with a positive slope, and $S_2 = A \cup B$, where $A := \{(a, 0) : a \geq 0\}, B := \{(0, b) : b \geq 0\}$; see Fig. 1. Consider (OP) with $f(w) = \frac{1}{2} \text{dist}(w, S_1)^2$, $g(w) = \delta_{S_2} = \min\{\delta_A(w), \delta_B(w)\}$, and $h \equiv 0$. Then it can be shown that the PDMC iterates with stepsize $\lambda = 1$ coincide with the MAP iterates; see also Section 4.3. Notice that this algorithm generates points confined in the union convex set $S_2$. To speed up the convergence of the algorithm to the solution, we conduct extrapolation if two consecutive iterates lie on the same convex set. As illustrated in Fig. 1, if $w^{k-1}$ and $w^k$ both lie on $A$ or $B$, we extrapolate along the direction $w^k - w^{k-1}$ to get an intermediate point $z^k$ before conducting alternating projections to obtain $w^{k+1}$. Intuitively, the iterates generated by this procedure tend to get closer to $S_1$ faster than when only (non-accelerated) MAP is used.

Inspired by the above example, we include a restriction in the stepsize used in Step 1 of Algorithm 1. In particular, we only proceed with the extrapolation step when two consecutive iterates “activate” the same components in $f$, $g$ and $h$. To formalize this idea, let

$$\hat{D} := \{D_i(f) \cap D_j(g) \cap D_m(h) : (i, j, m) \in I \times J \times M\}$$

and define

$$\chi_k := \begin{cases} 1 & \text{if } \phi_{\hat{D}}(w^k) \cap \phi_{\hat{D}}(w^{k-1}) \neq \emptyset \text{ and } k \geq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $\phi_{\hat{D}}$ is defined in (1.5). Then, as summarized in Algorithm 2, we simply replace the stepsize $t_k$ in Step 1 of Algorithm 1 by $t_k \chi_k$ to take into account the above-described restriction. It is clear that global
subsequential convergence of Algorithm 2 to a fixed point of $T_{\text{PDMC}}^\lambda$ directly follows from Theorem 2.2.

**Algorithm 2:** Accelerated proximal difference-of-min-convex algorithm for (OP).

Let $V = f + g - h$.

Choose $\sigma > 0$, $\lambda \in (0, 1/L] \cap (0, \bar{\lambda})$, and $w^0 \in \mathbb{E}$. Set $w^{-1} = w^0$ and $k = 0$.

**Step 1.** Set $z^k = w^k + t_k \chi_k p^k$, where $p^k = w^k - w^{k-1}$, $t_k \geq 0$ satisfies

$$V(z^k) \leq V(w^k) - \frac{\sigma}{2} t_k^2 \chi_k^2 \|p^k\|^2,$$

and $\chi_k$ is given by (3.18).

**Step 2.** Set $w^{k+1} \in T_{\text{PDMC}}^\lambda(z^k)$, $k = k+1$, and go back to Step 1.

We propose another acceleration scheme in Algorithm 3 motivated by Theorem 3.10 and Lemma 2.3 applied to the collection (3.17). Suppose $\{w^k\} \subseteq \text{dom}(g)$ is a sequence generated by (PDMC) such that $w^k \to w^*$, as in the settings described in Theorems 3.4, 3.6 and 3.8. From (3.3) and Lemma 2.3 there exists $N > 0$ large enough so that $\phi_D(w^k) \subseteq \phi_D(w^*)$ for all $k \geq N$. Thus, the iterates can be used to identify which components of the objective function are activated by the convergence point $w^* \in \text{Fix}(T_{\text{PDMC}}^\lambda)$, which is a candidate solution of (OP) by Theorem 3.10. Hence, in the search for $w^*$, we can switch to a simpler optimization problem as in (3.20) after a finite number of iterations, generalizing the conclusion of Theorem 3.8. Algorithms similar in spirit as Algorithm 3 for regularized optimization when the objective function is partly smooth are also studied in [Li et al., 2020; Lee, 2020; Lee and Wright, 2012] under the framework of manifold identification.

**Algorithm 3:** Proximal difference-of-min-convex algorithm with component identification for (OP).

Choose $w^0 \in \mathbb{E}$, $N \in \mathbb{N}$. Set Unchanged = 0, $k = 0$.

**Step 1.** Set Unchanged = $\chi_k(\text{Unchanged} + 1)$, where $\chi_k$ is given by (3.18).

**Step 2.** Compute $w^{k+1}$ according to the following rule:

1. If Unchanged $< N$: set $w^{k+1} \in T_{\text{PDMC}}^\lambda(w^k)$. Terminate if $w^{k+1} \in \text{Fix}(T_{\text{PDMC}}^\lambda)$; otherwise, set $k = k+1$ and go back to Step 1.

2. If Unchanged $= N$: pick $(i, j, m) \in \phi_D(w^k)$ and solve

$$w^{k+1} \in \arg \min_{z \in \mathbb{E}} f_i(z) + g_j(z) - h_m(z).$$

Terminate if $w^{k+1} \in \text{Fix}(T_{\text{PDMC}}^\lambda)$; otherwise, set Unchanged $= -1$, $w^{k+1} = w^k$, $k = k+1$, and go back to Step 1.

4 Applications to feasibility problems involving union convex sets

We now revisit the feasibility problem (FP) with $S_1$ and $S_2$ being union convex sets, say

$$S_1 = \bigcup_{i \in I} R_i^{(1)}, S_2 = \bigcup_{j \in J} R_j^{(2)},$$

where $|I|, |J| < \infty$, $R_k^{(l)}$ is convex, $\forall k, l$. (4.1)

Throughout this section, we assume that $S_1 \cap S_2 \neq \emptyset$. We establish the global subsequential convergence of the methods of averaged projections and alternating projections, which can be viewed as specific instances of the algorithms in the preceding section.
4.1 Method of averaged projections

Observe that (FP) can be reformulated as an optimization problem:
\[
\min_{w \in E} \frac{1}{2} \text{dist}(w, S_1)^2 + \frac{1}{2} \text{dist}(w, S_2)^2, \tag{4.2}
\]
and each term is a min-convex function if (4.1) holds. Indeed,
\[
f(w) := \frac{1}{2} \text{dist}(w, S_1)^2 = \min_{i \in I} \left\{ f_i(w) := \frac{1}{2} \text{dist} \left( w, R_i^{(1)} \right)^2 \right\}. \tag{4.3}
\]
By the convexity of \( R_i^{(1)} \), \( f_i \) is a convex function whose gradient, namely \( \nabla f_i(w) = w - P_{R_i^{(1)}}(w) \), is 1-Lipschitz continuous (see [Beck, 2017, Example 5.5]). On the other hand, we also have
\[
\frac{1}{2} \text{dist}(w, S_2)^2 = \frac{1}{2} \|w\|^2 - \left( \frac{1}{2} \|w\|^2 - \frac{1}{2} \text{dist}(w, S_2)^2 \right) =: g(w) - h(w). \tag{4.4}
\]
Note that \( h \) can also be expressed as
\[
h(w) = \max_{j \in J} \left\{ h_j(w) := \frac{1}{2} \|w\|^2 - \frac{1}{2} \text{dist} \left( w, R_j^{(2)} \right)^2 \right\},
\]
and is therefore convex. Thus, (4.4) is a difference-of-convex decomposition of \( \frac{1}{2} \text{dist}(w, S_2)^2 \), and by (4.1) \( g \) and \( h \) satisfy Assumption 3.1 (a), (c), and (d). Hence, we may use (PDMC) to solve (4.2) In this setting, \( f' \) defined in (3.6) can be simplified to
\[
f'(w) = \left\{ w - P_{R_i^{(1)}}(w) : \text{dist}(w, S_1) = \text{dist} \left( w, R_i^{(1)} \right) \right\} = w - P_{S_1}(w) \tag{4.5}
\]
while \( h' \) given by (3.7) reduces to \( h'(w) = P_{S_2}(w) \). Since \( \text{prox}_{\lambda g}(w) = \frac{1}{1 + \lambda} w \) for all \( w \in E \), (PDMC) simplifies to
\[
w^{k+1} \in T_{\text{MAveP}}^\lambda(w^k) := \left( \frac{1 - \lambda}{1 + \lambda} I + \frac{\lambda}{1 + \lambda} (P_{S_1} + P_{S_2}) \right) (w^k), \quad \lambda \in (0, 1]. \tag{4.6}
\]
When \( \lambda = 1 \), we denote \( T_{\text{MAveP}} := T_{\text{MAveP}}^1 \) and the above algorithm further simplifies to
\[
w^{k+1} \in T_{\text{MAveP}}(w^k) = \frac{(P_{S_1} + P_{S_2})(w^k)}{2},
\]
which we refer to as the method of averaged projections (MAveP). We have the following convergence guarantee for MAveP from our developed theory.

**Theorem 4.1.** Let \( S_1 \) and \( S_2 \) be union convex sets given by (4.1) and suppose that
\[
V_{ij}(w) := \frac{1}{2} \text{dist} \left( w, R_i^{(1)} \right)^2 + \frac{1}{2} \text{dist} \left( w, R_j^{(2)} \right)^2
\]
is coercive over \( E \) for any \((i, j) \in I \times J\). Then any sequence \( \{w^k\} \) generated by (4.6) with \( \lambda \in (0, 1] \) is bounded, and its accumulation points belong to \( \text{Fix}(T_{\text{MAveP}}^\lambda) \). Moreover, if \( T_{\text{MAveP}}^\lambda \) is single-valued at an accumulation point \( w^* \), then \( w^k \to w^* \).

**Proof.** First, note that Assumption 3.1 (a)-(d) are satisfied by \( f \), \( g \) and \( h \) defined in (4.3) and (4.4). On the other hand, the coerciveness assumption on \( V_{ij} \) results to satisfaction of Assumption 3.1 (e). Since \( g \) is convex, we obtain from Corollary 3.3 that \( \{w^k\} \) is bounded, with accumulation points being contained in \( \text{Fix}(T_{\text{MAveP}}^\lambda) \). Finally, \( I - \lambda \nabla f_i \) is nonexpansive by Remark 3.3 and so is \( \nabla h_j = P_{R_j^{(2)}} \). Hence, the last claim follows from Theorem 3.4. \( \square \)

\( ^E \) Each component is the convex conjugate of \( \|w\|^2/2 + \delta_{R_j^{(2)}}(w) \), and the maximum of convex functions is convex.


4.2 Method of alternating relaxed projections

Another projection algorithm for solving (FP) can be obtained by applying directly the FB algorithm to (4.2). Let \( f \) be given by (4.3), and denote \( g := \text{dist}(\cdot, S_2)^2/2 \) and \( g_j := \text{dist}(\cdot, R_j^{(2)})^2/2 \). From Beck [2017, Example 6.65], we have

\[
\text{prox}_{\lambda g}(w) = \frac{\lambda}{1 + \lambda} P_{S_2}(w) + \frac{1}{1 + \lambda} w, \quad \forall w \in \mathbb{E}.
\]

(4.7)

Using the optimality condition of (1.7) and Lemma 3.1, the above formula (4.7) for the proximal mapping of \( g_j \) extends to that of \( g \):

\[
\text{prox}_{\lambda g}(w) = \frac{\lambda}{1 + \lambda} P_{S_2}(w) + \frac{1}{1 + \lambda} w.
\]

Thus, (FB) becomes

\[
w^{k+1} \in T_{\lambda \text{MARP}}(w^k)
\]

\[
:= \frac{\lambda}{1 + \lambda} P_{S_2}((1 - \lambda)w^k + \lambda P_{S_1}(w^k)) + \frac{1}{1 + \lambda}((1 - \lambda)w^k + \lambda P_{S_1}(w^k))
\]

(4.8)

for \( \lambda \in (0, 1] \), which recovers a special instance of the method of alternating relaxed projections (MARP) studied in Bauschke et al. [2013]. Arguing as in the proof of Theorem 4.1 and using Theorem 3.6, we obtain the following convergence result.

**Theorem 4.2.** Under the hypotheses of Theorem 4.1, any sequence \( \{w^k\} \) generated by (4.8) with \( \lambda \in (0, 1) \) is bounded, and its accumulation points belong to \( \text{Fix}(T_{\lambda \text{MARP}}) \). Moreover, if \( T_{\lambda \text{MARP}} \) is single-valued at an accumulation point \( w^* \), then \( w^k \to w^* \).

4.3 Method of alternating projections

Another equivalent reformulation of the feasibility problem (FP) is

\[
\min_{w \in \mathbb{E}} f(w) + \delta_{S_2}(w),
\]

(4.9)

where \( f \) is defined by (4.3). By (4.5), the PS algorithm (PS) then takes the form

\[
w^{k+1} \in T_{\lambda \text{MAP}}(w^k) := P_{S_2}((1 - \lambda)w^k + \lambda P_{S_1}(w^k))
\]

(4.10)

with \( \lambda \in (0, 1] \). When \( \lambda = 1 \), we denote \( T_{\text{MAP}} := T_{\text{MAP}}^1 \) and the algorithm simplifies to

\[
w^{k+1} \in T_{\text{MAP}}(w^k) := (P_{S_2} \circ P_{S_1})(w^k),
\]

(4.11)

that is, the method of alternating projections (MAP). Recalling that \( 1d - \lambda \nabla f_i \) is nonexpansive (see the proof of Theorem 4.1), we obtain the following as a direct consequence of Theorems 3.7 and 3.8 and Proposition 3.9.

**Theorem 4.3.** Let \( S_1 \) and \( S_2 \) be given by (4.1), and suppose that \( f_i := \frac{1}{2} \text{dist}(\cdot, R_i^{(1)})^2 \) is coercive over \( R_j^{(2)} \) for all \( (i, j) \in I \times J \). Then any sequence \( \{w^k\} \) generated by (4.10) with \( \lambda \in (0, 1) \) is bounded, and its accumulation points belong to \( \text{Fix}(T_{\lambda \text{MAP}}) \). Moreover, if \( T_{\lambda \text{MAP}} \) is single-valued at an accumulation point \( w^* \), the following hold:

(a) \( w^k \to w^* \);

(b) There exists \( N \geq 0 \) such that \( \{w^k\}_{k=N}^{\infty} \subseteq \bigcup_{j:w^* \in R_j^{(2)}} R_j^{(2)} \); and

(c) If \( P_{R_j^{(1)}} \) is \( \kappa_{ij} \)-Lipschitz continuous on \( R_j^{(2)} \) with \( \kappa_{ij} \in [0, 1) \) for all \( (i, j) \in I \times J \), the convergence in (a) is locally linear.
Unfortunately, Theorem 4.3 applies only when \( \lambda \in (0, 1) \) because \( V := f + \delta_{S_2} \) may not be a Lyapunov function when \( \lambda = 1 \) (see [3.12]). Hence, we cannot guarantee the global (subsequential) convergence of the MAP scheme. This in fact is a major theoretical problem for considering MAP in (4.11). In the literature, particularly for nonconvex feasibility problems, it is often the case that global convergence results are only obtained for some relaxations of MAP with \( \lambda < 1 \) in [4.10] [Alcantara et al., 2022] [Attouch et al., 2010, 2013, Bauschke et al., 2013]. In Section 5.2.2, we overcome this challenge to show that (4.11) attains global convergence for a union-convex-feasibility reformulation of LCP problems. On the other hand, when \( S_2 \) is a convex set, we can obtain the conclusions of Theorem 4.3 (a) and (d) for (4.11) by utilizing Theorem 3.7.

