Computational Optimization and Applications (2025) 91:1185–1225
https://doi.org/10.1007/s10589-025-00680-1

An extended sequential quadratic method with extrapolation

Yongle Zhang[1] · **Ting Kei Pong**[2] · **Shiqi Xu**[3]

Received: 24 December 2023 / Accepted: 17 March 2025 / Published online: 7 April 2025

Abstract

We revisit and adapt the extended sequential quadratic method (ESQM) in Auslender (J Optim Theory Appl 156:183–212, 2013) for solving a class of difference-of-convex optimization problems whose constraints are defined as the intersection of level sets of Lipschitz differentiable functions and a simple compact convex set. Particularly, for this class of problems, we develop a variant of ESQM, called ESQM with extrapolation ($\mathrm{ESQM_e}$), which incorporates Nesterov's extrapolation techniques for empirical acceleration. Under standard constraint qualifications, we show that the sequence generated by $\mathrm{ESQM_e}$ clusters at a critical point if the extrapolation parameters are uniformly bounded above by a certain threshold. Convergence of the whole sequence and the convergence rate are established by assuming Kurdyka-Łojasiewicz (KL) property of a suitable potential function and imposing additional differentiability assumptions on the objective and constraint functions. In addition, when the objective and constraint functions are all convex, we show that linear convergence can be established if a certain exact penalty function is known to be a KL function with exponent $\frac{1}{2}$; we also discuss how the KL exponent of such an exact penalty function can be deduced from that of the *original* extended objective (i.e., sum of the objective and the indicator function of the constraint set). Finally, we perform numerical experiments to demonstrate the empirical acceleration of $\mathrm{ESQM_e}$ over a basic version of ESQM, and illustrate its effectiveness by comparing with the natural competing algorithm $\mathrm{SCP_{ls}}$ from Yu et al. (SIAM J Optim 31:2024–2054, 2021).

Keywords ESQM · Extrapolation · KL exponent · Linear convergence

✉ Ting Kei Pong
tk.pong@polyu.edu.hk

[1] Department of Mathematics, Visual Computing and Virtual Reality Key Laboratory of Sichuan Province, Sichuan Normal University, Chengdu, People's Republic of China

[2] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China

[3] Department of Mathematics, Sichuan Normal University, Chengdu, People's Republic of China

1 Introduction

Extrapolation techniques, due to their simplicity and easy adaptability, have been widely studied in recent years to empirically accelerate first-order methods; see, for example, [14, 21, 28, 32, 36] and references therein. Among them, Nesterov's extrapolation techniques [28–31] have been successfully applied to accelerate the proximal gradient algorithm [25] for minimizing $f+h$, with f being a convex *loss function* with Lipschitz continuous gradient, and h being a proper closed convex and possibly nonsmooth *regularizer* with *easy-to-compute* proximal operator. These studies led to the developments of various algorithms and softwares including the well-known algorithm FISTA [7] for linear inverse problems and the software TFOCS [8] for solving a large class of convex cone problems. Nesterov's extrapolation techniques have also been suitably adapted in subsequent works such as [37, 38] in some nonconvex settings, and most of these works also require the proximal operator of (part of) the regularizer to be easy to compute: in the case when h is the indicator function of some closed set D, this requirement amounts to saying that the projection onto D can be computed efficiently. In this paper, we consider a class of constrained optimization problems whose constraint sets *do not admit easy projections*, and investigate the adaptation of extrapolation techniques on empirically accelerating a classical algorithm for these problems.

Specifically, we consider the following difference-of-convex (DC) optimization problem with smooth inequality and simple geometric constraints:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & P(x) := P_1(x) - P_2(x) \\ \text{s.t.} \quad & g_i(x) \le 0, \quad i = 1, \ldots, m, \\ & x \in C, \end{aligned} \tag{1.1}$$

where $P_1 : \mathbb{R}^n \to \mathbb{R}$ and $P_2 : \mathbb{R}^n \to \mathbb{R}$ are convex, each $g_i : \mathbb{R}^n \to \mathbb{R}$ is smooth and ∇g_i is Lipschitz continuous, $C \subseteq \mathbb{R}^n$ is a nonempty compact convex set, and the feasible set $C \cap \mathscr{F}$ is nonempty, where $\mathscr{F} := \{x \in \mathbb{R}^n : g_i(x) \le 0,\ i = 1, \ldots, m\}$. This class of problems arises naturally in many applications. For example, in compressed sensing, the P can be a sparsity inducing regularizer such as the difference of ℓ_1 and ℓ_2 norms [39], the g_i, $i = 1, \ldots, m$, can be loss functions based on the noise models in the ith transmission channel, and the set C can be used to model some priors such as nonnegativity or boundedness.

Since projections onto the feasible set of (1.1) are not easy to compute, existing algorithms for (1.1) usually leverage the Lipschitz continuity of ∇g_i to build approximations for the feasible sets, leading to relatively easier subproblems. One natural approach for building approximations is to replace g_i by its quadratic majorants at the current iterate. Specific algorithms based on quadratically approximating g_i in (1.1) include the moving balls approximation algorithm [5] and its variants (see, e.g., [12, 40]). Another natural approach for building approximations is to make use of *affine approximations* to g_i at the current iterate, leading to subproblems with even simpler structures. This approach has its roots in the literature of sequential quadratic programming (SQP) method, and we refer the readers to [20] and references therein for more discussions of SQP. Here, we are interested in the framework described in [4], which focused on solving (1.1) when P and each g_i are twice continuously differentiable. We adapt the algorithmic framework described there to solve (1.1), and incorporate extrapolation techniques to empirically accelerate the algorithm. We call the resulting algorithm extended sequential quadratic method with extrapolation (ESQM$_\mathrm{e}$), following the use of the name ESQM in [4]. The algorithmic details will be presented in Sect. 3 below; in particular, in each iteration of ESQM$_\mathrm{e}$, the g_i in (1.1) is replaced by its affine approximation at a point *extrapolated from* the current iterate.

In this paper, we study the convergence properties of ESQM$_\mathrm{e}$ and perform numerical experiments to examine its computational efficiency. In particular, we show that the sequence generated by ESQM$_\mathrm{e}$ clusters at a critical point if the extrapolation parameters are uniformly bounded above by a certain threshold, under a set of constraint qualifications similarly used in [4]. We also construct a suitable potential function and establish the convergence of the whole sequence and its convergence rate by assuming Kurdyka-Łojasiewicz (KL) property of the potential function and additional differentiability conditions on P_2 and each g_i in (1.1). Furthermore, when $P_2 \equiv 0$ and each g_i is convex, we show that linear convergence can also be established if a certain exact penalty function of (1.1) is known to be a KL function with exponent $\frac{1}{2}$. We also discuss how the KL exponent of such an exact penalty function can be derived from that of the function $P + \delta_{C\cap\mathscr{F}}$ from (1.1) (see Sect. 2 for notation). Finally, we perform numerical experiments on compressed sensing models with different types of measurement noises taking the form of (1.1). Our experiments on random instances illustrate the empirical acceleration of ESQM$_\mathrm{e}$ over a basic version of ESQM, and also suggest that ESQM$_\mathrm{e}$ outperforms the natural competing algorithm SCP$_\mathrm{ls}$ from [40].

The remainder of the paper is organized as follows. We present notation and preliminary materials in Sect. 2. Our algorithm, ESQM$_\mathrm{e}$ is presented in Sect. 3, and its subsequential and sequential convergences are established in Sect. 4.1. We discuss the convergence behavior in the convex setting (i.e., $P_2 \equiv 0$ and each g_i is convex in (1.1)) in Sect. 4.2, and the relationship between the KL exponent of the function $P + \delta_{C\cap\mathscr{F}}$ from (1.1) and that of the exact penalty function used in the analysis in Sect. 4.2 is studied in Sect. 5. Numerical experiments are presented in Sect. 6.

2 Notation and preliminaries

In this paper, we let $\mathbb{R}$ and $\mathbb{R}_+$ denote the sets of real numbers and nonnegative real numbers respectively, and $\mathbb{N}$ is the set of positive integers. We also let $\mathbb{R}^n$ and $\mathbb{R}^n_+$ denote the n-dimensional Euclidean space and its nonnegative orthant respectively. For an $x \in \mathbb{R}$, we let $(x)_+$ denote $\max\{x, 0\}$. For an $x \in \mathbb{R}^n$, we let $\|x\|$ denote its Euclidean norm; moreover, for x and $y \in \mathbb{R}^n$, we let $\langle x, y\rangle$ denote their inner product.

For an extended-real-valued function $f : \mathbb{R}^n \to (-\infty, +\infty]$, we say that f is proper if $\operatorname{dom} f := \{x : f(x) < \infty\} \neq \emptyset$. A proper function f is said to be closed if it is lower semicontinuous. We use $x^k \xrightarrow{f} x$ to denote $x^k \to x$ and $f(x^k) \to f(x)$. For a proper closed function f, the regular subdifferential of f at $w \in \operatorname{dom} f$ is given by

$$\widehat{\partial} f(w) := \left\{ \xi \in \mathbb{R}^n : \liminf_{v \to w, v \neq w} \frac{f(v) - f(w) - \langle \xi, v - w \rangle}{\|v - w\|} \geq 0 \right\}.$$

The (limiting) subdifferential of f at $w \in \mathrm{dom}\, f$ is given by

$$\partial f(w) := \left\{ \xi \in \mathbb{R}^n : \exists w^k \xrightarrow{f} w, \xi^k \to \xi \text{ with } \xi^k \in \widehat{\partial} f(w^k) \text{ for each } k \right\},$$

and we set $\partial f(x) = \widehat{\partial} f(x) = \emptyset$ when $x \notin \mathrm{dom}\, f$. We also define $\mathrm{dom}\, \partial f := \{x \in \mathbb{R}^n : \partial f(x) \neq \emptyset\}$. The above subdifferential of f is consistent with the classical subdifferential of f when f is in addition convex; indeed, in this case, we have

$$\partial f(w) = \left\{ \xi \in \mathbb{R}^n : \langle \xi, v - w \rangle \leq f(v) - f(w) \ \ \forall v \in \mathbb{R}^n \right\};$$

see, for example, [35, proposition 8.12]. For a nonempty closed set $D \subseteq \mathbb{R}^n$, the indicator function δ_D is defined by

$$\delta_D(x) = \begin{cases} 0 & x \in D, \\ \infty & x \notin D. \end{cases}$$

The normal cone of D at $x \in D$ is defined by $\mathcal{N}_D(x) := \partial \delta_D(x)$. Finally, the distance from a point x to D is denoted by $\mathrm{dist}(x, D)$, and the convex hull of D is denoted by $\mathrm{conv}\, D$.

We next recall some important definitions that will be used in the sequel. We start by recalling the following constraint qualification for (1.1) (which was also used in [4]), and the (associated) first-order optimality conditions for (1.1).

Definition 2.1 (**RCQ**) We say that the Robinson constraint qualification holds at an $x \in \mathbb{R}^n$ for (1.1) if the following statement holds:

$$RCQ(x): \ \exists y \in C \text{ such that } g_i(x) + \langle \nabla g_i(x), y - x \rangle < 0 \ \ \forall i = 1, \ldots m.$$

Definition 2.2 (**Critical point**) For (1.1), we say that x is a critical point of (1.1) if $x \in C$ and there exists $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m) \in \mathbb{R}^m_+$ such that (x, λ) satisfies the following conditions:

(i) $g_i(x) \leq 0 \ \ \forall i = 1, \ldots, m$,
(ii) $\lambda_i g_i(x) = 0 \ \ \forall i = 1, \ldots, m$,
(iii) $0 \in \partial P_1(x) - \partial P_2(x) + \sum_{i=1}^m \lambda_i \nabla g_i(x) + \mathcal{N}_C(x)$.

One can show using similar arguments as in [40, Section 2] that if $RCQ(x)$ holds at every $x \in C \cap \mathscr{F}$, then any local minimizer of (1.1) is a critical point of (1.1).

Next, we recall the definitions of Kurdyka-Łojasiewicz (KL) property and exponent.

Definition 2.3 (**Kurdyka-Łojasiewicz (KL) property and exponent**) A proper closed function f is said to satisfy the KL property at $\bar{x} \in \mathrm{dom}\, \partial f$ if there exist $r \in (0, \infty]$, a neighborhood U of $\bar{x}$, and a continuous concave function $\phi : [0, r) \to \mathbb{R}_+$ satisfying $\phi(0) = 0$ such that:

(i) ϕ is continuously differentiable on $(0, r)$ with $\phi' > 0$;
(ii) for all $x \in U$ with $f(\bar{x}) < f(x) < f(\bar{x}) + r$, it holds that

$$\phi'(f(x) - f(\bar{x}))\text{dist}(0, \partial f(x)) \geq 1. \tag{2.1}$$

If f satisfies the KL property at $\bar{x} \in \text{dom}\,\partial f$ and the ϕ in (2.1) can be chosen as $\phi(\varsigma) = \rho\varsigma^{1-\alpha}$ for some $\rho > 0$ and $\alpha \in [0, 1)$, then we say that f satisfies the KL property with exponent α at $\bar{x}$.

A proper closed function f satisfying the KL property at every point in $\text{dom}\,\partial f$ is called a KL function. A proper closed function f satisfying the KL property with exponent $\alpha \in [0, 1)$ at every point in $\text{dom}\,\partial f$ is called a KL function with exponent α.

Many functions are known to satisfy the KL property. For instance, proper closed semi-algebraic functions satisfy the KL property with some exponent $\alpha \in [0, 1)$; see [10]. The KL property plays an important role in the global convergence analysis of first order methods and the exponent is important in establishing convergence rates; see, for example, [2, 3, 13, 23].

Finally, before ending this section, we recall two technical lemmas. The first lemma concerns the uniformized KL property (see [13, Section 3.5]) and is taken from [40, Lemma 3.10]. The second lemma is a special case of Robinson [33] concerning error bounds for convex functions, which will be used in Sect. 5 for studying the KL property of a penalty function associated with (1.1).

Lemma 2.1 *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a level-bounded proper closed convex function with $\Lambda := \text{Argmin}\, f \neq \emptyset$. Let $\underline{f} := \inf f$. Suppose that f satisfies the KL property at each point in Λ with exponent $\alpha \in [0, 1)$. Then there exist $\epsilon > 0$, $r_0 > 0$ and $c_0 > 0$ such that*

$$\text{dist}(x, \Lambda) \leq c_0(f(x) - \underline{f})^{1-\alpha}$$

for any $x \in \text{dom}\,\partial f$ satisfying $\text{dist}(x, \Lambda) \leq \epsilon$ and $\underline{f} \leq f(x) < \underline{f} + r_0$.

Lemma 2.2 *Let $h : \mathbb{R}^n \to \mathbb{R}^m$ with each component function h_i being convex. Let $\Omega := \{x \in \mathbb{R}^n : 0 \in h(x) + \mathbb{R}^m_+\}$ and suppose there exist $x^s \in \Omega$ and $\delta_0 > 0$ such that $\{y \in \mathbb{R}^m : \|y\| \leq \delta_0\} \subseteq h(x^s) + \mathbb{R}^m_+$. Then*

$$\text{dist}(x, \Omega) \leq \frac{\|x - x^s\|}{\delta_0}\text{dist}(0, h(x) + \mathbb{R}^m_+) \quad \forall x \in \mathbb{R}^n.$$

3 Algorithmic framework

In this section, we present our algorithm for solving (1.1). To describe our algorithm, following the discussion in [37, Section 3], for each i, notice that we can rewrite g_i (whose gradient is Lipschitz continuous) as $g_i = g_i^1 - g_i^2$, where g_i^1 and g_i^2 are two convex functions with Lipschitz continuous gradients. The next remark concerns the Lipschitz continuity moduli of ∇g_i, ∇g_i^1 and ∇g_i^2.

Remark 3.1 (Lipschitz continuity moduli) Here and throughout, we denote a Lipschitz continuity modulus of ∇g_i^1 by $L_{g_i} > 0$ and a Lipschitz continuity modulus of ∇g_i^2 by $\ell_{g_i} \geq 0$. In addition, by taking a larger L_{g_i} if necessary, we will assume without loss of generality that $L_{g_i} \geq \ell_{g_i}$. Then one can show that ∇g_i is Lipschitz continuous with a modulus L_{g_i}. We also define $L_g := \max\{L_{g_i} : i = 1, \ldots, m\}$ and $\ell_g = \max\{\ell_{g_i} : i = 1, \ldots, m\}$.

The algorithm we study in this paper is presented as Algorithm 1 below; here and throughout, for notational simplicity, for each $u, w \in \mathbb{R}^n$, we define

$$\mathrm{lin}_{g_i}(u, w) := g_i(w) + \langle \nabla g_i(w), u - w \rangle \quad \forall i = 1, \ldots, m, \quad \text{and} \quad \mathrm{lin}_{g_0}(u, w) := 0. \tag{3.1}$$

We identify our algorithm as an extended sequential quadratic method with extrapolation (ESQM$_\mathrm{e}$), where "extrapolation" refers to (3.3). This is because when $\beta_k \equiv 0$, our algorithm reduces to an instance of the ESQM proposed in [4], whose convergence was studied for solving (1.1) when the P and each g_i in (1.1) are in addition twice continuously differentiable.[1] Notice that (x^{k+1}, s^{k+1}) solves the subproblem in (3.4) if and only if $s^{k+1} = \max_{i=1,\cdots,m}[\mathrm{lin}_{g_i}(x^{k+1}, y^k)]_+$ and

$$x^{k+1} \in \mathop{\mathrm{Argmin}}_{x \in C} \; P_1(x) - \langle \xi^k, x \rangle + \theta_k \max_{i=1,\cdots,m} [\mathrm{lin}_{g_i}(x, y^k)]_+ + \frac{\theta_k L_g}{2}\|x - y^k\|^2. \tag{3.2}$$

Since problem (3.2) has a unique solution as an optimization problem with a nonempty closed convex feasible set and a (real-valued) strongly convex objective, we conclude that the subproblem in (3.4) has a unique solution. While this subproblem requires an iterative solver in general, we refer the readers to [41, Appendix A] for an efficient routine for solving the subproblem in (3.4) with some specific P_1 when $m = 1$.

The convergence properties of our algorithm will be studied in Sect. 4, and we end this section by presenting some useful facts concerning the subproblem (3.4). The first two items are simple observations already established in the preceding discussions, and they are stated here for easy reference later.

Lemma 3.1 *Suppose that $x^k \in C$ is generated at the beginning of the k-th iteration of Algorithm 1 for some $k \geq 0$. Then the following statements hold:*

(i) $s^{k+1} = \max_{i=1,\cdots,m}[\mathrm{lin}_{g_i}(x^{k+1}, y^k)]_+$.