**Theorem 4.4 (Global convergence of MAP for union convex set given in (4.1))**. Let \( S_1 \) be a union convex set given in (4.1) and \( S_2 \) be a convex set. If \( f_i = \text{dist}(\cdot, R_{i}^{(1)})^2/2 \) is coercive over \( S_2 \), then any sequence \( \{w^k\} \) generated by MAP in (4.11) is bounded, with accumulation points lying in \( \text{Fix}(T_{MAP}) \). If \( T_{MAP} \) is single-valued at an accumulation point \( w^* \), then \( w^k \to w^* \). The convergence is locally linear if \( P_{R_i}^{(1)} \) is \( \kappa_i \)-Lipschitz continuous on \( S_2 \) with \( \kappa_i < 1 \) for all \( i \in I \).

**Remark 4.1.** In the setting of Theorem 4.4, MAP first projects onto the union convex set \( S_1 \), and then onto the convex set \( S_2 \). However, we do note that in the literature (see Attouch et al., 2010, Hesse et al., 2014, for instance), MAP conventionally starts with the projection onto the convex set, before projecting onto the nonconvex set.

We now demonstrate by an example the relationship established in Section 3.1 among the sets containing the global/local minima and critical points of (4.9) and the fixed points of \( T_{MAP}^{\lambda} \).

**Example 4.1.** Consider (4.9) with \( S_1 = \{(a, 1) : a \in \mathbb{R}\} \) and \( S_2 \) given in Example 3.1. We note that the set \( S^* \) of local minima of (4.9) and \( \text{Fix}(T_{MAP}^{\lambda}) \) are given respectively by \( S^* := \{(0, 1)\} \cup \{(t, 0) : t > 0\} \) and \( \text{Fix}(T_{MAP}^{\lambda}) = \{(0, 1)\} \cup \{(t, 0) : t \geq \lambda\} \) for any \( \lambda \in (0, 1) \). Clearly, \( S^* \) is a subset of \( C^* \), the set of critical points of \( f + \delta_{S_2} \). Thus, we have \( \{(0, 1)\} = S_1 \cap S_2 \subseteq \text{Fix}(T_{MAP}^{\lambda}) \subseteq C^* \) for any \( \lambda \in (0, 1) \). It is then not difficult to verify the claims of Theorem 3.10 and Theorem 3.11. Moreover, observe that each point in \( \text{Fix}(T_{MAP}^{\lambda}) \) is nondegenerate, and we have \( \text{Fix}(T_{MAP}^{\lambda}) \subseteq \text{Fix}(T_{MAP}^{\lambda}) \) whenever \( \lambda \geq \lambda \), demonstrating Proposition 3.12.

The following remarks conclude this section.

**Remark 4.2.** (a) Let \( T \in \{T_{MAveP}^{\lambda}, T_{MARP}^{\lambda}, T_{MAP}^{\lambda}\} \). It is clear that \( S_1 \cap S_2 \subseteq \text{Fix}(T) \), which is also a consequence of Theorem 3.10 (c). When \( S_1 \cap S_2 = \text{Fix}(T) \), combining Theorems 4.1 to 4.3 asserts that MAveP, MARP and MAP are globally convergent to solutions of the feasibility problem under coerciveness assumptions. One instance for such an equality to hold is when \( S_1 \) and \( S_2 \) are both convex sets, but this is not always attainable for general nonconvex problems. We will deal with some nonconvex feasibility problems where equality of these two sets hold in Section 5.

(b) From Remark 2.2 and Algorithm 2, we know that latter iterations of MAveP, MARP and MAP can be used to reduce the feasibility reformulations (4.2) and (4.9) to simpler problems. In particular, similar to (3.20), (4.2) and (4.9) will be simplified to solving

\[
\min_{w \in \mathbb{E}} \frac{1}{2} \text{dist} \left( w, R_{i}^{(1)} \right)^2 + \frac{1}{2} \text{dist} \left( w, R_{j}^{(2)} \right)^2,
\]

and

\[
\min_{w \in \mathbb{E}} \frac{1}{2} \text{dist} \left( w, R_{i}^{(1)} \right)^2 + \delta_{R_{i}^{(1)}}(w),
\]

respectively, for some \( (i, j) \in I \times J \). That is, the original nonconvex feasibility problem (FP) can be reduced to an easier (convex) problem of finding the intersection of two convex sets \( R_{i}^{(1)} \) and \( R_{j}^{(2)} \).
5 Affine-union convex set feasibility problems

In this section, we establish global convergence of several algorithms for solving (FP) involving an affine set $S_1$, say

$$S_1 = \{ w \in E : Aw = b \},$$

where $A$ is a matrix with full row rank, $E = \mathbb{R}^q$ for some $q$, and $S_2$ is a union convex set. Specifically, we consider the sparse affine feasibility problem and a feasibility reformulation of the linear complementarity problem in Sections 5.1 and 5.2, respectively. Recall that in general, the feasibility problem (FP) can be reformulated as an optimization problem, either as (4.2) or (4.9). Other than these reformulations, the affine structure of $S_1$ given by (5.1) also allows for recasting the feasibility problem as

$$\min_{w \in E} \frac{1}{2} \|Aw - b\|^2 + \frac{1}{2} \text{dist}(w, S_2)^2,$$

and

$$\min_{w \in E} \frac{1}{2} \|Aw - b\|^2 + \delta_{S_2}(w).$$

To unify the analyses of algorithms for these four optimization reformulations, we first note that the projection onto $S_1$ is given by [Bauschke and Kruk, 2004, Lemma 4.1]

$$P_{S_1}(w) = w - A^\dagger(Aw - b).$$

Here, $A^\dagger$ is the Moore-Penrose inverse of $A$, which is given by $A^\dagger = A^T( AA^T)^{-1}$ since $A$ has full row rank. With this, we have

$$\frac{1}{2} \text{dist}(w, S_1)^2 \overset{(5.4)}{=} \frac{1}{2} \|A^\dagger(Aw - b)\|^2 = \frac{1}{2} w^T A^TQAw - w^T A^T Qb + \frac{1}{2} b^T Qb,$$

where $Q := (AA^T)^{-1}$. By denoting

$$f_Q(w) := \frac{1}{2} w^T A^TQAw - w^T A^T Qb + \frac{1}{2} b^T Qb,$$

we get $f_Q(w) = \|Aw - b\|^2/2$ if $Q = I$, and $f_Q(w) = \text{dist}(w, S_1)^2/2$ if $Q = (AA^T)^{-1}$. Thus, we may unify the convergence analyses of algorithms for (4.2) and (5.2) through varying $Q$ in

$$\min_{w \in \mathbb{R}^n} f_Q(w) + \frac{1}{2} \text{dist}(w, S_2)^2,$$

and similarly for (4.9) and (5.3) we may consider

$$\min_{w \in \mathbb{R}^n} f_Q(w) + \delta_{S_2}(w).$$

We note that (5.6) and (5.7) fit in the setting of Section 3 so the general convergence results apply. First, the gradient of $f_Q$ in (5.5)

$$\nabla f_Q(w) = A^TQ(Aw - b) = \begin{cases} A^\dagger(Aw - b) & \text{if } Q = (AA^T)^{-1}, \\ A^T(Aw - b) & \text{if } Q = I, \end{cases}$$

is Lipschitz-continuous with parameter

$$L_Q := \begin{cases} 1 & \text{if } Q = (AA^T)^{-1}, \\ \|A\|^2 & \text{if } Q = I, \end{cases}$$

so Assumption 3.1 (b) is satisfied. Moreover, we have the following:
(i) As noted in Section 4.1, $g$ and $h$ given in (4.4) satisfy Assumption 3.1 (c) and (d) since $S_2$ is a union convex set. By using this decomposition in (5.6) and then applying (PDMC), we get

$$w^{k+1} \in T_{PDMC}^\lambda(w^k) = \frac{1}{1+\lambda} \left( w^k - \lambda \nabla Q(w) + \lambda P_{S_2}(w^k) \right).$$  

(ii) A direct application of (FB) to (5.6) with $g = \text{dist}(\cdot, S_2)^2/2$, which satisfies Assumption 3.1 (c), leads to

$$w^{k+1} \in T_{FB}^\lambda(w^k) = \frac{\lambda}{1+\lambda} P_{S_2}(w^k - \lambda \nabla Q(w^k)) + \frac{1}{1+\lambda} (w^k - \lambda \nabla Q(w^k)).$$

(iii) The PS algorithm (PS) for solving (5.7) is given by

$$w^{k+1} \in T_{PS}^\lambda(w^k) = P_{S_2}(w^k - \lambda \nabla Q(w^k)).$$

If $Q = (AA^T)^{-1}$, the operators $T_{PDMC}^\lambda$, $T_{FB}^\lambda$ and $T_{PS}^\lambda$ above respectively coincide with $T_{MAP}^{\lambda_{AVE}}$, $T_{MAP}^{\lambda_{P}}$ and $T_{MAP}^{\lambda_{SRIP}}$ presented in Section 4.

**Remark 5.1.** By the convexity of $f_Q$, we have from Remark 3.3 that $1d - \lambda \nabla f_Q$ is nonexpansive. It is thus clear that the individual operators in the decompositions of $T_{PDMC}^\lambda$, $T_{FB}^\lambda$ and $T_{PS}^\lambda$ in the form (1.2) are all nonexpansive.

Before diving into more specific problem classes, we mention a few words about the notations that will be used in this section. We let $w_\perp := \max\{w, 0\}$, where the maximum is taken componentwise. $\text{Ran}(A)$ and $\text{Ker}(A)$ denote the range and kernel of a matrix $A \in \mathbb{R}^{p \times q}$. We let $\|A\|$ denote the operator norm of $A$, i.e., $\|A\|$ is the largest singular value of $A$. We also let $[q] = \{1, \ldots, q\}$, and for $\Lambda \subseteq [q]$, we denote by $A_{\Lambda}$ the submatrix of $A$ containing all of its columns indexed by $\Lambda$, $A_{A,\Lambda}$ the submatrix of $A$ containing its rows and columns indexed by $\Lambda$, and $\Lambda^c$ the complement set $\{i : i \in [q], i \not\in \Lambda\}$. Given $w \in \mathbb{R}^q$, $w_\Lambda \in \mathbb{R}^{|\Lambda|}$ denote the subvector of $w$ indexed by $\Lambda$.

### 5.1 Sparse affine feasibility

We consider the **sparse affine feasibility problem** (SAFP), which involves solving (FP) with

$$S_1 = \{w \in \mathbb{R}^n : Aw = b\}, \quad S_2 = A_\Lambda := \{w \in \mathbb{R}^n : \|w\|_0 \leq s\},$$

where $0 < s \leq n$, $A \in \mathbb{R}^{m \times n}$ has full row rank, $b \in \mathbb{R}^m$, and $\|w\|_0$ denotes the number of nonzero components of $w$. Hesse et al. [2014] have shown that $S_2 = A_\Lambda$ can be decomposed as

$$A_\Lambda = \bigcup_{t \in I_s} R_t, \quad I_s := \{t \subseteq [n] : t \text{ has } s \text{ elements}\}, \quad R_t := \text{Ran}(\mathbb{I}_{t,\Lambda}),$$

so $S_2$ is indeed a union convex set and the projection onto $S_2$ is given by

$$P_{S_2}(w) = \{ P_{R_t}(w) : t \in I_s \text{ such that } \min_{j \in t} |w_j| \geq \max_{j \in \Lambda^c} |w_j| \}.$$ In turn, we can use the algorithms (5.10)-(5.12) to solve the sparse affine feasibility problem.

Convergence analyses of existing methods for problems involving sparsity usually require near-orthonormality conditions on the matrix $A$, such as the **restricted isometry property (RIP)** introduced in Candès and Tao [2005]. A more general condition subsuming the RIP is the **scalable restricted isometry property (SRIP)** proposed in Beck and Teboulle [2011]. A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy SRIP of order $(d, \alpha)$ if there exist $\mu_d > \nu_d > 0$ with $\mu_d/\nu_d < \alpha$ such that

$$\nu_d \|w\|^2 \leq \|Aw\|^2 \leq \mu_d \|w\|^2, \quad \forall w \in A_d.$$ Using the results in Section 3, we only need a uniform lower bound for $\|Aw\|^2/\|w\|^2$ over $A_\Lambda$, as opposed to SRIP, to guarantee that PDMC, FB, and PS on (5.6) and (5.7) are globally subsequentially convergent to fixed points, which by Theorems 3.10 and 3.11 are candidate solutions to the feasibility problem. To establish the convergence results, we note the following simple but useful lemma.
Lemma 5.1. Let $A \in \mathbb{R}^{m \times n}$ and $Q \in \mathbb{R}^{m \times m}$ be with $\text{rank}(A) = m$, and let $\Lambda \subseteq [n]$. If $\text{rank}(A_{\cdot \Lambda}) = |\Lambda|$, then $\text{rank}((A^TQA)_{\cdot \Lambda}) = |\Lambda|$. Consequently, $\lambda_{\min}((A^TQA)^T_{\cdot \Lambda}(A^TQA)_{\cdot \Lambda}) > 0$.

Proof. Let $E = I_{\cdot \Lambda}$, then $A_{\cdot \Lambda} = AE$ and $(A^TQA)_{\cdot \Lambda} = A^TQA$. With the rank assumptions, the result immediately follows.

Theorem 5.2. Consider $(FP)$ with (5.13). Let $A \in \mathbb{R}^{m \times n}$ be of full row rank, $Q \in \{(AA^T), I\}$, and $f_Q$ and $L_Q$ be given by (5.5) and (5.9) respectively. Suppose there exists $\nu_s > 0$ such that

$$\nu_s \|w\|^2 \leq \|Aw\|^2, \quad \forall w \in A_s.$$  \hfill (5.16)

Then any sequence $\{w^k\}$ generated by the PDMC algorithm (5.10) with $\lambda \in (0, 1/L_Q]$ is bounded, with accumulation points lying on $\text{Fix}(T^\lambda_{PDMC})$. If $T^\lambda_{PDMC}$ is single-valued at an accumulation point $w^*$, then $w^k \to w^*$. The same conclusions hold for any sequence generated by the FB algorithm (5.11) with $\lambda \in (0, 1/L_Q)$.