(ii) *Problem (3.4) has a unique solution.*

(iii) *Let $g_0 := 0$. Then x^{k+1} is a component of the minimizer of the subproblem in (3.4) if and only if there exist $\lambda_i^k \geq 0$ for all $i \in I_k(x^{k+1})$ such that $\sum_{i \in I_k(x^{k+1})} \lambda_i^k = 1$ and*

$$0 \in \partial P_1(x^{k+1}) - \xi^k + \theta_k \sum_{i \in I_k(x^{k+1})} \lambda_i^k \nabla g_i(y^k) + \theta_k L_g (x^{k+1} - y^k) + \mathcal{N}_C(x^{k+1}),$$

[1] More precisely, when the P in (1.1) is smooth with Lipschitz gradient (say, with modulus L_P) and $\beta_k \equiv 0$, our algorithm applied to (1.1) with $P_1(x) := \frac{L_P}{2}\|x\|^2$ and $P_2(x) := \frac{L_P}{2}\|x\|^2 - P(x)$ becomes an instance of the ESQM in [4].

Algorithm 1 ESQM$_e$ for solving (1.1)

Step 0. Choose $x^{-1} = x^0 \in C$, $\theta_0 > 0$, $d > 0$, and $\{\beta_k\} \subseteq \left[0, \sqrt{\frac{L_g}{L_g+\ell_g}}\right)$ with $\bar{\beta} := \sup_k \beta_k < \sqrt{\frac{L_g}{L_g+\ell_g}}$, where $L_g = \max\{L_{g_i} : i = 1, \ldots, m\}$ and $\ell_g = \max\{\ell_{g_i} : i = 1, \ldots, m\}$ as in Remark 3.1. Set $k = 0$.

Step 1. Set

$$y^k = x^k + \beta_k(x^k - x^{k-1}). \tag{3.3}$$

Step 2. Take any $\xi^k \in \partial P_2(x^k)$ and compute

$$\begin{aligned}(x^{k+1}, s^{k+1}) \in \underset{(x,s)\in\mathbb{R}^{n+1}}{\operatorname{Argmin}} \quad & P_1(x) - \langle \xi^k, x\rangle + \theta_k s + \frac{\theta_k L_g}{2}\|x - y^k\|^2 \\ \text{s.t.} \quad & \operatorname{lin}_{g_i}(x, y^k) \le s, \quad i = 1, \ldots, m, \\ & (x, s) \in C \times \mathbb{R}_+,\end{aligned} \tag{3.4}$$

where lin_{g_i} is defined in (3.1).

Step 3. If $\operatorname{lin}_{g_i}(x^{k+1}, y^k) \le 0$ for all i, then $\theta_{k+1} = \theta_k$; otherwise $\theta_{k+1} = \theta_k + d$. Update $k \leftarrow k+1$ and go to step 1.

where

$$I_k(x) := \left\{\iota \in \{0, 1, \cdots, m\} : \operatorname{lin}_{g_\iota}(x, y^k) = \max_{i=0,1,\cdots,m} \operatorname{lin}_{g_i}(x, y^k)\right\}. \tag{3.3}$$

Proof Items (i) and (ii) were established in the discussions preceding this lemma.

We now prove (iii). Recall that x^{k+1} is a component of the minimizer of the subproblem in (3.4) if and only if it is a minimizer of the convex problem (3.2). Using $g_0 \equiv 0$ and [34, Theorem 23.8], this is further equivalent to

$$\begin{aligned}0 &\in \partial P_1(x^{k+1}) - \xi^k + \theta_k \partial\left(\max_{i=0,1,\cdots,m}\{\operatorname{lin}_{g_i}(\cdot, y^k)\}\right)(x^{k+1}) + \theta_k L_g(x^{k+1} - y^k) + \mathcal{N}_C(x^{k+1}) \\ &\overset{\text{(a)}}{=} \partial P_1(x^{k+1}) - \xi^k + \theta_k \operatorname{conv}\{\nabla g_i(y^k) : i \in I_k(x^{k+1})\} + \theta_k L_g(x^{k+1} - y^k) + \mathcal{N}_C(x^{k+1}),\end{aligned}$$

where (a) follows from [35, Exercise 8.31] with $I_k(\cdot)$ defined in (3.3). □

4 Convergence properties

4.1 Convergence analysis for ESQM$_e$

We first show that the successive changes of the $\{x^k\}$ generated by ESQM$_e$ vanish.

Theorem 4.1 *(Vanishing successive changes) Consider (1.1) and let $\{(x^k, y^k, \theta_k)\}$ be generated by Algorithm 1. Then the following statements hold:*

(i) *The sequence $\{x^k\}$ belongs to C and is bounded.*

(ii) *Let* $\bar{m} := \inf\{P(x) : x \in C\}$. *Then* $\bar{m} \in \mathbb{R}$ *and for any* $k \geq 1$,

$$Q(x^{k+1}, x^k, y^k, \theta_{k+1}) \leq Q(x^k, x^{k-1}, y^{k-1}, \theta_k) - \left(1 - \frac{L_g + \ell_g}{L_g}\beta_k^2\right)\frac{L_g}{2}\|x^k - x^{k-1}\|^2,$$

where

$$Q(x, y, z, \theta) := \frac{P(x) - \bar{m}}{\theta} + \max_{i=1,\cdots,m}\left[\mathrm{lin}_{g_i}(x, z)\right]_+ + \frac{L_g}{2}\|x - y\|^2 + \frac{L_g}{2}\|x - z\|^2.$$

(iii) *It holds that* $\sum_{k=1}^{\infty} \frac{L_g - (L_g + \ell_g)\beta_k^2}{2}\|x^k - x^{k-1}\|^2 < \infty$, *and* $\lim_{k\to\infty}\|x^k - x^{k-1}\| = 0$ *and* $\lim_{k\to\infty}\|x^k - y^k\| = 0$.

Proof (i): Note that $\{x^k\} \subseteq C$ according to (3.4). Since C is compact, $\{x^k\}$ is bounded.

(ii): Notice that the objective in (3.2) is strongly convex with x^{k+1} being its unique minimizer over C. Using this, and noting that $s^{k+1} = \max_{i=1,\ldots,m}\left[\mathrm{lin}_{g_i}(x^{k+1}, y^k)\right]_+$ (see Lemma 3.1(i)), we have for any $k \geq 0$ that

$$\begin{aligned}
&P_1(x^{k+1}) - \langle \xi^k, x^{k+1} - x^k\rangle + \theta_k s^{k+1} + \frac{\theta_k L_g}{2}\|x^{k+1} - y^k\|^2 \\
&= P_1(x^{k+1}) - \langle \xi^k, x^{k+1} - x^k\rangle + \theta_k \max_{i=1,\ldots,m}\left[\mathrm{lin}_{g_i}(x^{k+1}, y^k)\right]_+ + \frac{\theta_k L_g}{2}\|x^{k+1} - y^k\|^2 \\
&\leq P_1(x^k) + \theta_k \max_{i=1,\ldots,m}\left[\mathrm{lin}_{g_i}(x^k, y^k)\right]_+ + \frac{\theta_k L_g}{2}\|x^k - y^k\|^2 - \frac{\theta_k L_g}{2}\|x^{k+1} - x^k\|^2.
\end{aligned} \tag{4.1}$$

Meanwhile, from Remark 3.1 and the definition of lin_{g_i} in (3.1), we see that whenever $k \geq 1$,

$$\begin{aligned}
&\max_{i=1,\cdots,m}\left[\mathrm{lin}_{g_i}(x^k, y^k)\right]_+ \\
&= \max_{i=1,\cdots,m}\left[g_i^1(y^k) + \langle \nabla g_i^1(y^k), x^k - y^k\rangle - g_i^2(y^k) - \langle \nabla g_i^2(y^k), x^k - y^k\rangle\right]_+ \\
&\overset{(a)}{\leq} \max_{i=1,\cdots,m}\left[g_i^1(x^k) - g_i^2(x^k) + \frac{\ell_{g_i}}{2}\|x^k - y^k\|^2\right]_+ = \max_{i=1,\cdots,m}\left[g_i(x^k) + \frac{\ell_{g_i}}{2}\|x^k - y^k\|^2\right]_+ \\
&\overset{(b)}{\leq} \max_{i=1,\cdots,m}\left[\mathrm{lin}_{g_i}(x^k, y^{k-1}) + \frac{L_{g_i}}{2}\|x^k - y^{k-1}\|^2 + \frac{\ell_{g_i}}{2}\|x^k - y^k\|^2\right]_+ \\
&\overset{(c)}{\leq} \max_{i=1,\cdots,m}\left[\mathrm{lin}_{g_i}(x^k, y^{k-1})\right]_+ + \frac{L_g}{2}\|x^k - y^{k-1}\|^2 + \frac{\ell_g}{2}\|x^k - y^k\|^2,
\end{aligned} \tag{4.2}$$

where (a) holds because of the convexity of g_i^1 and the Lipschitz continuity of ∇g_i^2, (b) follows from the Lipschitz continuity of ∇g_i, and (c) holds because $L_g = \max\{L_{g_i} : i = 1, \ldots, m\}$ and $\ell_g = \max\{\ell_{g_i} : i = 1, \ldots, m\}$. Then, we obtain that when $k \geq 1$,

$$
\begin{aligned}
P(x^{k+1}) &= P_1(x^{k+1}) - P_2(x^{k+1}) \overset{\text{(a)}}{\leq} P_1(x^{k+1}) - \langle \xi^k, x^{k+1} - x^k \rangle - P_2(x^k) \\
&= P_1(x^{k+1}) - \langle \xi^k, x^{k+1} - x^k \rangle + \frac{\theta_k L_g}{2}\|x^{k+1} - y^k\|^2 - \frac{\theta_k L_g}{2}\|x^{k+1} - y^k\|^2 - P_2(x^k) \\
&\overset{\text{(b)}}{\leq} P_1(x^k) + \frac{\theta_k L_g}{2}\|x^k - y^k\|^2 - \theta_k s^{k+1} + \theta_k \max_{i=1,\cdots,m}\left[\mathrm{lin}_{g_i}(x^k, y^k)\right]_+ \\
&\quad - \frac{\theta_k L_g}{2}\|x^{k+1} - x^k\|^2 - \frac{\theta_k L_g}{2}\|x^{k+1} - y^k\|^2 - P_2(x^k) \\
&\overset{\text{(c)}}{\leq} P(x^k) + \theta_k\left(\max_{i=1,\cdots,m}\left[\mathrm{lin}_{g_i}(x^k, y^{k-1})\right]_+ + \frac{L_g}{2}\|x^k - y^{k-1}\|^2 + \frac{\ell_g}{2}\|x^k - y^k\|^2\right) \\
&\quad + \frac{\theta_k L_g}{2}\|x^k - y^k\|^2 - \theta_k s^{k+1} - \frac{\theta_k L_g}{2}\|x^{k+1} - x^k\|^2 - \frac{\theta_k L_g}{2}\|x^{k+1} - y^k\|^2,
\end{aligned}
$$

where (a) holds because P_2 is convex and $\xi^k \in \partial P_2(x^k)$, (b) holds thanks to (4.1), and (c) holds because of (4.2).

Rearranging terms in the above display and noting that $y^k - x^k = \beta_k(x^k - x^{k-1})$ for $k \geq 0$ (thanks to the definition of y^k in (3.3)), we have that for $k \geq 1$,

$$
\begin{aligned}
&P(x^{k+1}) + \theta_k s^{k+1} + \frac{\theta_k L_g}{2}\|x^{k+1} - x^k\|^2 + \frac{\theta_k L_g}{2}\|x^{k+1} - y^k\|^2 \\
&\leq P(x^k) + \theta_k \max_{i=1,\cdots,m}\left[\mathrm{lin}_{g_i}(x^k, y^{k-1})\right]_+ + \frac{\theta_k L_g}{2}\|x^k - y^{k-1}\|^2 \\
&\quad + \frac{\theta_k (L_g + \ell_g)}{2}\beta_k^2\|x^k - x^{k-1}\|^2 \\
&= P(x^k) + \theta_k \max_{i=1,\cdots,m}[\mathrm{lin}_{g_i}(x^k, y^{k-1})]_+ + \frac{\theta_k L_g}{2}\|x^k - x^{k-1}\|^2 \\
&\quad + \frac{\theta_k L_g}{2}\|x^k - y^{k-1}\|^2 - \left(1 - \frac{L_g + \ell_g}{L_g}\beta_k^2\right)\frac{\theta_k L_g}{2}\|x^k - x^{k-1}\|^2. \qquad (4.3)
\end{aligned}
$$

Since P is continuous and C is a nonempty compact set, we see that $\bar{m} = \inf\{P(x) : x \in C\} \in \mathbb{R}$. Then we can deduce from the definition of Q and the observation $s^{k+1} = \max_{i=1,\cdots,m}\left[\mathrm{lin}_{g_i}(x^{k+1}, y^k)\right]_+$ (thanks to Lemma 3.1(i)) that whenever $k \geq 1$,

$$
\begin{aligned}
&Q(x^{k+1}, x^k, y^k, \theta_{k+1}) \\
&= \frac{P(x^{k+1}) - \bar{m}}{\theta_{k+1}} + s^{k+1} + \frac{L_g}{2}\|x^{k+1} - x^k\|^2 + \frac{L_g}{2}\|x^{k+1} - y^k\|^2 \\
&\overset{\text{(a)}}{\leq} \frac{P(x^{k+1}) - \bar{m}}{\theta_k} + s^{k+1} + \frac{L_g}{2}\|x^{k+1} - x^k\|^2 + \frac{L_g}{2}\|x^{k+1} - y^k\|^2 \\
&\overset{\text{(b)}}{\leq} \frac{1}{\theta_k}\Bigg[P(x^k) - \bar{m} + \theta_k \max_{i=1,\cdots,m}[\mathrm{lin}_{g_i}(x^k, y^{k-1})]_+ + \frac{\theta_k L_g}{2}\|x^k - x^{k-1}\|^2 \\
&\quad + \frac{\theta_k L_g}{2}\|x^k - y^{k-1}\|^2 - \left(1 - \frac{L_g + \ell_g}{L_g}\beta_k^2\right)\frac{\theta_k L_g}{2}\|x^k - x^{k-1}\|^2\Bigg]
\end{aligned}
$$

$$= Q(x^k, x^{k-1}, y^{k-1}, \theta_k) - \left(1 - \frac{L_g + \ell_g}{L_g}\beta_k^2\right)\frac{L_g}{2}\|x^k - x^{k-1}\|^2, \tag{4.4}$$

where (a) holds because of the definition of $\bar{m}$ and the facts that $x^{k+1} \in C$ and $\{\theta_k^{-1}\}$ is nonincreasing, and (b) follows from (4.3) and the fact that $\frac{1}{\theta_k} > 0$.

(iii): Observe that, for any $k \geq 0$,

$$\begin{aligned}
&Q(x^{k+1}, x^k, y^k, \theta_{k+1})\\
&= \frac{P(x^{k+1}) - \bar{m}}{\theta_{k+1}} + \max_{i=1,\cdots,m}[\mathrm{lin}_{g_i}(x^{k+1}, y^k)]_+ + \frac{L_g}{2}\|x^{k+1} - x^k\|^2 + \frac{L_g}{2}\|x^{k+1} - y^k\|^2 \geq 0.
\end{aligned}$$

Combining the above display with item (ii), we have

$$\begin{aligned}
&\sum_{k=1}^{\infty}\left(1 - \frac{L_g + \ell_g}{L_g}\beta_k^2\right)\frac{L_g}{2}\|x^k - x^{k-1}\|^2\\
&\leq Q(x^1, x^0, y^0, \theta_1) - \liminf_{k\to\infty} Q(x^{k+1}, x^k, y^k, \theta_{k+1}) \leq Q(x^1, x^0, y^0, \theta_1) < \infty.
\end{aligned}$$

Finally, since $\sup_k \beta_k < \sqrt{\frac{L_g}{L_g + \ell_g}}$, we can deduce from the above display that

$$\lim_{k\to\infty}\|x^k - x^{k-1}\| = 0.$$

Combining this with the definition of y^k in (3.3), we can obtain further that $\lim_{k\to\infty}\|y^k - x^k\| = \lim_{k\to\infty}\beta_k\|x^k - x^{k-1}\| = 0$. □

Next, we recall the following assumption involving the RCQ in Definition 2.1. This assumption was first introduced in [4, Assumption (A1)] for studying ESQM.

Assumption 4.1 For (1.1), the $RCQ(x)$ holds at every $x \in C \cap \mathscr{F}$, and for every $x \in C\backslash\mathscr{F}$, there cannot exist u_i, $i \in I(x)$, such that

$$u_i \geq 0 \;\; \forall i \in I(x), \quad \sum_{i\in I(x)} u_i = 1, \quad \left\langle \sum_{i\in I(x)} u_i \nabla g_i(x), z - x \right\rangle \geq 0 \;\; \forall z \in C, \tag{4.5}$$

where $I(x) := \left\{\iota \in \{1, \ldots, m\} : g_\iota(x) = \max_{i=1,\ldots,m}[g_i(x)]_+\right\}$.[2]

Remark 4.1 (i) Using [4, Remark 2.1], one can deduce that if Assumption 4.1 holds, then for any $x \in C$, there cannot exist u_i, $i \in I(x)$, such that (4.5) holds.

(ii) From [4, Remark 2.2], we know that if the $RCQ(x)$ holds at every $x \in C$, then Assumption 4.1 holds.

[2] We would like to point out that while our definition of $I(x)$ seems to look slightly different from the corresponding definition, namely $T(x)$, in [4] (see the discussions before [4, Eq. (6)]), one can check that the two definitions are equivalent.

Using Assumption 4.1 and Theorem 4.1, we will prove in the next theorem that the sequence $\{\theta_k\}$ in Algorithm 1 is bounded. The same conclusion was established for ESQM in [4, Theorem 3.1(b)].

Theorem 4.2 *(*Boundedness of $\{\theta_k\}$*) Consider (1.1) and suppose that Assumption 4.1 holds. Let $\{(s^k, \theta_k)\}$ be generated by Algorithm 1, $\mathfrak{A} := \{k \in \mathbb{N} : \theta_{k+1} > \theta_k\}$, and let $|\mathfrak{A}|$ denote the cardinality of $\mathfrak{A}$. Then $|\mathfrak{A}|$ is finite, i.e., there exists $N_0 \in \mathbb{N}$ such that $\theta_k \equiv \theta_{N_0}$ whenever $k \geq N_0$. Moreover, $s^{k+1} = 0$ whenever $k \geq N_0$.*

Proof Suppose to the contrary that $|\mathfrak{A}| = \infty$. Then by the definition of θ_k in Step 3 of Algorithm 1, we have $\lim_{k\to\infty} \theta_k = \infty$ and $\lim_{k\to\infty} \theta_k^{-1} = 0$.