Proof. Using Theorem 3.4 the convergence of (5.10) can be shown by proving that $V_i := f_Q + \text{dist}(\cdot, R_i)^2/2$ is coercive for all $i \in I_s$, where $I_s$ and $R_i$ are given in (5.14). That is, given $\{w^k\}$ such that $\|w^k\| \to \infty$, we need to show that $V_i(w^k) \to \infty$. Suppose otherwise, then $\{V_i(w^k)\}$ must have a bounded subsequence, and we assume without loss of generality that the whole sequence is bounded. Since

$$V_i(w^k) = f_Q(w^k) + \frac{1}{2}\|w^k - P_{R_i}(w^k)\|^2 = f_Q(w^k) + \frac{1}{2}\|w^k\|, \quad \forall w^k \in A_s.$$  \hfill (5.17)

$\{(w^k)_{c}\}$ must be bounded. It follows that $\|(w^k)_{c}\| \to \infty$ since we are given that $\|w^k\| \to \infty$. Meanwhile, if $Q = (AA^T)^{-1}$, we have

$$f_Q(w^k) = \frac{1}{2}\|(A^T A)_{c}c(w^k) + (A^T A)_{c}c(w^k) - A^T b\|^2 \geq \frac{1}{2}\lambda_{\min}(A^T A)\|(w^k)_{c}\|^2 + \frac{1}{2}\|(A^T A)_{c}c(w^k) - A^T b\|^2$$

$$\quad - \|(A^T A)_{c}c\| \cdot \|(w^k)_{c}\| \cdot \|(A^T A)_{c}c(w^k) - A^T b\|.$$  \hfill (5.19)

On the other hand, if $Q = I$, we obtain by a similar computation that

$$f_Q(w^k) \geq \frac{1}{2}\lambda_{\min}(A^T A)\|(w^k)_{c}\|^2 + \frac{1}{2}\|A_{c}c(w^k) - b\|^2$$

$$\quad - \|A_{c}\| \cdot \|(w^k)_{c}\| \cdot \|A_{c}c(w^k) - b\|.$$  \hfill (5.20)

By (5.16) it is clear that $\text{rank}(A_{c}) = |c|$. Thus, by Lemma 5.1 $\lambda_{\min}(A^T A) > 0$ and $\lambda_{\min}(A^T A) > 0$. Thus, letting $k \to \infty$ in (5.19) and (5.20) we obtain that $f_Q(w^k) \to \infty$, and so by (5.17) $V_i(w^k) \to \infty$, which is a contradiction. Hence, $V_i$ is coercive, as desired. Since (5.11) is applied to the same objective function (5.6) which we have just shown to be coercive, the convergence of the FB method is an immediate consequence of Theorem 3.5.

We now show the linear convergence of the PS algorithm for solving (5.7).

Theorem 5.3. Consider the setting of Theorem 5.2. Then any sequence $\{w^k\}$ generated by (5.12) with $\lambda \in (0, 1/L_Q)$ is bounded, with accumulation points lying on $\text{Fix}(T^\lambda_{PS})$. If $T^\lambda_{PS}$ is single-valued at an accumulation point $w^*$, then $w^k \to w^*$ with a local linear rate.

Proof. Given any $i \in I_s$ and any sequence $\{w^k\}$ that lies in $R_i$ such that $\|w^k\| \to \infty$, clearly $(w^k)_{c} = 0$ for all $k$. Consequently, by noting that $\lambda_{\min}(A^T A)_{c}c$ and $\lambda_{\min}(A^T A)_{c}c$ are both strictly positive from the proof of Theorem 5.2, we obtain from (5.19) and (5.20) that $f_Q(w^k) \to \infty$. Thus, $f_Q$ is
coercive over \( R_i \), showing that Assumption 3.1 (e) is fulfilled. To complete the proof, it suffices to show by Proposition 3.9 that \( \text{Id} - \lambda \nabla f_Q \) is a contraction over \( R_i \). Suppose that \( Q = (A^T A)^{-1} \) and \( w, w' \in R_i \), then

\[
\| \nabla f_Q(w) - \nabla f_Q(w') \|^2 = \| A^T A (w - w') \|^2 \\
= \| (A^T A)_{:,i} (w - w') \|^2 \\
\geq \lambda_{\min} (A^T A)_{:,i} (A^T A)_{:,i} \| (w - w') \|^2 \\
= \lambda_{\min} (A^T A)_{:,i} (A^T A)_{:,i} \| w - w' \|^2. \tag{5.21}
\]

Similarly, for \( Q = I \) and \( w, w' \in R_i \), we have

\[
\| \nabla f_Q(w) - \nabla f_Q(w') \|^2 \geq \lambda_{\min} (A^T A)_{:,i} (A^T A)_{:,i} \| w - w' \|^2. \tag{5.22}
\]

As (3.14) still holds when \( f_i \) is replaced by \( f_Q \), (5.21) and (5.22) further lead to

\[
\| (w - \lambda \nabla f(w)) - (w' - \lambda \nabla f(w')) \| \leq \kappa_i \| w - w' \|, \quad \forall w, w' \in R_i,
\]

where

\[
\kappa_i^2 = \begin{cases} 
1 + (\lambda^2 - 2\lambda) \lambda_{\min} (A^T A)_{:,i} & \text{if } Q = (A^T A)^{-1} \\
1 + (\lambda^2 - 2\lambda) \lambda_{\min} (A^T A)_{:,i} & \text{if } Q = I.
\end{cases}
\]

Since the second term is negative for \( \lambda \in (0, 1/L_Q) \), \( \kappa_i \in [0, 1) \) and the conclusion follows.

\[\square\]

### 5.2 Linear complementarity problem

We now turn our attention to the linear complementarity problem (LCP) described in [1.3] and consider the feasibility problem reformulation [FP] with [1.4]. We note that \( A \) given in [1.4] has full row rank for any matrix \( M \). Observe that \( S_1 \) is an affine set and \( S_2 \) also has some sparsity structure, as in the one considered in the previous section. In particular, \( S_2 \subset A_n \). However, \( S_2 \) has additional properties that distinguishes it from \( A_n \), including the nonnegativity of its vectors as well as the complementarity between \( (w_1, \ldots, w_n) \) and \( (w_{n+1}, \ldots, w_{2n}) \).

As shown in [Alcantara et al., 2022] Proposition 2.2, \( z \in P_{S_2}(w) \) if and only if

\[
(z_j, z_{n+j}) \in \begin{cases} 
\{(0, (w_{n+j})_+)\} & \text{if } w_j < w_{n+j}, \\
\{((w_j)_+, 0)\} & \text{if } w_j > w_{n+j}, \quad \forall j \in [n].
\end{cases}
\]

We also get from [Alcantara et al., 2022] Section 3.1 that \( S_2 \) can be decomposed as a union of closed convex sets:

\[
S_2 = \bigcup_{\iota \in \mathcal{I}} R_{\iota}, \quad \text{where } R_{\iota} := \text{Ran}(\iota, i) \cap \mathbb{R}^{2n}_+ \tag{5.23}
\]

\( \mathbb{R}^{2n}_+ \) denotes the set of nonnegative vectors in \( \mathbb{R}^{2n} \), and \( \mathcal{I} \) is the set of all \( \iota \subseteq [2n] \) expressible as \( \iota = \Lambda_1 \cup \Lambda_2 \) for some \( \Lambda_1 \subset [n] \) and \( \Lambda_2 = \{ n + j : j \in [n], j \notin \Lambda_1 \} \). It is also clear that for any \( \iota \in \mathcal{I} \) and \( w \in \mathbb{R}^{2n} \), the projection \( z \) of \( w \) onto \( R_{\iota} \) is given by

\[
(z_j, z_{n+j}) = \begin{cases} 
((w_j)_+, 0) & \text{if } j \in \iota, \\
(0, (w_{n+j})_+) & \text{if } j \notin \iota, \quad \forall j \in [n].
\end{cases} \tag{5.24}
\]

21
5.2.1 Nondegenerate LCPs

Since $S_1$ given by (1.4) is an affine set and $S_2$ is a union convex set, we may consider again the algorithms (5.10) and (5.12) for solving the feasibility reformulation of the LCP.

In Section 5.1, the condition (5.16) relaxed from SRIP was used to establish the convergence of these algorithms. For the feasibility reformulation of LCP, a property similar to (5.16) can be obtained through some assumptions on $M$ that are more conventional in the LCP literature.

Definition 5.1. A matrix $M \in \mathbb{R}^{n \times n}$ is said to be a nondegenerate matrix if all of its principal minors are nonzero. In other words, the determinant of $M_{\lambda \lambda}$ is nonzero for all $\lambda \subseteq [n]$.

Lemma 5.4. Let $M \in \mathbb{R}^{n \times n}$ be a nondegenerate matrix, and $A = [M - \mathbb{I}]$. Then there exists $\nu > 0$ such that

$$\nu \|w\|^2 \leq \|Aw\|^2, \forall w \in S_2 \cup (-S_2),$$

where $S_2$ is given by (1.4).

Proof. If $t \in \mathcal{I}$ and $w \in R_t \cup (-R_t)$, then,

$$\|Aw\|^2 = \|A_{\cdot t}w_t\|^2 \geq \lambda_{\min}(A_{\cdot t}^T A_{\cdot t}) \|w_t\|^2 = \lambda_{\min}(A_{\cdot t}^T A_{\cdot t}) \|w_t\|^2 = \nu_t \|w\|^2,$$

where $\nu_t = \lambda_{\min}(A_{\cdot t}^T A_{\cdot t})$. Meanwhile, nondegeneracy of $M$ implies that the square matrix $A_{\cdot t}$ is nonsingular by [Alcantara et al., 2022, Lemma 2.10], so that $\nu_t > 0$. By taking $\nu = \min_{t \in \mathcal{T}} \nu_t$ and noting (5.23) we get the desired inequality.

With the above lemma, we can easily obtain the convergence of PDMC (5.10), FB (5.11) and PS (5.12) for the feasibility reformulation of LCPs.

Theorem 5.5. Let $M \in \mathbb{R}^{n \times n}$ be a nondegenerate matrix, $b \in \mathbb{R}^n$, $A = [M - \mathbb{I}]$, $Q \in \{(AA^T)^{-1}, \mathbb{I}\}$, and $f_Q$ and $L_Q$ be given by (5.5) and (5.9), respectively, then for (FP) with (1.4):

(a) Any sequence $\{w^k\}$ generated by (5.10) with $\lambda \in (0,1/L_Q]$ is bounded, with accumulation points lying on $\text{Fix}(T_{\text{PD}})$. If $T_{\text{PD}}^\lambda$ is single-valued at an accumulation point $w^*$, then $w^k \to w^*$ at a linear rate. The same conclusions hold for (5.11) with $\lambda \in (0,1/L_Q)$.

(b) Any sequence $\{w^k\}$ generated by (5.12) with $\lambda \in (0,1/L_Q)$ is bounded, with accumulation points lying on $\text{Fix}(T_{\text{PS}}^\lambda)$. If $T_{\text{PS}}^\lambda$ is single-valued at an accumulation point $w^*$, then $w^k \to w^*$ at a linear rate.

Proof. Let $t \in \mathcal{I}$. We define $V_t := f_Q + \text{dist}(\cdot, R_t)^2/2$, and see from (5.24) that

$$V_t(w) = f_Q(w) + \frac{1}{2} \|w_t - [w_t]_+\|^2 + \frac{1}{2} \|w_\epsilon\|^2 \geq f_Q(w) + \frac{1}{2} \|w_\epsilon\|^2. \quad (5.25)$$

Using (5.25) and Lemma 5.4, the rest of the proof follows from arguments analogous to those in the proofs of Theorems 5.2 and 5.3.

5.2.2 LCPs involving P-matrices

To further obtain global full-sequence convergence, we consider a special class of nondegenerate matrices.

Definition 5.2. A matrix $M \in \mathbb{R}^{n \times n}$ is said to be a $P$-matrix if all of its principal minors are positive.

It is known that (1.3) has a unique solution for any $b \in \mathbb{R}^n$ when $M$ is a $P$-matrix [Cottle et al., 1992, Theorem 3.3.7]. Consequently, $S_1 \cap S_2$ contains a single point when $M$ is a $P$-matrix for $S_1$ and $S_2$ defined in (1.4). Some important applications of LCP involving $P$-matrices can be found in Schäfer [2004]. In this section, our goal is to prove the following global convergence result.
Theorem 5.6. Let $M$ be $P$-matrix, $b \in \mathbb{R}^n$, $A = [M - \mathbb{I}]$, and consider $\text{FP}$ with (1.4). Denote by $w^*$ the unique point in $S_1 \cap S_2$ and let $Q \in \{(AA^T)^{-1}, \mathbb{I}\}$, and $f_Q$ and $L_Q$ be given by (5.5) and (5.9), respectively.

(a) Any sequence generated by (5.10) with $\lambda \in (0, 1/L_Q)$ converges to $w^*$. The same conclusion holds true for any sequence generated by (5.11) with $\lambda \in (0, 1/L_Q)$.

(b) Any sequence generated by (5.12) with $\lambda \in (0, 1/L_Q)$ converges to $w^*$ with a local linear rate. Moreover, the objective function (5.7) converges to the global optimum of $0$ with a local linear rate.

Some remarks on Theorem 5.6 are in order.

Remark 5.2. (a) Since a $P$-matrix is necessarily nondegenerate, we already know from Theorem 5.5(b) that PS algorithm with $\lambda \in (0, 1/L_Q)$ generates a bounded sequence that is globally subsequentially convergent. Theorem 5.6 (b), on the other hand, indicates that the full sequence generated with any stepsize $\lambda \in (0, 1/L_Q)$ (i.e., including $\lambda = 1/L_Q$) converges linearly to $w^*$.

(b) We highlight that for a sequence generated by the PS algorithm, we obtain an additional property that the objective function values $f_Q(w^k)$ decreases to zero linearly as well.

Now we describe the flow for proving Theorem 5.6. First, as $P$-matrices are nondegenerate, global convergence in Theorem 5.6(a) will directly follow from Theorem 5.5(a) if we can show that $\text{Fix}(T_{FB}^\lambda) = S_1 \cap S_2 = \{w^*\}$. Next, we note that the linear convergence of the PS iterates, by virtue of Proposition 3.9, Theorem 5.5(b), and Remark 3.2, can be achieved by showing that

(i) $\text{Fix}(T_{PS}^\lambda) = S_1 \cap S_2$ for any $\lambda \in (0, 1/L_Q)$; and

(ii) the function

$$V(w) := f_Q(w) + \delta_{S_2}(w)$$

is a Lyapunov function for the PS for any $\lambda \in (0, 1/L_Q)$.