We first claim that for each i, there exists $n_i \in \mathbb{N}$ such that for all $k \geq n_i$,

$$g_i(y^k) + \langle \nabla g_i(y^k), x^{k+1} - y^k \rangle \leq 0,$$

where $\{(x^k, y^k)\}$ is generated by Algorithm 1.

Suppose not. Then there exists $i_0 \in \{1, \dots, m\}$ and (infinite) subsequences $\{x^{k_j}\}$ and $\{y^{k_j}\}$ such that

$$g_{i_0}(y^{k_j}) + \langle \nabla g_{i_0}(y^{k_j}), x^{k_j+1} - y^{k_j} \rangle > 0 \quad \forall j.$$

Using this and recalling the definition of $I_k(\cdot)$ in (3.3), we have that

$$\mathrm{lin}_{g_i}(x^{k_j+1}, y^{k_j}) > 0 \ \ \forall i \in I_{k_j}(x^{k_j+1}), \ \forall j.$$

In particular, $0 \notin I_{k_j}(x^{k_j+1})$ (see (3.1)). Now, in view of the finiteness of $\left\{I_{k_j}(x^{k_j+1})\right\}$ (since $I_{k_j}(x^{k_j+1}) \subseteq \{1, \dots, m\}$ for all j), by passing to a further subsequence if necessary, we deduce that there exists a nonempty subset $I_0 \subseteq \{1, \dots, m\}$ such that $I_{k_j}(x^{k_j+1}) \equiv I_0$ for all j. That is, for all $i \in I_0$,

$$\mathrm{lin}_{g_i}(x^{k_j+1}, y^{k_j}) = \max_{l=0,1,\dots,m} \left\{\mathrm{lin}_{g_l}(x^{k_j+1}, y^{k_j})\right\} > 0 \ \forall j. \tag{4.6}$$

In addition, from Lemma 3.1(iii), we have that for each k_j, there exist $\lambda_i^{k_j} \geq 0$ for each $i \in I_{k_j}(x^{k_j+1}) \equiv I_0$, such that $\sum_{i\in I_0} \lambda_i^{k_j} = 1$ and

$$0 \in \theta_{k_j}^{-1}(\partial P_1(x^{k_j+1}) - \xi^{k_j}) + L_g(x^{k_j+1} - y^{k_j}) + \sum_{i\in I_0} \lambda_i^{k_j} \nabla g_i(y^{k_j}) + \mathcal{N}_C(x^{k_j+1}). \tag{4.7}$$

Now, since the sequences $\{x^k\} \subseteq C$ and $\{\lambda_i^{k_j}\}$ (for each $i \in I_0$) are bounded, by passing to a further subsequence if necessary, we assume that $\lim_{j\to\infty} x^{k_j} = x^*$ for some x^* and that for each $i \in I_0$, $\lim_{j\to\infty} \lambda_i^{k_j} = \bar{\lambda}_i$ for some $\bar{\lambda}_i$. Then $x^* \in C$, $\bar{\lambda}_i \geq 0$ (for each $i \in I_0$), $\sum_{i\in I_0} \bar{\lambda}_i = 1$ and $I_0 \subseteq \left\{\iota \in \{0, 1, \cdots, m\} : g_\iota(x^*) = \max_{i=0,1,\cdots,m} g_i(x^*)\right\}$

(thanks to (4.6), (3.1) and Theorem 4.1(iii), and recall that $g_0 \equiv 0$ from Lemma 3.1(iii)). Since $0 \notin I_0$, we see that

$$I_0 \subseteq I(x^*) = \left\{ \iota \in \{1, \cdots, m\} : g_\iota(x^*) = \max_{i=1,\cdots,m} [g_i(x^*)]_+ \right\},$$

where $I(x)$ was defined in Assumption 4.1. Passing to the limit in (4.7), and noting that $\lim_{j\to\infty} \theta_{k_j}^{-1} = 0$, $\lim_{k\to\infty} \|x^{k+1} - y^k\| = \lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$ (thanks to Theorem 4.1(iii)) and the fact that $\{\partial P_1(x^{k_j+1})\}$ and $\{\xi^{k_j}\}$ are uniformly bounded (thanks to the real-valuedness and convexity of P_1, P_2, the compactness of C and [34, Theorem 24.7]), we have upon invoking the closedness of $x \mapsto \mathcal{N}_C(x)$ that

$$0 \in \sum_{i\in I_0} \bar{\lambda}_i \nabla g_i(x^*) + \mathcal{N}_C(x^*),$$

which implies that

$$\left\langle \sum_{i\in I_0} \bar{\lambda}_i \nabla g_i(x^*), x - x^* \right\rangle \geq 0 \quad \forall x \in C.$$

Since $I_0 \subseteq I(x^*)$, this contradicts Assumption 4.1 in view of Remark 4.1(i).

Therefore, if $|\mathfrak{A}| = \infty$, then it must hold that for each i, there exists $n_i \in \mathbb{N}$, such that for any $k \geq n_i$,

$$g_i(y^k) + \langle \nabla g_i(y^k), x^{k+1} - y^k \rangle \leq 0.$$

Let $N_* := \max_{i=1,\ldots,m} n_i$. Then for all $i \in \{1, \ldots, m\}$ and for any $k \geq N_*$, we have

$$g_i(y^k) + \langle \nabla g_i(y^k), x^{k+1} - y^k \rangle \leq 0.$$

In view of this and the definition of θ_k in Step 3 of Algorithm 1, we must have $\theta_k \equiv \theta_{N_*}$ for all $k \geq N_*$, which contradicts $\theta_k \to \infty$. Thus, it must hold that $|\mathfrak{A}| < \infty$.

Since $|\mathfrak{A}|$ is finite, there exists $N_0 \in \mathbb{N}$, such that $\theta_k \equiv \theta_{N_0}$ whenever $k \geq N_0$. From Step 3 of Algorithm 1, we know that for each i, $g_i(y^k) + \langle \nabla g_i(y^k), x^{k+1} - y^k \rangle \leq 0$, for all $k \geq N_0$. Then Lemma 3.1(i) asserts that $s^{k+1} = 0$ for any $k \geq N_0$. □

We are now ready to prove that any cluster point of the $\{x^k\}$ generated by Algorithm 1 is a critical point of (1.1).

Theorem 4.3 (Subsequential convergence) Consider (1.1) and suppose that Assumption 4.1 holds. Let $\{x^k\}$ be generated by Algorithm 1. Then for any accumulation point $\bar{x}$ of $\{x^k\}$, there exists $\bar{\lambda}_i \geq 0$ for each $i \in \tilde{I}(\bar{x})$ such that $\sum_{i\in\tilde{I}(\bar{x})} \bar{\lambda}_i = 1$ and

$$0 \in \partial P_1(\bar{x}) - \partial P_2(\bar{x}) + \theta_{N_0} \sum_{i\in\tilde{I}(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}) + \mathcal{N}_C(\bar{x}), \tag{4.8}$$

where $\tilde{I}(\bar{x}) := \left\{\iota \in \{0, 1, \cdots, m\} : g_\iota(\bar{x}) = \max_{i=0,1,\cdots,m}\{g_i(\bar{x})\}\right\}$, $g_0 = 0$, and θ_{N_0} is defined in Theorem 4.2; moreover, $\bar{x}$ is a critical point of (1.1).

Proof Suppose that $\bar{x}$ is an accumulation point of $\{x^k\}$ with $\lim_{j\to\infty} x^{k_j} = \bar{x}$ for some convergent subsequence $\{x^{k_j}\}$. Let $\{\xi^k\}$ be generated in Algorithm 1 and $\{\lambda_i^k\}$ with $i \in I_k(x^{k+1})$ be as in Lemma 3.1(iii). Then, in view of the finiteness of $\{I_{k_j}(x^{k_j+1})\}$ (since $I_{k_j}(x^{k_j+1}) \subseteq \{0, 1, \ldots, m\}$ for all j), by passing to a further subsequence if necessary, we see that there exists a nonempty subset $I_0 \subseteq \{0, 1, \ldots, m\}$ such that $I_{k_j}(x^{k_j+1}) \equiv I_0$. Moreover, $\{\lambda_i^{k_j}\}$ for each $i \in I_{k_j}(x^{k_j+1}) \equiv I_0$ is bounded as sequences of nonnegative numbers at most 1, and $\{\xi^k\}$ is bounded thanks to the real-valuedness and convexity of P_2 and [34, Theorem 24.7]. Passing to a further subsequence if necessary, we assume without loss of generality that $\lim_{j\to\infty} \lambda_i^{k_j} = \bar{\lambda}_i \geq 0$ for each $i \in I_0$ and $\lim_{j\to\infty} \xi^{k_j} = \bar{\xi}$; moreover, the property of $\{\lambda_i^{k_j}\}$ with $i \in I_{k_j}(x^{k_j+1}) \equiv I_0$ guaranteed by Lemma 3.1(iii) asserts that for all j, it holds that

$$\begin{aligned} &0 \in \partial P_1(x^{k_j+1}) - \xi^{k_j} + \theta_{k_j} L_g(x^{k_j+1} - y^{k_j}) + \theta_{k_j} \sum_{i\in I_0} \lambda_i^{k_j} \nabla g_i(y^{k_j}) + \mathcal{N}_C(x^{k_j+1}) \\ &\text{and } \sum_{i\in I_0} \lambda_i^{k_j} = 1, \quad \lambda_i^{k_j} \geq 0 \ \forall i \in I_{k_j}(x^{k_j+1}) \equiv I_0. \end{aligned} \tag{4.9}$$

In addition, in view of (3.4), we obtain that for each j,

$$g_i(y^{k_j}) + \langle \nabla g_i(y^{k_j}), x^{k_j+1} - y^{k_j}\rangle \leq s^{k_j+1} \quad \forall i = 1, \ldots, m. \tag{4.10}$$

Now, note that $\lim_{k\to\infty} \|x^k - x^{k-1}\| = \lim_{k\to\infty} \|x^{k+1} - y^k\| = 0$ (thanks to Theorem 4.1(iii)), $s^{k_j+1} = 0$ and $\theta_{k_j} \equiv \theta_{N_0}$ whenever $k_j \geq N_0$ (thanks to Theorem 4.2). Passing to the limit in (4.10) and (4.9), we see that

$$g_i(\bar{x}) \leq 0 \ \forall i = 1, \ldots, m, \ \sum_{i\in I_0} \bar{\lambda}_i = 1, \quad \bar{\lambda}_i \geq 0 \ \forall i \in I_0, \tag{4.11}$$

and

$$0 \in \partial P_1(\bar{x}) - \partial P_2(\bar{x}) + \theta_{N_0} \sum_{i\in I_0} \bar{\lambda}_i \nabla g_i(\bar{x}) + \mathcal{N}_C(\bar{x}). \tag{4.12}$$

where we also invoked the closedness of ∂P_1, ∂P_2 and $\mathcal{N}_C$ to deduce (4.12). Furthermore, we have from the definition of $I_{k_j}(x^{k_j+1})$ in (3.3) (and recall that $I_{k_j}(x^{k_j+1}) \equiv I_0$) and Theorem 4.1(iii) that

$$I_0 \subseteq \tilde{I}(\bar{x}) := \left\{\iota \in \{0, 1, \cdots, m\} : g_\iota(\bar{x}) = \max_{i=0,1,\cdots,m}\{g_i(\bar{x})\}\right\}. \tag{4.13}$$

Then the inclusion (4.8) follows from (4.12) and (4.11) upon noting $I_0 \subseteq \tilde{I}(\bar{x})$ (see (4.13)) and defining $\bar{\lambda}_i = 0$ for $i \in \tilde{I}(\bar{x}) \backslash I_0$.

Finally, let $\hat{\lambda}_i := \theta_{N_0}\bar{\lambda}_i \geq 0$ for all $i \in I_0 \cap \{1, \cdots, m\}$, and $\hat{\lambda}_i = 0$ for all $i \in \{1, \cdots, m\} \setminus I_0$. Then by (4.11) and $I_0 \subseteq \tilde{I}(\bar{x})$ (see (4.13)), we have that

$$\hat{\lambda}_i g_i(\bar{x}) = 0 \;\; \forall i = 1, \ldots, m; \tag{4.14}$$

indeed, for each $i \in I_0$, we have $g_i(\bar{x}) = 0$, and for each $i \notin I_0$, we have $\hat{\lambda}_i = 0$.

Notice that $\nabla g_0(\bar{x}) = 0$ (thanks to $g_0 \equiv 0$). Using the definition of $\hat{\lambda}_i$ and (4.12), we have

$$0 \in \partial P_1(\bar{x}) - \partial P_2(\bar{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\bar{x}) + \mathcal{N}_C(\bar{x}). \tag{4.15}$$

Combining (4.11), (4.14), (4.15) and the above definition of $\hat{\lambda}$, we conclude that $\bar{x}$ is a critical point of (1.1). □

We next derive the global convergence property of the $\{x^k\}$ generated by Algorithm 1. We will need to make use of the following function,

$$H(x, y, z) := \frac{P(x) - \bar{m}}{\hat{\theta}} + \max_{i=1,\ldots,m} [\mathrm{lin}_{g_i}(x, z)]_+ + \frac{L_g}{2}\|x - y\|^2 + \frac{L_g}{2}\|x - z\|^2 + \delta_C(x), \tag{4.16}$$

where $\bar{m}$ is defined in Theorem 4.1(ii), and $\hat{\theta} := \theta_{N_0}$ with N_0 defined in Theorem 4.2. Our analysis follows the nowadays standard convergence arguments based on Kurdyka-Łojasiewicz property; see, for example, [2, 3, 13]. In essence, under Assumption 4.1, we will show that H has sufficient descent along the sequence $\{(x^{k+1}, x^k, y^k)\}$ for all sufficiently large k, and H is constant on the set of accumulation points of $\{(x^{k+1}, x^k, y^k)\}$. We will also show that $\mathrm{dist}(0, \partial H(x^{k+1}, x^k, y^k))$ is suitably bounded by successive changes of the iterates by imposing additional differentiability assumptions on each g_i and P_2. These together with an additional assumption that H satisfies the KL property will be used to establish global convergence of the $\{x^k\}$ generated by Algorithm 1.

We start with a remark concerning the sufficient descent property.

Remark 4.2 (Sufficient descent) Consider (1.1) and suppose that Assumption 4.1 holds. Notice from the definition of Q in Theorem 4.1(ii) and that of H in (4.16) that $H(x, y, z) = Q(x, y, z, \hat{\theta}) + \delta_C(x)$. Now, according to Theorem 4.2, we have $\theta_k \equiv \theta_{N_0} = \hat{\theta}$ for all $k \geq N_0$. Thus, we have $H(x^k, x^{k-1}, y^{k-1}) = Q(x^k, x^{k-1}, y^{k-1}, \hat{\theta})$ for all $k \geq N_0$, where $\{(x^k, y^k)\}$ is generated by Algorithm 1. Then one can see that the sequence $\{H(x^{k+1}, x^k, y^k)\}_{k \geq N_0}$ is nonincreasing thanks to Theorem 4.1(ii), and it holds that

$$H(x^{k+1}, x^k, y^k) \leq H(x^k, x^{k-1}, y^{k-1}) - \frac{L_g - (L_g + \ell_g)\bar{\beta}^2}{2}\|x^k - x^{k-1}\|^2 \quad \forall k \geq N_0,$$

where $\bar{\beta} = \sup_k \beta_k$, and notice that $L_g > (L_g + \ell_g)\bar{\beta}^2$ thanks to the choice of $\{\beta_k\}$.

Lemma 4.1 *Consider (1.1) and suppose that Assumption 4.1 holds. Let $\{(x^k, y^k)\}$ be generated by Algorithm 1, H be defined in (4.16), and Ω be the set of accumulation points of $\{(x^{k+1}, x^k, y^k)\}$. Then Ω is a nonempty compact set, $\omega := \lim_{k\to\infty} H(x^{k+1}, x^k, y^k)$ exists, and $H \equiv \omega$ on Ω.*

Proof From Theorem 4.1(i), we have that the set of accumulation points of $\{x^k\}$, denoted by Λ, is a nonempty compact set. Since $\lim_{k\to\infty} \|x^k - x^{k-1}\| = \lim_{k\to\infty} \|x^k - y^k\| = 0$ thanks to Theorem 4.1(iii), one can see that $\Omega = \{(\bar{x}, \bar{x}, \bar{x}) : \bar{x} \in \Lambda\}$, which is a nonempty compact set.

Next, according to Remark 4.2, the sequence $\{H(x^{k+1}, x^k, y^k)\}_{k\geq N_0}$ is nonincreasing. Moreover, one can see from the definition of H (see (4.16)) that $\{H(x^{k+1}, x^k, y^k)\}$ is bounded from below (by zero). Thus, $\omega := \lim_{k\to\infty} H(x^{k+1}, x^k, y^k)$ exists.

For any $(\bar{x}, \bar{x}, \bar{x}) \in \Omega$, let $\{x^{k_j}\}$ be a convergent subsequence with $\lim_{j\to\infty} x^{k_j} = \bar{x}$. Since P and each g_i are continuous, and $\lim_{k\to\infty} \|x^k - x^{k-1}\| = \lim_{k\to\infty} \|x^k - y^k\| = 0$ (see Theorem 4.1(iii)), we obtain that

$$
\begin{aligned}
H(\bar{x}, \bar{x}, \bar{x}) &= \frac{P(\bar{x}) - \bar{m}}{\hat{\theta}} + \max_{i=1,\cdots,m} \left[\mathrm{lin}_{g_i}(\bar{x}, \bar{x})\right]_+ \\
&= \lim_{j\to\infty} \frac{P(x^{k_j+1}) - \bar{m}}{\hat{\theta}} + \max_{i=1,\cdots,m} \left[\mathrm{lin}_{g_i}(x^{k_j+1}, y^{k_j})\right]_+ \\
&\quad + \frac{L_g}{2}\|x^{k_j+1} - x^{k_j}\|^2 + \frac{L_g}{2}\|x^{k_j+1} - y^{k_j}\|^2 \\
&= \lim_{j\to\infty} H(x^{k_j+1}, x^{k_j}, y^{k_j}) = \lim_{k\to\infty} H(x^{k+1}, x^k, y^k) = \omega.
\end{aligned}
$$

Since $(\bar{x}, \bar{x}, \bar{x}) \in \Omega$ is arbitrary, we conclude that $H \equiv \omega$ on Ω. □

Next, we introduce an assumption for deriving a bound on $\mathrm{dist}(0, \partial H(x^{k+1}, x^k, y^k))$. This assumption was also used in [38, 40] and is satisfied in many applications; see [38].