The characterization of the fixed point sets, which is also of independent interest, is shown below.

Theorem 5.7. Consider the setting of Theorem 5.6. If $M$ is a $P$-matrix, then

$$\text{Fix}(T_{PDMC}^\lambda) = \text{Fix}(T_{FB}^\lambda) = \text{Fix}(T_{PS}^\lambda) = S_1 \cap S_2, \quad \forall \lambda \in (0, 1/L_Q).$$

On the other hand, the next theorem proves that $V$ given by (5.26) is indeed a Lyapunov function for $T_{PS}^\lambda$ for all $\lambda \in (0, 1/L_Q)$.

Theorem 5.8. Consider the setting of Theorem 5.6. Let $w \in S_2 \setminus \text{Fix}(T_{PS}^\lambda)$ and $w^+ \in T_{PS}^\lambda(w)$ where $\lambda \in (0, 1/L_Q)$, If $M$ is a $P$-matrix, then $f_Q(w^+) < f_Q(w)$. Consequently, $V$ given by (5.26) is a Lyapunov function for the PS for any stepsize $\lambda \in (0, 1/L_Q)$.

With Theorems 5.7 and 5.8 we are now ready to prove Theorem 5.6.

Proof of Theorem 5.6. As mentioned above, part (a) is a direct consequence of Theorems 5.5 and 5.7. As for (b), linear convergence of $\{w^k\}$ to $w^*$ is a direct consequence of applying Proposition 3.9, Theorem 5.5(b) and Remark 3.2 in conjunction with Theorems 5.8 and Theorem 5.7. Thus, it remains to show that $f_Q(w^k) \to 0$ at a linear rate. First, note that the conclusions of Theorem 3.8 hold even for the stepsize $\lambda = 1/L_Q$ since Theorem 5.8 shows that (5.26) is a Lyapunov function for any $\lambda \in (0, 1/L_Q)$ (see also Remark 3.2). Hence, we know that there exists $N \geq 0$ such that $w^k \in R_i$ for all $k \geq N$ and for any $i \in I$ such that $w^* \in R_i$, where $w^*$ is the unique point in $S_1 \cap S_2$. It follows that $w^k - w^* \in S_2 \cup \{-S_2\}$ for all $k \geq N$. Suppose now that $Q = (AA^T)^{-1}$. By using Lemma 5.4 and noting that $AA^T = \mathbb{I}$, we then get

$$\nu\|w^k - w^*\|^2 \leq \|A(w^k - w^*)\|^2 = \|(AA^T)A(w^k - w^*)\|^2 \leq \|A\|^2 \cdot \|A^T A(w^k - w^*)\|^2,$$
that is,
\[ \frac{\nu}{\|A\|^2} \|w^k - w^*\|^2 \leq \|A^\dagger A (w^k - w^*)\|^2, \quad \forall k \geq N. \] (5.27)

On the other hand, if \( Q = I \), we immediately get from Lemma 5.4 that
\[ \nu \|w^k - w^*\|^2 \leq \|A(w^k - w^*)\|^2, \quad \forall k \geq N. \] (5.28)

Since \( w^* \in S_1 \), with (5.27) and (5.28), we obtain that
\[ \frac{1}{2} \|w^k - w^*\|^2 \leq \eta f_Q(w^k) \quad \forall k \geq N \] where \( \eta = \begin{cases} \frac{\|A\|^2}{\nu} & \text{if } Q = (AA^\top)^{-1} \\ \frac{1}{\nu} & \text{if } Q = I. \end{cases} \) (5.29)

Then,
\[ f_Q(w^{k+1}) \begin{aligned} &\leq \min_{z \in S_2} \frac{Q\lambda}{Q_f}(z, w^k) \\ &\leq \min_{\alpha \in [0,1]} f_Q(\alpha w^* + (1 - \alpha)w^k) + \frac{1}{2\lambda} \|w^k - (\alpha w^* + (1 - \alpha)w^k)\|^2 \\ &\leq (1 - \alpha) f_Q(w^k) + \frac{\alpha^2}{2\lambda} \|w^k - w^*\|^2, \quad \forall \alpha \in [0,1], \end{aligned} \] (5.30)

where the last inequality is from the convexity of \( f_Q \) and the fact that \( f_Q(w^*) = 0 \) since \( w^* \in S_1 \). Applying (5.29) to the inequality above then gives
\[ f_Q(w^{k+1}) \leq \left(1 - \alpha + \frac{\eta}{\lambda} \alpha^2\right) f_Q(w^k), \quad \forall \alpha \in [0,1]. \] (5.30)

The claim now follows by minimizing the right-hand side of (5.30). \( \Box \)

### 5.2.3 Acceleration schemes

When we apply the acceleration scheme described in Algorithm 2 to the algorithms (5.10)–(5.12), Theorem 2.2 only guarantees their iterates to be globally subsequentially convergent. In the case that \( M \) is a \( P \)-matrix, we show in the following theorem that we can further obtain global convergence to the unique solution for such accelerated algorithms.

**Theorem 5.9.** Under the hypotheses of Theorem 5.6, the following hold:

(a) Any sequence generated by the PDMC algorithm (5.10) with stepsize \( \lambda \in (0, 1/L_Q] \) accelerated by means of Algorithm 2 converges to \( w^* \). The same conclusion holds true for the accelerated version of the FB algorithm (5.11) with stepsize \( \lambda \in (0, 1/L_Q) \).

(b) Any sequence \( \{w^k\} \) generated by the PS algorithm (5.12) with stepsize \( \lambda \in (0, 1/L_Q] \) accelerated by means of Algorithm 2 converges to \( w^* \). Moreover, \( f_Q(w^k) \to 0 \) at a linear rate.

Notice that for the PS algorithm, we obtain a linear rate of convergence for the objective values, which together with (5.28) further shows that \( w^k \) converges to \( w^* \) at an \( R \)-linear rate.

For Algorithm 2 applied to the projected subgradient method (5.12), since the function \( f_Q \) in (5.5) is quadratic, we can get a closed form for the stepsize \( t_k \) in Algorithm 2 that will satisfy (3.19). In particular, consider the following convex quadratic function:
\[ q_k(t) := f_Q(w^k + tp^k) = \frac{1}{2} (Ap^k)^T Q(Ap^k)t^2 + \nabla f_Q(w^k)^T p^k t + f_Q(w^k). \]
Clearly, if $\nabla f_Q(w^k)^T p^k \geq 0$, $p^k$ is not a descent direction, so there is no positive number $t$ such that $q_k(t) < q_k(0)$. On the other hand, if $\nabla f_Q(w^k)^T p^k < 0$, we can easily verify that

$$q_k(t) \leq q_k(0) - \frac{\sigma}{2} \|p^k\|^2 t^2,$$

\forall t \text{ such that } 0 \leq t \leq \frac{-2 \nabla f_Q(w^k)^T p^k}{(Ap^k)^T Q (Ap^k) + \sigma \|p^k\|^2} =: t_k^{(2)}.

(5.31)

On the other hand, as the objective function is actually $f_Q + \delta_{S_2}$, we also need to ensure that $w^k + tp^k \in S_2$ by projecting $t$ to a suitable range. As we conduct extrapolation at the $k$th iteration only when there is $\iota \in I$ such that $w^{k-1}, w^k \in R_\iota$, clearly for $t \geq 0$, we just need

$$t \leq t_k^{(1)} := \min \left\{ -w_j^k/p_j^k : p_j^k < 0 \right\}.$$

(5.32)

Such a closed-form step size then results in Algorithm 4, whose convergence guarantees directly follow from Theorem 5.9 (b).

**Algorithm 4:** Accelerated projected subgradient algorithm for LCP

Choose $\sigma > 0$, $w^0 \in S_2$, $Q \in \{(AA^T)^{-1}, I\}$ and $\lambda \in (0, 1/L_Q]$, where $L_Q$ is given by (5.9) and $A$ by (1.4). Set $w^{-1} = w^0$ and $k = 0$.

**Step 1.** Set $z^k = w^k + t_k \chi_k p^k$, where $p^k = w^k - w^{k-1}$,

$$\chi_k = \begin{cases} 1 & \text{if } \exists \iota \in I \text{ such that } w^{k-1}, w^k \in R_\iota, \\ 0 & \text{otherwise} \end{cases},$$

and $t_k = \chi_k \max\{0, \min\{t_k^{(1)}, t_k^{(2)}\}\}$, with $t_k^{(1)}, t_k^{(2)}$ defined in (5.31) and (5.32).

**Step 2.** Set $w^{k+1} \in T^\lambda_{PS}(z^k)$, $k = k + 1$, and go back to Step 1.

**Remark 5.3.** For the sparse affine feasibility problem, we can use the same algorithm as above, but we need not compute $t_k^{(1)}$ since $R_\iota$ given by (5.14) is a subspace.

We end this section with the following remarks with regard to the acceleration technique by component identification as described in Algorithm 3.

**Remark 5.4.** Similar to Remark 4.2 (b), we note that latter iterations of the algorithms (5.10) (5.12) indicate which $\iota \in I$ can be used to reduce the original problems (5.6) and (5.7) into the simpler problem of finding a point in $S_1 \cap R_\iota$. For SAFP, the latter problem amounts to solving the linear system $A_* w_* = b$, while for the LCP case, finding $S_1 \cap R_\iota$ is equivalent to solving the system $Aw^* = b$ and $w_{\Lambda,c} = 0$, which is simply an $n \times n$ system of linear equations. If the obtained solution satisfies $w^*_i \geq 0$, then $w^*$ is indeed a solution of the original feasibility problem. For (5.10) (5.12) and their extrapolation-accelerated versions by Algorithm 2 we know from Theorems 5.6 and 5.9 that these algorithms will converge to the unique point $w^*$ in $S_1 \cap S_2$ when $M$ is a $P$-matrix. Thus, theoretically, we know that Algorithm 3 described will indeed output the solution $w^*$.

6 Related works

We now contrast the results presented in the foregoing sections with existing literature on related topics.

**Fixed point algorithms.** A rigorous treatment of the relationship between Lyapunov functions and convergence behavior of fixed point iterations of an upper semicontinuous map $T$ can be found in Kellet and Teel 2005. As opposed to the Lyapunov functions defined in Kellet and Teel 2005, where the goal is to establish the “robust stability” of the fixed point iterations, the notion we introduced in Definition 2.1.
is substantially more relaxed. In particular, we only require strict monotonicity at a point not in \( \text{Fix}(T) \), since this property is enough to establish the global subsequential convergence of the algorithm to a fixed point of \( T \) (Theorem 2.1). In addition, we extend our focus to those operators \( T \) that are union upper semicontinuous, which are not necessarily upper semicontinuous themselves. Moreover, due to practical interests, an accelerated fixed point algorithm with guaranteed global subsequential convergence is also provided in our work.

Similar classes of operators \( T \) that can be expressed as \( (1.2) \) are studied in \cite{Dao and Tam, 2019, Tam, 2018}. The setting in the said works involve nonexpansive and paracontracting mappings \( T_i \) (and therefore continuous), in which case, local convergence of the fixed point algorithm \((\text{FPA})\) was established. With the aid of Lyapunov functions, we have significantly extended this framework to include upper semicontinuous mappings \( T_i \) that need not be nonexpansive but are calm at accumulation points (Theorem 2.4).

**Optimization.** The class of min-\( \rho \)-convex functions introduced in Definition 3.1 is a generalization of the class of min-convex functions defined in \cite{Dao and Tam, 2019}. In particular, this generalization extends to taking the minimum of a finite number of weakly-convex functions in the case that \( \rho < 0 \).

The possibly nonsmooth and nonconvex optimization problem \((\text{OP})\) under Assumption 3.1 along with \( \text{(PDMC)} \) provides a unified setting for several other optimization problems and algorithms in the literature. In particular, it covers several scenarios corresponding to the case of \(|I| = 1|:

\begin{enumerate}
    \item When \( g \equiv h \equiv 0 \), the PDMC algorithm coincides with the classical method of steepest descent. A more general setting is when \( g = \delta_S \) and \( h \equiv 0 \) where \( S \) is a convex set, in which case PDMC algorithm reduces to the traditional projected gradient algorithm. The global convergence of these algorithms hold under the usual Lipschitz continuity property of the gradient of \( f \), as in Assumption 3.1(b).

    \item When \( g \) is min-convex and \( h \equiv 0 \), PDMC simplifies to the forward-backward algorithm studied in \cite{Dao and Tam, 2019, Assumption 3.1 (b)}. Local convergence of this algorithm is established in the said work under Assumption 3.1(b).

    \item If \( f \) and \( g \) are convex, and \(|I| = |M| = 1\), PDMC reduces to the proximal difference-of-convex algorithm studied in \cite{Wen et al., 2018} with the exception that our setting requires the differentiability of \( h \), as in Assumption 3.1(d). For this algorithm, Assumption 3.1 (b) is also important to guarantee global subsequential convergence.
\end{enumerate}

On the other hand, to the best of our knowledge, there are no current works handling the case \(|I| > 1|.

Moreover, the acceleration schemes provided in Section 3.3 which fully utilizes the piecewise structure of the functions involved, are the first of their kinds for nonconvex optimization problems.

**Nonconvex feasibility problems.** Due to difficulties that come with nonconvexity, existing body of literature on projection algorithms for solving nonconvex feasibility problems mainly focuses on local convergence when the algorithms are initialized close enough to a solution. For instance, the local convergence of MAP for the intersection of union convex sets was established in \cite{Dao and Tam, 2019}. In fact, using the same framework, one can also obtain local convergence of MAveP and MARP presented in Sections 4.1 and 1.2 respectively. Local linear convergence of MAP and MAveP to solutions of the feasibility problem was studied in \cite{Lewis et al., 2009} using the notion of strong regularity. On the other hand, global convergence for these algorithms on nonconvex sets largely remained unknown, and our present work provides sufficient conditions to attain global convergence to candidate solutions of the feasibility problem involving union convex sets. Furthermore, we also established local linear convergence of MAP under a Lipschitz continuity assumption that is more easily verifiable and sometimes weaker than strong regularity. For example, for the feasibility formulation of LCP, in the course of proving Theorem 5.6 in Appendix D, the proof of Proposition D.2 shows that if \( w^* \) is nondegenerate and \( w^* \in S_1 \cap S_2 \), then \( S_1 \) and \( S_2 \) have a “linearly regular intersection at \( w^* \)” as defined in \cite{Lewis et al., 2009} (see also the proof of \cite{Alcantara et al., 2022} Theorem 3.19). Consequently, local linear convergence to \( w^* \) follows from \cite{Lewis et al., 2009} Theorem 5.16 provided that \( w^* \in S_1 \cap S_2 \) is nondegenerate and \( M \) is a nondegenerate matrix. However, nondegeneracy of \( w^* \) is essential to guaranteeing
that result, but this is not verifiable a priori. \textbf{Theorem 5.6} on the other hand, asserts that this linear rate is achieved whether or not $w^*$ is nondegenerate.

**Sparse affine feasibility problem.** The SRIP condition defined in (5.15) is introduced by Beck and Teboulle [2011]. They showed that when SRIP of order $(d, \alpha)$ with $d = 2s$ holds, the SAFP has a unique solution provided that $S_1 \cap S_2 \neq \emptyset$, and if $\alpha = \sqrt{2}$ additionally, global and linear convergence of a projected gradient algorithm to the unique solution of the SAFP holds. This algorithm coincides with the PS algorithm (5.12) with $Q = I$ and some specific $\lambda$ dependent on the SRIP parameters $\mu_2s$ and $\nu_2s$. On the other hand, if $\alpha = 2$, it is shown in Hesse et al. [2014] that MAP is globally convergent to the unique solution provided that $AA^T = I$. In Section 5.1 we provided a unified setting to analyze these two algorithms. Moreover, we have shown in \textbf{Theorem 5.3} that under significantly weaker assumptions than SRIP of some order, we obtain global convergence to candidate solutions of the SAFP. Finally, we also remark that the algorithms PDMC (5.10) and FB (5.11) are new methods in the SAFP literature. In addition, the acceleration schemes described in Section 3.3 applied to PDMC, FB and PS are also novel contributions to this topic.

**Linear complementarity problem.** There are two well-known projection-based algorithms for solving complementarity problems, namely the \textit{basic projection algorithm} (BPA) and the \textit{extragradient algorithm} (EGA) (see Facchinei and Pang [2003, Algorithm 12.1.1 and 12.1.9]). BPA is suitable for problems where the matrix $M$ associated with the LCP (1.3) is a positive definite matrix, in the sense that $x^T M x > 0$ for all nonzero vector $x$. On the other hand, EGA can handle positive semidefinite LCPs, i.e., $x^T M x \geq 0$ for all $x$. Meanwhile, all the algorithms we have proposed in Section 5.2 are new projection methods for LCP with guaranteed global subsequential convergence for nondegenerate LCPs and guaranteed global full convergence to the solution set for $P$-LCPs. The classes of nondegenerate and $P$-matrices both include the set of positive definite matrices, and therefore the proposed methods can solve those LCPs that are in the scope of BPA. On the other hand, both the sets of nondegenerate and $P$-matrices contain matrices that are not positive semidefinite, and are therefore solvable by our proposed approaches but not by EGA.

7 Numerical experiments

This section presents numerical experiments on sparse affine feasibility and linear complementarity problems to support the established theoretical convergence of the proposed algorithms. We also demonstrate that the proposed acceleration schemes [Algorithms 2 and 3] can significantly improve the efficiency of the non-accelerated algorithms. All methods compared are implemented in MATLAB, and all experiments are conducted on a machine running Ubuntu 20.04 and MATLAB R2021b with 64GB memory and an Intel Xeon Silver 4208 CPU with 8 cores and 2.1 GHz.

\textbf{Nomenclature for algorithms.} As mentioned in Section 5, PDMC, FB, and PS respectively reduce to MAveP, MARP, and MAP when $Q$ is set to $(AA^T)^{-1}$. On the other hand, when $Q = I$, we use the original notations PDMC, FB, and PS to denote the algorithms. We include a prefix “A” and/or a suffix “+” to signify that Algorithm 2 and/or Algorithm 3 are incorporated in the algorithms.

7.1 Sparse affine feasibility problem

We use our proposed algorithms to solve SAFP with synthetic and real datasets described below. Comparisons with the proximal gradient method by Beck and Teboulle [2011], which we denote by PG-BT, are also presented.

The stepsizes of PDMC/MAveP, FB/MARP and PS/MAP are set to $\lambda = \tau/L_Q$ with $\tau = 1.0999, 0.999$, respectively (see Theorems 5.2 and 5.3). The line search parameter in Algorithm 2 is set to $\sigma = 10^{-2}$. The parameter $N$ in Algorithm 3 is set to 50 and 100 when $Q = (AA^T)^{-1}$ and $Q = I$, respectively, and the linear

\footnote{Symmetric $P$-matrices must be positive definite, but nonsymmetric $P$-matrices might have all principal minors positive while being indefinite. See Cottle et al. [1992, Example 3.3.2] for an example.}
system described in [Remark 5.4] for dealing with Step 2.2 of Algorithm 3 is handled by solving
\[ A^\top_i A_{i,:} w_i = A_{i,:} \; i^\top b \]
by the conjugate gradient (CG) method (see, for example, [Nocedal and Wright 2006, Chapter 5]). We note that for the SAFP problems, the cost of one CG iteration is \( O(ms) \) and the number of CG iterations in one round of Step 2.2 of Algorithm 3 is upper bounded by \( s \), so the overall cost of invoking the CG procedure once is at most \( O(ms^2) \) although in practice we often observe that CG terminates within few iterations, while one step of \( \text{(PDMC)} \) is \( O(mn) \). In the experiments, we also observe that empirically the CG procedure takes almost negligible amount of running time in the whole procedure. When Algorithm 3 is used in combination with Algorithm 2, we set \( N \) to half its specified value when only Algorithm 3 is used. All algorithms are initialized with \( w^0 = A^\top b \), and the residual is measured by
\[ \text{Residual} := \frac{1}{2} \| Aw^k - b \|^2 + \frac{1}{2} \text{dist}(w^k, S_2)^2, \]
which is 0 if and only if \( w^k \) is a solution to the SAFP.

**Synthetic data.** We generate standard random test problems involving a matrix \( A \in \mathbb{R}^{m \times n} \) with entries sampled from the standard normal distribution, and an \( s \)-sparse signal \( w^* \) such that the nonzero entries \( w^*_i \) are generated as \( w^*_i = \eta_1 10^{-m_2} \) with \( \alpha = 5 \), \( \eta_1 = \pm 1 \) with probability 0.5, and \( \eta_2 \) is uniformly generated in \([0, 1] \) [Becker et al., 2011]. After generating \( A \) and \( w^* \), we set \( b = Aw^* \) so that \( w^* \) is a solution of the SAFP. The results of running time and total iterations required for reducing (7.1) below \( 10^{-6} \) for ten independent trials with \( n = 10000 \), \( m = 2500 \), and \( s = 625 \) are summarized in Table 1, where “Average CI Iterations” refers to the average number of times Step 2.2 of Algorithm 3 is executed, while “Average CI Time” pertains to the average time required to execute the said component identification (CI) step.

We see from Table 1 that the acceleration schemes Algorithms 2 and 3 reduce both the running time and the number of iterations of the algorithms. Notice that the algorithms corresponding to \( Q = (AA^\top)^{-1} \) are superior to those that correspond to \( Q = I \) in this experiment. Moreover, the non-accelerated MARP and MAP algorithms are already more efficient than PG-BT, while for MAveP, its accelerated versions, namely AMAveP and AMAveP+, provide better performance. On the other hand, when \( Q = I \), only the accelerated versions of PS have better performance than PG-BT. Finally, we observe that for this experiment, Algorithm 2 has faster convergence than Algorithm 3 and incorporating component identification to Algorithm 2 only resulted to minimal improvements in convergence time. Component identification in this experiment mainly only helped to reduce the residual to a much lower level after the stopping criterion of \( 10^{-6} \) is almost reached.

**Real-world datasets.** We then consider three public real-world datasets [4]. They are: colon-cancer \((m = 62, n = 2000)\), duke breast-cancer \((m = 44, n = 7129)\) and leukemia \((m = 38, n = 7129)\). We set \( s \) to 5% of the total number of features \( n \). A key difference between this experiment and the previous one is that for real-world data, when \( s < m \), we are not guaranteed that there is a solution to (FP) due to the existence of noise in the data. Therefore, \( s > m \) is needed in this case, while we can safely assign \( s < m \) in the previous experiment. The results are summarized in Figs. 2 and 3.

Similar to the results on synthetic datasets, the acceleration schemes Algorithms 2 and 3 reduce both the running time and the number of iterations of the algorithms, except for the duke breast-cancer dataset where component identification in Algorithm 3 did not take place. The algorithms corresponding to \( Q = (AA^\top)^{-1} \) also provided performance better than when \( Q \) is set to \( Q = I \), or when PG-BT is used.

### 7.2 Linear complementarity problem

We consider some standard LCP test problems as follows.

**LCP1.** [Qi et al., 2000, Example 2] \( M \) is a tridiagonal matrix with \( M_{ii} = 4 \) for all \( i \in [n] \) and \( M_{ij} = -1 \) when \( |i - j| = 1 \), and \( b = (1, 1, \ldots, 1)^\top \).

Table 1: Performance of non-accelerated and accelerated algorithms for ten independent trials of SAFP with synthetic data. For the running time and residual of each method, we report their average±standard deviation. The blank entries indicate that the algorithm reached the maximum number of iterations (namely 10000) but did not obtain a solution with Residual of at most $10^{-6}$. Ave. CI Iters. refers to the average number of times Step 2.2 in Algorithm 3) is run, while Ave. CI Time is the average amount of time required to finish one CI iteration.

<table>
<thead>
<tr>
<th>Method</th>
<th>Ave. Time (seconds)</th>
<th>Ave. Ave. Residual</th>
<th>Ave. CI Iters</th>
<th>Ave. CI Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAveP</td>
<td>2793.7</td>
<td>42.5 ± 1.7</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AMAveP</td>
<td>722.4</td>
<td>11.9 ± 0.4</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>MAveP+</td>
<td>2396.9</td>
<td>36.5 ± 1.7</td>
<td>3.5</td>
<td>0.015</td>
</tr>
<tr>
<td>AMAveP+</td>
<td>695.0</td>
<td>11.3 ± 0.4</td>
<td>1</td>
<td>0.014</td>
</tr>
<tr>
<td>APDMC</td>
<td>8076.1</td>
<td>133.6 ± 5.0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>APDMC+</td>
<td>7639.8</td>
<td>126.3 ± 4.7</td>
<td>1.8</td>
<td>0.014</td>
</tr>
<tr>
<td>MARP</td>
<td>779.5</td>
<td>12.7 ± 0.2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AMARP</td>
<td>299.6</td>
<td>5.4 ± 0.1</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>MARP+</td>
<td>592.2</td>
<td>9.7 ± 0.2</td>
<td>1</td>
<td>0.014</td>
</tr>
<tr>
<td>AMARP+</td>
<td>287.1</td>
<td>5.2 ± 0.1</td>
<td>1</td>
<td>0.015</td>
</tr>
<tr>
<td>AFB</td>
<td>8246.7</td>
<td>140.9 ± 5.0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AFB+</td>
<td>7763.7</td>
<td>123.2 ± 2.3</td>
<td>2.5</td>
<td>0.014</td>
</tr>
<tr>
<td>MAP</td>
<td>673.6</td>
<td>10.6 ± 0.2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>AMAP</td>
<td>263.4</td>
<td>4.5 ± 0.3</td>
<td>NA</td>
<td>7.8e-07</td>
</tr>
<tr>
<td>MAP+</td>
<td>600.1</td>
<td>9.5 ± 0.2</td>
<td>1.2</td>
<td>0.014</td>
</tr>
<tr>
<td>AMAP+</td>
<td>250.1</td>
<td>4.3 ± 0.3</td>
<td>1</td>
<td>0.014</td>
</tr>
<tr>
<td>APS</td>
<td>417.5</td>
<td>5.7 ± 0.3</td>
<td>NA</td>
<td>8.4e-07</td>
</tr>
<tr>
<td>APS+</td>
<td>402.9</td>
<td>5.5 ± 0.3</td>
<td>1</td>
<td>0.014</td>
</tr>
<tr>
<td>PG-BT</td>
<td>847.0</td>
<td>15.3 ± 0.6</td>
<td>NA</td>
<td>9.5e-07</td>
</tr>
</tbody>
</table>

LCP2. [Kanzow, 1996] Example 7.1] $M$ is an upper triangular matrix with $M_{ii} = 1$ for all $i \in [n]$, $M_{ij} = 2$ for all $i < j$, and $b = (1, 1, \ldots, 1)^T$. Since BPA is only applicable when $M + M^T$ is positive definite [Facchinei and Pang, 2003, Theorem 12.1.2], we will only compare our algorithm with EGA in this case.

LCP3. [Kanzow, 1996] Example 7.3] $M$ and $b$ are generated as follows: The entries of $b$ are independently sampled from uniform random with range $(-500, 500)$. $M$ is a $P$-matrix given by $M = A_1^T A_1 + A_2 + \text{diag(\eta)}$, where $A_1, A_2 \in \mathbb{R}^{n \times n}$ are matrices with entries independently sampled from uniform random in $(-5, 5)$, $A_2$ is skew-symmetric, and each entry of $\eta \in \mathbb{R}^n$ is independently taken from uniform random of $(0, 0.3)$.

We set the stepsizes of MAveP, MARP, and MAP to $\lambda = 1, 0.999, 1$ (see [Theorem 5.6]) and set the values of the parameters $s$ and $N$ as in the preceding section. Matlab’s backslash operator is used to handle the linear system described in Remark 5.4] with the cost of $O(n^3)$ for our problem, which is of the same order as the overhead of computing $Q$ and $L_Q$. On the other hand, the cost of one iteration of [PDMC] is $O(n^2)$. We will see in the experimental results that although component identification in this case is slightly more expensive than its counterpart in the SAFP experiment, it still takes only a small portion of the overall running time of the algorithms. We set $n = 5000$ in all of the experiments, and divide both $M$ and $b$ by the same scalar $\|M\|_1/\sqrt{n}$. This normalization is due to the geometric observation that for $n = 1$, projection algorithms tend to converge faster to a solution when the slope is in a moderate range.

We compare our algorithms with BPA and EGA mentioned in [Section 6]. Instead of $(7.1)$, we use the following standard measure of residual in LCP [Facchinei and Pang, 2003, Proposition 1.5.8] to facilitate fair
Figure 2: Comparisons of running time of non-accelerated and accelerated algorithms for solving SAFP with real-world datasets.

comparisons with these two algorithms:

\[
\text{Residual} := \| \min(x^k, Mx^k - b) \|. \tag{7.2}
\]

We report the running time and iterations required for reducing (7.2) below $10^{-6}$. We note that for the feasibility reformulation of the LCP (see Section 5.2), the first $n$ coordinates of $w^k$ correspond to $x^k$.