Assumption 4.2 Each g_i in (1.1) is twice continuously differentiable. The function P_2 is continuously differentiable on an open set U_0 containing $\mathcal{X}$, and ∇P_2 is locally Lipschitz continuous on U_0, where $\mathcal{X}$ is the set of critical points of (1.1).

Now, we present the following bound on $\mathrm{dist}(0, \partial H(x^{k+1}, x^k, y^k))$.

Lemma 4.2 *Consider (1.1) and suppose that Assumptions 4.1 and 4.2 hold. Let $\{(x^k, y^k)\}$ be generated by Algorithm 1 and H be defined in (4.16). Then there exist $\tau > 0$ and $N_1 \in \mathbb{N}$ such that for all $k \geq N_1$, we have*

$$
\mathrm{dist}(0, \partial H(x^{k+1}, x^k, y^k)) \leq \tau(\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|).
$$

Proof Let Λ be the set of accumulation points of $\{x^k\}$. Then Λ is nonempty and compact in view of Theorem 4.1(i), and $\Lambda \subseteq \mathcal{X}$ thanks to Theorem 4.3, where $\mathcal{X}$ is defined in Assumption 4.2. Moreover, we have $\lim_{k\to\infty} \mathrm{dist}(x^k, \Lambda) = 0$. Since $\Lambda \subseteq \mathcal{X} \subset U_0$ (where U_0 is defined in Assumption 4.2) and Λ is compact, there exist a bounded open set U_1 and an $N_2 \in \mathbb{N}$ such that $x^k \in U_1$ for all $k \geq N_2$ and the closure of U_1 is contained in U_0.[3]

Next, let N_0 be defined as in Theorem 4.2. Since P_2 is continuously differentiable on U_0 and $x^k \in U_1 \subset U_0$ for any $k \geq N_1 := \max\{N_0, N_2\}$, we obtain from [35, Theorem 8.6] that for any $k \geq N_1$,

$$\begin{aligned}
&\partial H(x^{k+1}, x^k, y^k) \supseteq \widehat{\partial} H(x^{k+1}, x^k, y^k)\\
&\overset{(a)}{\supseteq} \begin{bmatrix} \frac{1}{\hat{\theta}}\widehat{\partial} P(x^{k+1}) + \mathcal{N}_C(x^{k+1}) + L_g(x^{k+1} - x^k) + L_g(x^{k+1} - y^k) \\ -L_g(x^{k+1} - x^k) \\ -L_g(x^{k+1} - y^k) \end{bmatrix} + \widehat{\partial}\Xi(x^{k+1}, x^k, y^k)\\
&\overset{(b)}{=} \begin{bmatrix} \frac{1}{\hat{\theta}}\partial P(x^{k+1}) + \mathcal{N}_C(x^{k+1}) + L_g(x^{k+1} - x^k) + L_g(x^{k+1} - y^k) \\ -L_g(x^{k+1} - x^k) \\ -L_g(x^{k+1} - y^k) \end{bmatrix} + \partial\Xi(x^{k+1}, x^k, y^k)\\
&\overset{(c)}{\supseteq} \begin{bmatrix} \frac{1}{\hat{\theta}}\partial P(x^{k+1}) + \sum\limits_{i \in I_k(x^{k+1})} \lambda_i^k \nabla g_i(y^k) + \mathcal{N}_C(x^{k+1}) + L_g(x^{k+1} - x^k) + L_g(x^{k+1} - y^k) \\ -L_g(x^{k+1} - x^k) \\ \sum_{i \in I_k(x^{k+1})} \lambda_i^k \nabla^2 g_i(y^k)(x^{k+1} - y^k) - L_g(x^{k+1} - y^k) \end{bmatrix},
\end{aligned} \tag{4.17}$$

where $\Xi(x, y, z) := \max_{i=1,\ldots,m}[\mathrm{lin}_{g_i}(x, z)]_+$, and $I_k(x^{k+1})$ and λ_i^k are defined as in Lemma 3.1(iii); here, (a) holds because of the subdifferential calculus rules in [35, Proposition 10.5, Corollary 10.9] and the regularity of the normal cone of C in [35, Theorem 6.9], (b) holds because $\partial\Xi = \widehat{\partial}\Xi$ (thanks to [35, Example 7.28]) and $\partial P(x^{k+1}) = \widehat{\partial} P(x^{k+1})$ (thanks to the regularity of P_1 as asserted in [35, Proposition 8.12], the assumption that P_2 is continuously differentiable at $x^{k+1} \in U_0$ and the subdifferential calculus rule [35, Exercise 8.8(c)]), and (c) follows from [35, Proposition 10.5, Exercise 8.31] and the fact that $\sum_{i \in I_k(x^{k+1})} \lambda_i^k = 1$ and $\lambda_i^k \geq 0$ for all $i \in I_k(x^{k+1})$.

On the other hand, according to Theorem 4.2 and the definition of $\hat{\theta}$ in (4.16), we have that $\theta_k \equiv \theta_{N_0} = \hat{\theta}$ for any $k \geq N_0$. Using this together with the property of λ_i^k from Lemma 3.1(iii) and the differentiability assumption on P_2, we obtain that for all $k \geq N_1$,

$$0 \in \partial P_1(x^{k+1}) - \nabla P_2(x^k) + \hat{\theta} \sum_{i \in I_k(x^{k+1})} \lambda_i^k \nabla g_i(y^k) + \hat{\theta} L_g(x^{k+1} - y^k) + \mathcal{N}_C(x^{k+1}).$$

Rearranging terms in the above display, we see that

$$\nabla P_2(x^k) - \hat{\theta} \sum_{i \in I_k(x^{k+1})} \lambda_i^k \nabla g_i(y^k) - \hat{\theta} L_g(x^{k+1} - y^k) \in \partial P_1(x^{k+1}) + \mathcal{N}_C(x^{k+1}). \tag{4.18}$$

[3] The existence of such a U_1 can be argued as follows: since Λ is compact, there exists $\epsilon > 0$ such that $\{x \in \mathbb{R}^n : \mathrm{dist}(x, \Lambda) < \epsilon\} \subseteq U_0$. Then set $U_1 := \{x \in \mathbb{R}^n : \mathrm{dist}(x, \Lambda) < \epsilon/2\}$.

Since P_2 is continuously differentiable in U_0 (and hence at x^k and x^{k+1} when $k \geq N_1$), we obtain for any $k \geq N_1$ that

$$\begin{aligned}
&\frac{1}{\hat{\theta}}\left(-\hat{\theta} L_g(x^k - y^k) + \nabla P_2(x^k) - \nabla P_2(x^{k+1})\right)\\
&= \frac{1}{\hat{\theta}}\left(\hat{\theta} L_g(x^{k+1} - x^k) - \nabla P_2(x^{k+1}) + \hat{\theta} \sum_{i\in I_k(x^{k+1})} \lambda_i^k \nabla g_i(y^k)\right)\\
&\quad + \frac{1}{\hat{\theta}}\left(\nabla P_2(x^k) - \hat{\theta} \sum_{i\in I_k(x^{k+1})} \lambda_i^k \nabla g_i(y^k) - \hat{\theta} L_g(x^{k+1} - y^k)\right)\\
&\overset{(a)}{\in} \frac{1}{\hat{\theta}}\left(\hat{\theta} L_g(x^{k+1} - x^k) - \nabla P_2(x^{k+1}) + \hat{\theta} \sum_{i\in I_k(x^{k+1})} \lambda_i^k \nabla g_i(y^k)\right) + \frac{1}{\hat{\theta}} \partial P_1(x^{k+1}) + \mathcal{N}_C(x^{k+1})\\
&= \frac{1}{\hat{\theta}} \partial P(x^{k+1}) + \sum_{i\in I_k(x^{k+1})} \lambda_i^k \nabla g_i(y^k) + \mathcal{N}_C(x^{k+1}) + L_g(x^{k+1} - x^k) \qquad (4.19)
\end{aligned}$$

where (a) follows from (4.18), and the last equality holds thanks to [35, Exercise 8.8(c)] and the fact that $P = P_1 - P_2$.