We see from Figs. 4 and 5 and Table 2 that indeed, the proposed acceleration schemes significantly reduce the required time and number of iterations to solve the generated LCPs. The only exception is LCP1, on which the non-accelerated algorithms with $Q = (AA^T)^{-1}$ already require only few iterations to solve the
problem to the desired accuracy, so there is no much room for improvement. In particular, the stopping condition of (7.2) being smaller than $10^{-6}$ is met before our specified $N$ for Algorithm 2 for these algorithms, so component identification is not executed at all. For LCP1, we see that the proposed methods are very competitive against BPA and significantly better than the relatively slow EGA. For LCP2, EGA and the proposed algorithms corresponding to $Q = I$ have very slow convergence, while those that correspond to $Q = (AA^T)^{-1}$ and accelerated by Algorithm 2 either with or without component identification, showed good performance. Finally, for LCP3, EGA has a better performance than the base forms of our proposed algorithms, but is outperformed by our acceleration by extrapolation schemes, again either with or without component identification, along with $Q = (AA^T)^{-1}$. For those methods with component identification, the

Figure 3: Comparisons of iterations of non-accelerated and accelerated algorithms for solving SAFP with real-world datasets.
residual of their final outputs presented in Table 2 tend to be much lower than our stopping condition. The reason is that the linear system solver is non-iterative and cannot be terminated exactly at the point where the required residual tolerance is met. This is especially prominent in the residuals of those algorithms that conduct component identification without extrapolation. Interestingly, for the extrapolation schemes, the addition of component identification does not change the residual much. A closer examination revealed that in such cases, component identification is sometimes not triggered, making the algorithms that combined both extrapolation and component identification empirically identical to those with extrapolation only.

Overall speaking, the proposed acceleration scheme in Algorithm 1 using extrapolation is indeed very effective in reducing the running time and iterations of fixed-point maps, while the component identification part in Algorithm 3 is more useful when highly-accurate solutions are required.

8 Conclusion

In this work, we analyze the global subsequential convergence of fixed point iterations of union upper semicontinuous operators, and prove global convergence under a local Lipschitz condition. We show that this class of fixed point algorithms in fact covers several iterative methods for solving optimization and feasibility problems alike, and therefore global convergence of these methods is a consequence of the derived theory for the general setting of fixed point problems. In particular, we establish the global convergence of proximal algorithms for minimizing a class of nonconvex nonsmooth functions, specifically those that can be expressed as the sum of a piecewise smooth mapping and another function that is the difference of a
min-$\rho$-convex and a convex function. Linear convergence is also proven under a mild calmness condition. We also prove global convergence of traditional projection methods for solving feasibility problems involving union convex sets. Acceleration methods via extrapolation and component identification are proposed by utilizing the special structure of the defining operators of the algorithms. Using these results, we derive several algorithms for solving sparse affine feasibility and linear complementarity problems that are new to the literature. Numerical evidence illustrated the empirical performance of the proposed algorithms, and our proposed acceleration schemes indeed provide significant improvement over the non-accelerated ones in terms of the running time and the number of iterations required to solve the problems. In the near future, we plan to further study how to better initialize the step size in the extrapolation step. Another interesting future work is to obtain an iteration bound for the component identification result, and then to further develop global iteration complexities of the discussed algorithms on top of the identification bound.

References


Table 2: Performance of non-accelerated and accelerated algorithms for ten independent trials with LCP3 for reducing the residual to below $10^{-6}$. The proposed accelerated and non-accelerated algorithms corresponding to $Q = I$ are omitted as all of them failed to make the residual below $10^{-6}$ in 10000 iterations. Ave. CI Iters. refers to the average number of times Step 2.2 in Algorithm 3 is run, while Ave. CI Time is the average amount of time required to finish one CI iteration. The average residuals of AMAveP+, AMARP+ and AMAP+ are larger than that of MAveP+, MARP+ and MAP+, as the former algorithms did not enter Step 2.2 of Algorithm 3 for some test problems (as average CI is smaller than 1), i.e., no component identification occurred.

<table>
<thead>
<tr>
<th>Method</th>
<th>Ave. Iters</th>
<th>Time (seconds)</th>
<th>Ave. CI Iters</th>
<th>Ave. CI Time</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAveP</td>
<td>3909.6</td>
<td>79.5 ± 5.3</td>
<td>NA</td>
<td>NA</td>
<td>1.0e-06 ± 7.0e-10</td>
</tr>
<tr>
<td>AMAveP+</td>
<td>504.4</td>
<td>11.6 ± 0.4</td>
<td>NA</td>
<td>NA</td>
<td>9.8e-07 ± 2.2e-08</td>
</tr>
<tr>
<td>MAveP</td>
<td>2209.9</td>
<td>64.6 ± 9.6</td>
<td>16.4</td>
<td>1.2</td>
<td>2.2e-15 ± 1.0e-16</td>
</tr>
<tr>
<td>AMAveP+</td>
<td>451.4</td>
<td>11.5 ± 0.5</td>
<td>0.9</td>
<td>1.2</td>
<td>1.3e-07 ± 3.2e-07</td>
</tr>
<tr>
<td>MARP</td>
<td>1956.0</td>
<td>40.8 ± 2.6</td>
<td>NA</td>
<td>NA</td>
<td>1.0e-06 ± 1.8e-09</td>
</tr>
<tr>
<td>AMARP</td>
<td>355.0</td>
<td>8.6 ± 1.2</td>
<td>NA</td>
<td>NA</td>
<td>9.7e-07 ± 3.3e-08</td>
</tr>
<tr>
<td>MARP+</td>
<td>1116.8</td>
<td>31.2 ± 4.3</td>
<td>6.3</td>
<td>1.2</td>
<td>2.2e-15 ± 1.0e-16</td>
</tr>
<tr>
<td>AMARP+</td>
<td>306.9</td>
<td>8.7 ± 0.6</td>
<td>0.9</td>
<td>1.2</td>
<td>8.9e-08 ± 2.8e-07</td>
</tr>
<tr>
<td>MAP</td>
<td>979.0</td>
<td>20.9 ± 1.3</td>
<td>NA</td>
<td>NA</td>
<td>1.0e-06 ± 2.3e-09</td>
</tr>
<tr>
<td>AMAP</td>
<td>244.1</td>
<td><strong>6.3 ± 0.4</strong></td>
<td>NA</td>
<td>NA</td>
<td>8.8e-07 ± 1.2e-07</td>
</tr>
<tr>
<td>MAP+</td>
<td>577.1</td>
<td>16.1 ± 2.1</td>
<td>2.8</td>
<td>1.2</td>
<td>2.2e-15 ± 1.0e-16</td>
</tr>
<tr>
<td>AMAP+</td>
<td>358.0</td>
<td>7.1 ± 0.5</td>
<td>0.8</td>
<td>1.2</td>
<td>1.4e-07 ± 2.9e-07</td>
</tr>
<tr>
<td>BPA</td>
<td>–</td>
<td>–</td>
<td>NA</td>
<td>NA</td>
<td>–</td>
</tr>
<tr>
<td>EGA</td>
<td>914.4</td>
<td>11.8 ± 0.4</td>
<td>NA</td>
<td>NA</td>
<td>9.9e-07 ± 7.5e-09</td>
</tr>
</tbody>
</table>


A Importance of Assumption 1.1

As mentioned in Section 1, Assumption 1.1 is critical in ensuring that limits of convergent sequences generated by (FPA) are fixed points of $T$. We further demonstrate its importance in the following example.

Example A.1.

1. Let $E = \mathbb{R}^2$, $S_1 = \{(a,0) : a \in \mathbb{R}\}$ and $S_2 = \{(a,a) : a \neq 0\}$. Consider

$$T(w) := \begin{cases} (P_{S_2} \circ P_{S_1})(w) & \text{if } w \in D_1 = \{(a,b) : a \neq 0\}, \\ \{\bar{w}\} & \text{if } w \in D_2 = \{(0,b) : b \in \mathbb{R}\}, \end{cases}$$

where $\bar{w}$ is some fixed nonzero point in $E$. Here, each $T_j$ is continuous on the set $D_j$, where $D_2$ is closed but $D_1$ is an open set, so the closedness requirement in Assumption 1.1 is not met. Given any $w^0 = (w_0^0, w_0^1)$ with $w_0^1 \neq 0$, it is clear that the sequence generated by (FPA) converges to 0, but $0 \notin T(0) = \{\bar{w}\}$, that is, $0 \notin \text{Fix}(T)$.

2. Let $D_0 = \{0\}$, $D_1 = \{w : \|w\| \geq 1\}$, and $D_{j+1} = \{w \in E : \frac{1}{j+1} \leq \|w\| \leq \frac{1}{j}\}$ for all $j \geq 1$ so that $E = \bigcup_{j=0}^{\infty} D_j$. Define $T$ by

$$T(w) = \begin{cases} P_{D_j}(w) = \left\{ \frac{w}{j\|w\|} \right\} & \text{if } w \in D_j, j \geq 1 \\ \{\bar{w}\} & \text{if } w \in D_0 \end{cases},$$

where $\bar{w} \neq 0$ is fixed. Observe that all $D_j$’s are closed, and each $T_j$ on $D_j$ is continuous, but the finiteness requirement on the collection of sets $D_j$ in Assumption 1.1 is violated. We then see that given any $w^0 \neq 0$, (FPA) generates a sequence that converges to $0 \notin T(0) = \{\bar{w}\}$.

We prove in Lemma B.1 that under Assumption 1.1 any accumulation point of a convergent sequence produced by (FPA) if such exists, indeed belongs to $\text{Fix}(T)$.

B Proofs of results in Section 2

We first provide a formal definition of upper semicontinuity.

Definition B.1. [Aubin and Frankowska, 2009, Definition 1.4.1] Let $X,Y$ be metric spaces, $F : X \rightrightarrows Y$ be a set-valued map, and $w \in X$ be a point such that $F(w) \neq \emptyset$. Then $F$ is said to be upper semicontinuous (usc) at $w \in X$ if for any neighborhood $U$ of $F(w)$, there exists $\delta > 0$ such that for all $z \in X$ with $\|z - w\| < \delta$, we have $F(z) \subseteq U$. Moreover, $F$ is said to be usc if it is usc at each point in $X$.

Remark B.1. Consider $w \in X$ such that $F$ is usc at $w$ and $F(w)$ is compact, and let $\{w^k\} \subseteq X$ be such that $w^k \to w$. From the above definition, we know that given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|z - w\| < \delta$ implies $F(z) \subseteq B(F(w), \varepsilon)$. Since $w^k \to w$, there exists $N > 0$ such that $\bigcup_{k=N}^{\infty} F(w^k) \subseteq B(F(w), \varepsilon)$. Thus, any sequence $\{z^k\}_{k=N}^{\infty}$ with $z^k \in F(w^k)$ is bounded, and we may extract a subsequence $\{z^{k_j}\}_{j=0}^{\infty}$ such that $z^{k_j} \to z$ for some $z \in Y$. Since $F$ is usc at $w$ and $F(w)$ is bounded, it is not difficult to verify from the definition that $z \in F(w)$.

B.1 Proofs of Theorems 2.1 and 2.2

Note that the fixed point algorithm (FPA) is a special case of Algorithm 1 with $t_k \equiv 0$ in Step 1, and hence Theorem 2.1 will follow directly from Theorem 2.2. To establish the results, we need the following two lemmas.
Lemma B.1. Let \( \{w^k\} \) be a sequence generated by Algorithm 1. If \( w^* \) is an accumulation point of this sequence, there exists \( (w^*)^+ \in T(w^*) \) such that \( (w^*)^+ \) is also an accumulation point of \( \{w^k\} \). In particular, if \( w^k \to w^* \), then \( w^* \in \text{Fix}(T) \).

Proof. Let \( \{w^{kj}\}_{j=0}^\infty \) be a subsequence of \( \{w^k\} \) such that \( w^{kj} \to w^* \). Now consider \( \{w^{kj+1}\}_{j=0}^\infty \), where \( w^{kj+1} \in T(z^{kj}) \). Since the index set \( I \) is finite, there exists \( i \in I \) and a subsequence \( \{w^{kjr}\}_{r=0}^\infty \) of \( \{w^{kj+1}\}_{j=0}^\infty \) such that \( w^{kjr+1} \in T(z^{kjr}) \). Then by the definition of \( T \), we have \( \{z^{kjr}\}_{r=0}^\infty \subseteq D_i \). Moreover, we note from (2.2) and Definition 2.1 that

$$
\frac{\sigma}{2}t_{kj}^2 \|p^{kj}\|^2 \leq V(w^{kj+1}) - V(w^{kj+1+1}).
$$

By summing the inequality above from \( r = 0 \) to infinity, we see that the monotonicity (from the algorithm) and the lower-boundedness of \( \iota \) imply \( t_{kj} \chi_{kj} \|p^{kj}\| \to 0 \), so \( z^{kj} \to w^* \). Therefore, by the closedness of \( D_i \), we get \( w^* \in D_i \) so that \( T_i(w^*) \subseteq T(w^*) \). Moreover, since \( T_i \) is usc at \( w^* \), we have from Remark B.1 that \( \{w^{kjr+1}\}_{r=0}^\infty \) has a subsequence converging to some \( (w^*)^+ \in T(w^*) \). Thus, \( (w^*)^+ \in T(w^*) \), as desired. Finally, if \( w^k \to w^* \), we must have \( w^* = (w^*)^+ \in T(w^*), \) i.e., \( w^* \in \text{Fix}(T) \).

Proof of Theorem 2.2. First, for any generated sequence \( \{w^k\} \) of (FPA) since \( \{V(w^k)\} \) is monotonically decreasing by (2.1) and \( V \) is bounded below, \( \{V(w^k)\} \) converges to a finite value. If \( w^* \) is an accumulation point of \( \{w^k\} \), we have from Lemma B.1 that there exists another accumulation point \( (w^*)^+ \in T(w^*) \) of \( \{w^k\} \). Since \( \{V(w^k)\} \) is convergent, by taking the corresponding subsequences of \( \{w^k\} \) that converge to \( w^* \) and \( (w^*)^+ \), we must have \( V(w^*) = V((w^*)^+) \) by the continuity of \( V \). By Definition 2.1 we then conclude that \( w^* \in \text{Fix}(T) \).

B.2 Proof of Lemma 2.3

Proof. Given \( \iota \notin \phi_D(w^*) \), there exists \( \delta > 0 \) such that \( B(w^*, \delta) \cap D_i = \emptyset \). Otherwise, we can construct a sequence \( \{w^k\} \subseteq D_i \) such that \( w^k \to w^* \), and since \( D_i \) is closed, \( w^* \in D_i \). That is, \( \iota \notin \phi_D(w^*) \), which is a contradiction. Setting \( \delta := \min\{\delta_i : \iota \notin \phi_D(w^*)\} \), we see that \( B(w^*, \delta) \cap D_i = \emptyset \) for all \( \iota \notin \phi_D(w^*) \). In other words, if \( \iota \notin \phi_D(w) \) (i.e., \( w \in D_i \)) and \( w \in B(w^*, \delta) \cap \mathbb{R} \), then \( w \in \phi_D(w^*) \). Thus, the proof is complete.

B.3 Proof of Theorem 2.4

Proof. Since \( w^* \in \text{Fix}(T) \) by Theorem 2.1 and \( T \) is single-valued at \( w^* \), we know that \( T_i(w^*) = w^* \) for all \( \iota \in \phi_D(w^*) \). Using Lemma 2.3 we can find \( \delta > 0 \) such that \( \phi_D(w) \subseteq \phi_D(w^*) \) for all \( B(w^*, \delta) \). We can then find a subsequence \( \{w^{kjr}\}_{r=0}^\infty \subseteq B(w^*, \delta) \) of \( \{w^k\} \) such that \( w^{kjr} \to w^* \). Let \( \iota_0 \in \phi_D(w^k) \) such that \( w^{kjr} \in T_i(w^k) \). Since \( w^k \in B(w^*, \delta) \), we have \( \phi_D(w^k) \subseteq \phi_D(w^*) \) so that \( \iota_0 \in \phi_D(w^*) \) and \( w^* \in T_i(w^k) \). By (2.3)

$$
\|w^{kjr} - w^*\| \leq \kappa_{i_0}\|w^{kjr} - w^*\| \leq \kappa\delta,
$$

where \( \kappa := \max\{\kappa_i : \iota \in \phi_D(w^*)\} \). Thus, \( w^{kjr} \in B(w^*, \delta) \) and we may proceed inductively to conclude that \( \|w^{kjr} - w^*\| \leq \|w^{kjr} - w^*\| \) for all \( k \geq k_0 \) and

$$
\|w^k - w^*\| \leq \kappa^{k-k_0}\|w^{kjr} - w^*\|, \quad \forall k \geq k_0.
$$

Thus, \( \{\|w^k - w^*\|\}_{k=k_0}^\infty \) is a decreasing sequence that is bounded below, and is therefore convergent. Since \( \|w^{kjr} - w^*\| \to 0 \), it follows that \( \{\|w^k - w^*\|\}_{k=0}^\infty \) also converges to 0, that is, \( w^k \to w^* \).

C Missing proofs in Section 3

C.1 Proof of Proposition 3.9

Proof. We already have that \( w^k \to w^* \) by Theorem 3.6. From Theorem 2.4 we know that there exists \( N \geq 0 \) such that for each \( k \geq N \), we can find \( \iota = (i,j) \in I \times J \) (dependent on \( k \)) such that \( w^{k+1} = T_{ij}(w^k) \) and
\( w^* = T_i(w^*) \), where \( T_i = P_{R_j} \circ (I - \lambda \nabla f_i) \). Then
\[
\|w^{k+1} - w^*\| = \|(P_{R_j} \circ (I - \lambda \nabla f_i))(w^k) - (P_{R_j} \circ (I - \lambda \nabla f_i))(w^*)\|
\leq \|(I - \lambda \nabla f_i)(w^k - w^*)\|
\]

Taking a larger \( N \), if necessary, we have from Theorem 3.8 that there exists \( m \in J \) such that \( w^k, w^* \in R_m \).

By hypothesis, we then obtain from the above inequality that \( \|w^{k+1} - w^*\| \leq \kappa_{im} \|w^k - w^*\| \leq \kappa \|w^k - w^*\| \), where \( \kappa = \max\{\kappa_{ij} : w^* \in R_j\} \). This completes the proof.

C.2 Proof of Theorem 3.10

Proof. Let \( w \) be a local minimum of \( f + g - h \), and let \((i, j, m) \in I \times J \times M\) be such that \( w \in D_i(f) \cap D_j(g) \cap D_m(h) \). Then there exists \( \delta > 0 \) such that
\[
(f_i + g_j - h_m)(w) = (f + g - h)(w) \leq (f + g - h)(z) \leq (f_i + g_j - h_m)(z), \quad \forall z \in B(w, \delta),
\]
where the last inequality follows from the definition of \( f, g \) and \( h \). That is, \( w \) is a local minimum of \( f_i + g_j - h_m \), which proves Theorem 3.10(a). Meanwhile, consider any \( \lambda \in (0, \bar{\lambda}) \cap (0, 1/L] \). Computation similar to that for obtaining (3.10) gives
\[
(f_i + g_j - h_m)(z) \leq f_i(w) + \langle \nabla f_i(w), z - w \rangle + \frac{L_i}{2} \|z - w\|^2 - h_m(w) - \langle \nabla h_m(w), z - w \rangle + g_j(z)
\leq f_i(w) + \langle \nabla f_i(w), z - w \rangle + \frac{1}{2\lambda} \|z - w\|^2 - h_m(w) - \langle \nabla h_m(w), z - w \rangle + g_j(z)
=: Q_{f_i}^\lambda(z, w) - L_{h_m}(z, w) + g_j(z),
\]
where \( Q_{f_i} \) and \( L_{h_m} \) are defined similar to those in the proof of Theorem 3.2, and the second inequality holds since \( \lambda \in (0, 1/L] \) and \( L \geq L_i \). Minimizing both sides of the above inequality over all \( z \in B(w, \delta) \), we have from the local minimality of \( w \) that
\[
(f_i + g_j - h_m)(w) \leq \min_{z \in B(w, \delta)} Q_{f_i}^\lambda(z, w) - L_{h_m}(z, w) + g_j(z)
\leq Q_{f_i}^\lambda(w, w) - L_{h_m}(w, w) + g_j(w)
=: (f_i + g_j - h_m)(w).
\]
That is, \( w \) is also a local minimum of \( Q_{f_i}^\lambda(z, w) - L_{h_m}(z, w) + g_j(z) \). Since \( \lambda \in (0, \bar{\lambda}) \), \( Q_{f_i}^\lambda(z, w) - L_{h_m}(z, w) + g_j(z) \) is a convex function in \( z \) (see also Remark 3.1(e)), and therefore globally minimized at \( z = w \). Hence, we conclude that (see also (3.11))
\[
w = \text{prox}_{\lambda g_j}(w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w)),
\]
for all \((i, j, m) \in I \times J \times M\) such that \( w \in D_i(f) \cap D_j(g) \cap D_m(h) \) and all \( \lambda \in (0, \bar{\lambda}) \cap (0, 1/L] \). By (3.13), it suffices to show that there exists \( \varepsilon > 0 \) and some \((i, j, m) \in I \times J \times M\) such that \( w \in D_i(f) \cap D_j(g) \cap D_m(h) \) and
\[
M_{\lambda}^\lambda(w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w)) = M_{g_j}^\lambda(w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w)), \quad \forall \lambda \in (0, \bar{\lambda}) \cap (0, \varepsilon),
\]
in order to prove part (b) of the theorem. To this end, let \((i, m) \in I \times M\) be any index such that \( w \in D_i(f) \cap D_m(h) \). Note that if
\[
z_{\lambda} \in \text{prox}_{\lambda g_j}(w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w)), \quad \lambda \in (0, \bar{\lambda}),
\]
38
then we obtain from \(3.12\) that
\[
V(w) - V(w^*) \geq V(w) - V(z_\lambda) \geq \frac{1 - \lambda L}{2\lambda} \|w - z_\lambda\|^2,
\] (C.3)
where \(w^*\) is an optimal solution to \((\text{OP})\). Meanwhile, taking \(D := \{D_j(g) : j \in J\}\) and defining \(\phi_D\) as in \((1.5)\), we know from Lemma 2.3 that there exists \(\eta > 0\) such that \(\phi_D(z) \subseteq \phi_D(w)\) for all \(z \in B(w, \eta) \cap \text{dom}(g)\). Using \((\text{C.3})\) we can find \(\varepsilon > 0\) small enough so that \(\|w - z_\lambda\| < \eta\) for any \(\lambda \in (0, \varepsilon]\). Now, fix \(\lambda \in (0, \lambda) \cap (0, \varepsilon]\) and let \(j \in \phi_D(z_\lambda)\). Then \(g(z_\lambda) = g_j(z_\lambda)\) and we have
\[
M^\lambda_j(w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w)) \geq \frac{1}{2\lambda} \|z_\lambda - (w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w))\|^2
\]
\[
= M_j^\lambda(w) - \lambda \nabla f_i(w) + \lambda \nabla h_m(w)) \geq M_j^\lambda(w - \lambda \nabla f_i(w) + \lambda \nabla h_m(w))
\]
where the last inequality holds due to Lemma 3.1 (a). Since \(z_\lambda \in \text{dom}(g), \phi_D(z_\lambda) \subseteq \phi_D(w)\) so that \(w \in D_j(g)\). That is, we have shown that \((\text{C.1})\) holds, completing the proof of Theorem 3.10 (b). Part (c), on the other hand, follows immediately from \((\text{C.3})\).

C.3 Proof of Proposition 3.12

**Proof.** Since \(w \in \text{Fix}(T^\lambda_{PS})\) is nondegenerate, there exists \(i \in I\) such that \(w \in D_i(f)\) and a unique \(j \in J\) such that \(w = P_{R_j}(w - \lambda \nabla f_i(w))\). Hence, \(-\nabla f_i(w) \in N_{R_j}(w)\), where \(N_{R_j}(w)\) is the normal cone to \(R_j\) at \(w\) in the sense of convex analysis. For each \(j \in J\), let \(D_j := \{z \in \mathbb{E} : \text{dist}(z, S) = \text{dist}(z, R_j)\}\). Define \(\phi_D(z) := \{j : z \in D_j\}\), then by Lemma 2.3 and the nondegeneracy of \(w\), there exists \(\delta > 0\) such that \(\phi_D(z) = \{j\}\) for all \(z \in B(w, \delta)\). Now, set \(\varepsilon = \delta/(2\|\nabla f_i(w)\|)\) and take any \(\lambda \in (0, \varepsilon]\), then \(\|w - (w - \lambda \nabla f_i(w))\| < \delta\) so that \(\phi_D(w - \lambda \nabla f_i(w)) = \{j\}\). Moreover, \(P_{R_j}(w) = w = \lambda \nabla f_i(w)\) due to the fact that \(-\nabla f_i(w) \in N_{R_j}(w)\). Since \(w - \lambda \nabla f_i(w) \
\)D Missing proofs in Section 5.2

D.1 Proof of Theorem 5.8

We will first develop necessary tools for separately considering different cases of \(w\) and \(w^+ \in T^\lambda_{PS}(w)\). The following lemma will be our key tool to proving the desired result.

**Lemma D.1.** Let \(w \in S_2 \setminus \text{Fix}(T^\lambda_{PS})\) and \(w^+ \in T^\lambda_{PS}(w)\) where \(\lambda \in (0, 1/LQ]\). If \(w \in R_i\), \(\hat{w} := P_{R_i}(w - \lambda \nabla f_Q(w))\), and \(w \neq \hat{w}\), then \(f_Q(w^+) < f_Q(w)\).

**Proof.** From \((3.11)\) we have
\[
T^\lambda_{PS}(w) = P_S(w - \lambda \nabla f_Q(w)) = \arg \min_{z \in S_2} Q^\lambda_{f_Q}(z, w).
\] (D.1)
By the convexity of \(R_i\) and the definition of \(\hat{w}\), we have \(\langle w - \lambda \nabla f_Q(w) - \hat{w}, w - \hat{w}\rangle \leq 0\). Then
\[
\lambda \langle \nabla f_Q(w), \hat{w} - w \rangle = \langle \lambda \nabla f_Q(w) - (w - \hat{w}), \hat{w} - w \rangle + \langle w - \hat{w}, \hat{w} - w \rangle \leq -\|\hat{w} - w\|^2.
\]
Thus, from the definition of \(Q^\lambda_{f_Q}\) in \((3.9)\)
\[
Q^\lambda_{f_Q}(\hat{w}, w) \leq f_Q(w) - \frac{\lambda}{\lambda} \|\hat{w} - w\|^2 + \frac{1}{2\lambda} \|\hat{w} - w\|^2 = f_Q(w) - \frac{1}{2\lambda} \|\hat{w} - w\|^2 < f_Q(w),\]
(D.2)
where the last inequality holds because \(w \neq \hat{w}\). Further, we have from \((3.9)\) and \((D.1)\) that \(f_Q(w^+) \leq Q^\lambda_{f_Q}(w^+, w) \leq Q^\lambda_{f_Q}(\hat{w}, w)\), combining which with \((D.2)\) then gives the desired result.

\]
39
In particular, the above lemma shows that if \( w \notin \text{Fix}(T^\lambda_{PS}) \) and \( w^+ \in T^\lambda_{PS}(w) \) belong to the same convex set \( R_\iota \), then \( f_Q(w^+) < f_Q(w) \). In the other case, we will show that there exists some \( \iota \in \mathcal{I} \) such that \( w \in R_\iota \) and the vector \( \bar{w} = P_{R_\iota}(w - \lambda \nabla f_Q(w)) \) is distinct from \( w \). First, we consider the instance when \( w \) is a nondegenerate point of \( S_2 \) as defined in Definition 3.2. That is, there is a unique \( \iota \in \mathcal{I} \) such that \( w \in R_\iota \), so \((w_j, w_{n+j}) \neq (0,0)\) for all \( j \in [n] \).

**Proposition D.2.** Let \( w \in S_2 \setminus \text{Fix}(T^\lambda_{PS}) \) and \( w^+ \in T^\lambda_{PS}(w) \) where \( \lambda \in (0,1/L_Q) \). If \( M \) is a nondegenerate matrix, and \( w \) is nondegenerate, then \( f_Q(w^+) < f_Q(w) \).

**Proof.** Let \( \iota \in \mathcal{I} \) such that \( w \in R_\iota \), and let \( \hat{w} \) be as in [Lemma D.1]. To prove the claim, we only need to show that \( w \neq \hat{w} \) and the result follows from Lemma D.1. Suppose to the contrary that \( w = \hat{w} \), that is, \( w = P_{R_\iota}(\bar{w}) \) for \( \bar{w} := w - \lambda \nabla f_Q(w) \). Since \( w = P_{R_\iota}(\bar{w}) \) and \((w_j, w_{n+j}) \neq (0,0)\) for all \( j \in [n] \), we have from [Equation (5.24)] that \( w_j = \bar{w}_j > 0 \) and \( w_{n+j} = 0 \) if \( j \in \iota \), and \( w_{n+j} = \bar{w}_{n+j} > 0 \) and \( w_j = 0 \) otherwise. Thus,

\[
(\bar{w}_j, \bar{w}_{n+j}) - (w_j, w_{n+j}) = \begin{cases} 
(0, \bar{w}_{n+j}) & \text{if } j \in \iota, \\
(\bar{w}_j, 0) & \text{if } j \notin \iota, 
\end{cases} \quad \forall j \in [n].
\]

Hence,

\[
(\bar{w}_j - w_j)(\bar{w}_{n+j} - w_{n+j}) = 0, \quad \forall j \in [n]. \tag{D.3}
\]

Since \( \bar{w} - w = -\lambda \nabla f_Q(w) \), we have \( \bar{w} - w \in \text{Ran}(A^T) \) from [Equation (5.8)]. Thus, \( \bar{w} - w = \text{Ker}(A)^\perp \). By this together with [Equation (D.3)] and the nondegeneracy of \( M \), we have from [Alcantara et al., 2022 Proposition 2.11] that \( \bar{w} - w = 0 \). Consequently, we have \( \nabla f_Q(w) = 0 \), and since \( A \) is of full row rank, it follows from [Equation (5.8)] that \( Aw - b = 0 \). That is, \( w \in S_1 \), and in turn, we get \( w \in S_1 \cap S_2 \). This is a contradiction since \( w \notin \text{Fix}(T^\lambda_{PS}) \). Hence, \( w \neq \hat{w} \), as desired. \( \square \)

If \( w \) is degenerate, [Equation (D.3)] does not hold anymore, which prohibits the use of [Alcantara et al., 2022 Proposition 2.11]. To deal with this case, we need the following lemma.

**Lemma D.3.** Let \( w \in S_2 \setminus \text{Fix}(T^\lambda_{PS}) \) and \( w^+ \in T^\lambda_{PS}(w) \), where \( \lambda \in (0,1/L_Q) \), and suppose that \( w \) is degenerate. Let \( \iota \in \mathcal{I} \) such that \( w \in R_\iota \) and suppose that \( w = \hat{w} \), where \( \hat{w} = P_{R_\iota}(w - \lambda \nabla f_Q(w)) \). Denote \( \hat{w} := w - \lambda \nabla f_Q(w) \) and

\[
\Gamma(w) := \{ j \in [n] \mid w_j = w_{n+j} = 0 \text{ and } (\bar{w}_j, \bar{w}_{n+j}) \notin \mathbb{R}_2^- \}, \tag{D.4}
\]

where \( \mathbb{R}_2^- := \{(x_1, x_2) : x_1, x_2 \leq 0\} \), then \( \Gamma(w) \neq \emptyset \) implies \( f_Q(w^+) < f_Q(w) \).

**Proof.** Define \( \Gamma(w)_1 := \Gamma(w) \cap \iota \) and \( \Gamma(w)_2 := \Gamma(w) \cap \iota^c \). Note that since \( \Gamma(w) \neq \emptyset \), either \( \Gamma(w)_1 \) or \( \Gamma(w)_2 \) is nonempty. Now, since \( w = P_{R_\iota}(w) \), we obtain from [Equation (3.24)] that

\[
\begin{cases}
\bar{w}_j < 0, \quad \bar{w}_{n+j} > 0 & \text{if } j \in \Gamma(w)_1 \\
\bar{w}_j > 0, \quad \bar{w}_{n+j} < 0 & \text{if } j \in \Gamma(w)_2. \tag{D.5}
\end{cases}
\]

Let \( \iota' \in \mathcal{I} \) be given by \( \iota' = \Lambda_1 \cup \Lambda_2 \), where \( \Lambda_1 = (\iota \cap \Gamma(w)^c) \cup \Gamma(w)_2 \) and \( \Lambda_2 = \{n+j : j \in [n], j \notin \Lambda_1\} \). Namely, for all \( j \in [n] \) with \((w_j, w_{n+j}) = (0,0), \) \( \iota' \) picks the one in \( \{j, n+j\} \) not included in \( \iota \). Then \( w \in R_{\iota'} \), and by setting \( w' := P_{R_{\iota'}}(w - \lambda \nabla f_Q(w)) = P_{R_{\iota'}}(\hat{w}) \), we have from the definition of \( \iota' \) and [Equation (D.5)] that

\[
(w_j', w_{n+j}') = \begin{cases}
(w_j, w_{n+j}) & \text{if } j \notin \Gamma(w) \\
(0, \bar{w}_{n+j}) & \text{if } j \in \Gamma(w)_1 \\
(\bar{w}_j, 0) & \text{if } j \in \Gamma(w)_2
\end{cases}.
\]

Since \( \bar{w}_{n+j} \neq 0 \) for \( j \in \Gamma(w)_1 \) and \( \bar{w}_j \neq 0 \) for \( j \in \Gamma(w)_2 \), we see that \( w \neq w' \). By Lemma D.1, \( f_Q(w^+) < f_Q(w) \). \( \square \)

40
Finally, we also need the following lemma before proving Theorem 5.8

**Lemma D.4.** Cottle et al. [1992, Theorem 3.3.4] $M \in \mathbb{R}^{n \times n}$ is a $P$-matrix if and only if whenever $x_j(Mx) \leq 0$ for all $j \in [n]$, we have $x = 0$.

**Proof of Theorem 5.8.** If $w$ is nondegenerate, the result immediately follows from Proposition D.2 since a $P$-matrix is necessarily nondegenerate. Assume that $w$ is degenerate and let $i \in I$ such that $w \in R_i$. If $w \neq \bar{w}$, where $\bar{w} := P_{R_i}(w - \lambda \nabla f_Q(w))$, the result immediately follows from Lemma D.1 Suppose now that $w = \bar{w}$. We claim that $\Gamma(w)$ given by (D.4) is nonempty. To this end, consider the following index sets:

$$
I_1(w) := \{ j \in [n] : w_j = w_{n+j} = 0 \},
$$

$$
I_2(w) := \{ j \in [n] : w_j > 0 \text{ and } w_{n+j} = 0 \},
$$

$$
I_3(w) := \{ j \in [n] : w_j = 0 \text{ and } w_{n+j} > 0 \}.
$$

Since $w \in S_2$, $I_1 \cup I_2 \cup I_3 = [n]$. If $j \in I_2 \cup I_3$, we have from the equation $w = P_{R_i}(\bar{w})$ and (5.24) that

$$
\begin{align*}
w_j &= \bar{w}_j > 0 \quad \text{if } j \in I_2, \\
w_{n+j} &= \bar{w}_{n+j} > 0 \quad \text{if } j \in I_3.
\end{align*}
$$

(D.6)

Meanwhile, as in the proof of Proposition D.2 we have $z := \bar{w} - w \in \text{Ran}(A^T)$. From the formula of $A$, it is not difficult to verify that $\text{Ran}(A^T) = \text{Ker}([I \ M^T])$. Thus, by letting $z = (u, v) \in \mathbb{R}^n \times \mathbb{R}^n$, we have $u + M^Tv = 0$, that is,

$$
(M^Tv)_j = -u_j, \quad \forall j \in [n].
$$

By multiplying both sides by $v_j$, we then obtain

$$
v_j(M^Tv)_j = -u_jv_j = -(\bar{w}_j - w_j)(\bar{w}_{n+j} - w_{n+j}) = \begin{cases} 
-\bar{w}_j\bar{w}_{n+j} & \text{if } j \in I_1 \\
0 & \text{if } j \in I_2 \cup I_3.
\end{cases}
$$

(D.7)

where the last equality follows from (D.6) and the definition of the index set $I_1$. Now, if $\Gamma(w) = \emptyset$, then $\vec{w}_j, \vec{w}_{n+j} \leq 0$ for all $j \in I_1$, and the above equation implies that $v_j(M^Tv)_j \leq 0$ for all $j \in [n]$. By Lemma D.4, $v = 0$ since $M$ is a $P$-matrix, which in turn gives $u = 0$. That is, we have $\bar{w} - w = 0$. As shown in the proof of Proposition D.2 this implies that $w \in S_1 \cap S_2$, which is a contradiction since $w \notin \text{Fix}(T_{P_R}^\lambda)$. Hence, we must have $\Gamma(w) \neq \emptyset$ and by Lemma D.3, we get $f_Q(w^+) < f_Q(w)$. This completes the proof. \square

**D.2 Proof of Theorem 5.7.**

**Proof.** First, we show that

$$
w \in P_{S_2}(\bar{w}) \quad \text{and} \quad \bar{w} - w \in \text{Ran}(A^T) \quad \implies \quad w = \bar{w} \in S_1 \cap S_2.
$$

(D.8)

Indeed, let $z := \bar{w} - w$ and denote $z = (u, v) \in \mathbb{R}^n \times \mathbb{R}^n$. From the proof of Theorem 5.8 we know that (D.7) holds. Since $w \in P_{S_2}(\bar{w})$, we have that $(\bar{w}_j, \bar{w}_{n+j}) \in \mathbb{R}_+^2$ for all $j \in I_1$. Hence, we obtain from the same arguments in the proof of Theorem 5.8 that $w = \bar{w} \in S_1 \cap S_2$. We now consider the three cases separately:

(i) Suppose $z \in \text{Fix}(T_{PDMC}^\lambda)$. Then from the formula of $T_{PDMC}^\lambda$ in (5.10), it can be verified that $z = -\nabla f_Q(z) + z'$, where $z' \in P_{S_2}(z)$. By (5.8) and (D.8) we get the desired result.

(ii) Let $z \in \text{Fix}(T_{FB}^\lambda)$ and denote $\bar{z} := z - \lambda \nabla f_Q(z)$ and $w := ((1 + \lambda)z - \bar{w})/\lambda$. From the formula of $T_{FB}^\lambda$ in (5.11) we can derive that $w \in P_{S_2}(\bar{w})$. We then have $\bar{w} - w = (1 + \lambda)(\bar{w} - z)/\lambda = -(1 + \lambda)\nabla f(z)$, so that $\bar{w} - w \in \text{Ran}(A^T)$ by (5.8). By (D.8) we have $w = \bar{w} \in S_1 \cap S_2$. From the formula of $w$, we obtain that $z = w$ so that $z \in S_1 \cap S_2$.

(iii) If $z \in \text{Fix}(T_{P_R}^\lambda)$, from (5.12) we have $z \in P_{S_2}(\bar{w})$ where $\bar{w} = z - \lambda \nabla f(z)$. From the latter equation and (5.8) we obtain $\bar{w} - z \in \text{Ran}(A^T)$ so that $z = \bar{w} \in S_1 \cap S_2$, as desired. \square
D.3 Proof of Theorem 5.9

Proof. (a) Let \( \{w^k\} \) be a sequence generated by the accelerated PDMC algorithm. Suppose \( \iota \in \mathcal{I} \) such that \( w^* \notin R_\iota \), and suppose there exists infinitely many terms in \( \{w^k\} \) such that \( \text{dist}(w^k, S_2) = \text{dist}(w^k, R_\iota) \).

That is, there exists a subsequence \( \{w^{k_j}\}_{j=0}^\infty \subseteq D_\iota \) where \( D_\iota := \{w \in \mathbb{R}^{2n} : \text{dist}(w, S_2) = \text{dist}(w, R_\iota)\} \).

Since \( V = f_Q + \text{dist}(\cdot, S_2)^2/2 \) is coercive (see Theorem 5.5), and \( V(w^k) \) is decreasing by (3.19) and (3.12), \( \{w^k\} \) is bounded. Thus, \( \{w^{k_j}\}_{j=0}^\infty \) has an accumulation point, which by Theorem 2.2 belongs to \( \text{Fix}(T^\lambda_{\text{PDMC}}) \). Since \( \text{Fix}(T^\lambda_{\text{PDMC}}) = S_1 \cap S_2 \) by Theorem 5.7 this accumulation point is equal to \( w^* \). But since \( D_\iota \) is closed, it follows that \( w^* \in D_\iota \). That is, \( \text{dist}(w^*, R_\iota) = \text{dist}(w^*, S_2) = 0 \). This contradicts the choice of \( \iota \). Hence, there are only finitely many terms of \( \{w^k\} \) in \( D_\iota \) if \( w^* \notin R_\iota \).

Now, let \( \iota \in \mathcal{I}' \) and \( \{w^{k_j}\}_{j=0}^\infty \) be a subsequence in \( D_\iota \). Since \( V(w^k) \to 0 \), we have \( \|(w^{k_j})_c\| \to 0 \) and \( f_Q(w^{k_j}) \to 0 \) from (3.25). If \( Q = (AA^T)^{-1} \), then by noting that \( A^\dagger b = A^\dagger Aw^* = (A^\dagger A)_{(:,:w^*}, \) we have

\[
f_Q(w^{k_j}) \leq \frac{1}{2}\|((A^\dagger A)_{(:,:)c}(w^{k_j})_c - (A^\dagger A)_{(:,:)c}(w^*)_c\|^2 + c_{k_j},
\]

where

\[c_{k_j} := ((A^\dagger A)_{(:,:)c}(w^{k_j})_c - (A^\dagger A)_{(:,:)c}(w^*)_c, (A^\dagger A)_{(:,:)c}(w^{k_j})_c) + \frac{1}{2}\|((A^\dagger A)_{(:,:)c}(w^{k_j})_c\|^2.

By the boundedness of \( \{w^{k_j}\}_{k=0}^\infty \) and the fact that \( \|(w^{k_j})_c\| \to 0 \), we see that \( c_{k_j} \to 0 \). It follows from (D.9) that \( \|(A^\dagger A)_{(:,:)c}(w^{k_j})_c - (A^\dagger A)_{(:,:)c}(w^*)_c\|^2 \to 0 \). Meanwhile, we have from (5.27) that

\[
\nu \|A\|^2 \|(w^{k_j})_c - (w^*)_c\|^2 \leq \|(A^\dagger A)_{(:,:)c}(w^{k_j})_c - (A^\dagger A)_{(:,:)c}(w^*)_c\|^2,
\]

and therefore \( w^{k_j} \to w^* \). By utilizing (5.28) and following the same arguments, we also obtain that \( w^{k_j} \to w^* \) if \( Q = I \). In summary, we have shown that for any \( \iota \in \mathcal{I}' \), the subsequence of \( \{w^k\} \) confined in \( D_\iota \) converges to \( w^* \). Since \( \mathcal{I}' \) is finite, it follows that the full sequence \( \{w^k\} \) converges to \( w^* \).

The proof for the convergence of the accelerated FB algorithm is exactly the same as that of the accelerated PDMC.

(b) By arguing as in the proof of Theorem 3.8 and by noting that \( \text{Fix}(T^\lambda_{\text{PDMC}}) = \{w^*\} \) by Theorems 2.2 and 5.7 it follows that only those sets \( R_\iota \) such that \( w^* \in R_\iota \) can possibly contain infinitely many terms of \( \{w^k\} \). Arguing as in the proof of Theorem 5.6 we get the linear convergence of \( f(w^k) \) to 0. By (5.29) we conclude that \( w^k \to w^* \). This completes the proof. \( \square \)