Combining (4.17) and (4.19), for any $k \geq N_1$, we have

$$\begin{bmatrix} \frac{1}{\hat{\theta}}\left(-\hat{\theta} L_g(x^k - y^k) + \nabla P_2(x^k) - \nabla P_2(x^{k+1})\right) + L_g(x^{k+1} - y^k) \\ -L_g(x^{k+1} - x^k) \\ \sum_{i\in I_k(x^{k+1})} \lambda_i^k \nabla^2 g_i(y^k)(x^{k+1} - y^k) - L_g(x^{k+1} - y^k) \end{bmatrix} \in \partial H(x^{k+1}, x^k, y^k).$$

Since ∇P_2 is locally Lipschitz continuous on U_0 (and hence Lipschitz continuous on the bounded open set U_1, say, with modulus L_{P_2}), we see for any $k \geq N_1$ that

$$\begin{aligned}
&\operatorname{dist}\left(0, \partial H(x^{k+1}, x^k, y^k)\right)^2\\
&\leq \left\| \frac{1}{\hat{\theta}}\left(-\hat{\theta} L_g(x^k - y^k) + \nabla P_2(x^k) - \nabla P_2(x^{k+1})\right) + L_g(x^{k+1} - y^k) \right\|^2 + \|L_g(x^{k+1} - x^k)\|^2\\
&\quad + \left\| \sum_{i\in I_k(x^{k+1})} \lambda_i^k \nabla^2 g_i(y^k)(x^{k+1} - y^k) - L_g(x^{k+1} - y^k) \right\|^2\\
&\leq 3L_g^2\|x^k - y^k\|^2 + \frac{3}{\hat{\theta}^2} L_{P_2}^2 \|x^{k+1} - x^k\|^2 + 3L_g^2\|x^{k+1} - y^k\|^2 + L_g^2\|x^{k+1} - x^k\|^2\\
&\quad + 2\left\| \sum_{i\in I_k(x^{k+1})} \lambda_i^k \nabla^2 g_i(y^k) \right\|^2 \|x^{k+1} - y^k\|^2 + 2L_g^2\|x^{k+1} - y^k\|^2.
\end{aligned}$$

The desired conclusion now follows immediately from the above display, the definition and the boundedness of $\{y^k\}$ (thanks to Theorem 4.1(i) and (3.3)) and the continuity of $\nabla^2 g_i$ (thanks to Assumption 4.2). □

Now, we present the convergence rate of the $\{x^k\}$ generated by Algorithm 1 under suitable assumptions. The proof is routine and we refer the readers to, for example, the proofs of Theorems 4.2 and 4.3 of [38].

Theorem 4.4 *(Global convergence and convergence rate of Algorithm 1 in nonconvex setting) Consider (1.1). Suppose that Assumptions 4.1 and 4.2 hold, and the H in (4.16) is a KL function. Let $\{(x^k, y^k)\}$ be generated by Algorithm 1 and Ω be the set of accumulation points of $\{(x^{k+1}, x^k, y^k)\}$. Then $\{x^k\}$ converges to a critical point $\bar{x}$ of (1.1). Moreover, if H satisfies the KL property with exponent $\alpha \in [0, 1)$ at every point in Ω, then there exists $\underline{N} \in \mathbb{N}$ such that the following statements hold.*

(i) *If $\alpha = 0$, then $\{x^k\}$ converges finitely, i.e., $x^k \equiv \bar{x}$ for all $k > \underline{N}$.*
(ii) *If $\alpha \in (0, \frac{1}{2}]$, then there exist $a_0 \in (0, 1)$ and $a_1 > 0$ such that*

$$\|x^k - \bar{x}\| \leq a_1 a_0^k \ \ \forall k > \underline{N}.$$

(iii) *If $\alpha \in (\frac{1}{2}, 1)$, then there exists $a_2 > 0$ such that*

$$\|x^k - \bar{x}\| \leq a_2 k^{-\frac{1-\alpha}{2\alpha-1}} \ \ \forall k > \underline{N}.$$

4.2 Convergence analysis in convex setting

We study the convergence properties of Algorithm 1 under the following convex settings.

Assumption 4.3 Suppose that in (1.1), $P_2 = 0$ and $g_1, \ldots, g_m$ are convex.[4]

Assumption 4.4 The Slater condition holds for $C \cap \mathscr{F}$ in (1.1), i.e., there exists $\hat{x} \in C$ with $g_i(\hat{x}) < 0$ for $i = 1, \ldots, m$.

Remark 4.3 If each g_i is convex and Assumption 4.4 holds, then $RCQ(x)$ holds at every $x \in C$, which implies that Assumption 4.1 holds thanks to Remark 4.1(ii).

Now, we present the convergence properties of Algorithm 1 under Assumptions 4.3 and 4.4. Unlike our convergence rate result in Theorem 4.4 which was based on the KL property of the function H in (4.16), our analysis in this section is based on the KL property of the following function:

$$F_\eta(x) := \frac{1}{\eta}(P_1(x) - \hat{m}) + \delta_C(x) + \max_{i=1,\cdots,m} [g_i(x)]_+ , \tag{4.20}$$

where $\eta > 0$ and $\hat{m} := \inf\{P_1(x) : x \in C\} \in \mathbb{R}$. Compared with H, the explicit KL exponent of F_η is generically readily obtainable (from that of $P_1 + \delta_{C \cap \mathscr{F}}$), as we will discuss in Sect. 5.

[4] Under this assumption, we also set $\ell_{g_i} = 0$ for all i in Algorithm 1.

Theorem 4.5 *[Convergence rate of Algorithm 1 in convex setting] Consider (1.1) and suppose that Assumptions 4.3 and 4.4 hold. Let $\{(x^k,\theta_k)\}$ be generated by Algorithm 1. Then the following statements hold.*

(i) *For any $k \geq 1$,*

$$E(x^{k+1},x^k,\theta_{k+1}) \leq E(x^k,x^{k-1},\theta_k) - \frac{(1-\beta_k^2)L_g}{2}\|x^k - x^{k-1}\|^2,$$

where $E(x,y,\theta) := \frac{1}{\theta}\Big(P_1(x) - \hat{m} + \delta_C(x) + \theta \max_{i=1,\cdots,m}[g_i(x)]_+ + \frac{\theta L_g}{2}\|x - y\|^2\Big)$ with $\hat{m}$ defined as in (4.20).

(ii) *Let Ω be the set of accumulation points of $\big\{(x^{k+1},x^k,\theta_k)\big\}$. Then Ω is a nonempty compact set, $\bar{\omega} := \lim_{k\to\infty} E(x^{k+1},x^k,\theta_k)$ exists, and $E \equiv \bar{\omega}$ on Ω.*

(iii) *If the function[5] $F_{\hat{\theta}}$ is a KL function with exponent $\frac{1}{2}$, then $\{x^k\}$ converges to a minimizer x^* of (1.1), and there exist $c_0 > 0$, $s \in (0,1)$ and $k_0 \in \mathbb{N}$ such that*

$$\|x^k - x^*\| \leq c_0 s^k \quad \forall k > k_0.$$

Proof Using the strong convexity of the objective in (3.2) (note that $\xi^k = 0$ as $P_2 = 0$) and the fact that x^{k+1} minimizes this objective over C, we obtain that for any $x \in C$,

$$\begin{aligned}
&P_1(x^{k+1}) + \theta_k \max_{i=1,\cdots,m}\Big[\mathrm{lin}_{g_i}(x^{k+1},y^k)\Big]_+ + \frac{\theta_k L_g}{2}\|x^{k+1} - y^k\|^2 \\
&\leq P_1(x) + \theta_k \max_{i=1,\cdots,m}\Big[\mathrm{lin}_{g_i}(x,y^k)\Big]_+ + \frac{\theta_k L_g}{2}\|x - y^k\|^2 - \frac{\theta_k L_g}{2}\|x - x^{k+1}\|^2.
\end{aligned} \tag{4.21}$$

Now we are ready to prove the three items one by one.

(i): For any $k \geq 1$, we see that

$$\begin{aligned}
&\frac{1}{\theta_{k+1}}\Big(P_1(x^{k+1}) - \hat{m}\Big) + \max_{i=1,\cdots,m}\Big[g_i(x^{k+1})\Big]_+ \overset{\text{(a)}}{\leq} \frac{1}{\theta_k}\Big(P_1(x^{k+1}) - \hat{m}\Big) + \max_{i=1,\cdots,m}\Big[g_i(x^{k+1})\Big]_+ \\
&\overset{\text{(b)}}{\leq} \frac{P_1(x^{k+1}) - \hat{m}}{\theta_k} + \max_{i=1,\cdots,m}\Big[\mathrm{lin}_{g_i}(x^{k+1},y^k) + \frac{L_{g_i}}{2}\|x^{k+1} - y^k\|^2\Big]_+ \\
&\overset{\text{(c)}}{\leq} \frac{P_1(x^{k+1}) - \hat{m}}{\theta_k} + \max_{i=1,\cdots,m}\Big[\mathrm{lin}_{g_i}(x^{k+1},y^k)\Big]_+ + \frac{L_g}{2}\|x^{k+1} - y^k\|^2 \\
&\overset{\text{(d)}}{\leq} \frac{P_1(x^k) - \hat{m}}{\theta_k} + \max_{i=1,\cdots,m}\Big[\mathrm{lin}_{g_i}(x^k,y^k)\Big]_+ + \frac{L_g}{2}\|x^k - y^k\|^2 - \frac{L_g}{2}\|x^{k+1} - x^k\|^2 \\
&\overset{\text{(e)}}{\leq} \frac{P_1(x^k) - \hat{m}}{\theta_k} + \max_{i=1,\cdots,m}\Big[g_i(x^k)\Big]_+ + \frac{L_g}{2}\|x^k - y^k\|^2 - \frac{L_g}{2}\|x^{k+1} - x^k\|^2 \\
&= \frac{1}{\theta_k}\Big(P_1(x^k) - \hat{m} + \theta_k \max_{i=1,\cdots,m}\Big[g_i(x^k)\Big]_+\Big) + \frac{\beta_k^2 L_g}{2}\|x^k - x^{k-1}\|^2 - \frac{L_g}{2}\|x^{k+1} - x^k\|^2
\end{aligned}$$

[5] i.e., the F_η in (4.20) with $\eta = \hat{\theta}$, where $\hat{\theta}$ is given in (4.16).

$$= E(x^k, x^{k-1}, \theta_k) - \frac{(1-\beta_k^2)L_g}{2}\|x^k - x^{k-1}\|^2 - \frac{L_g}{2}\|x^{k+1} - x^k\|^2,$$

where (a) holds thanks to $\theta_k \le \theta_{k+1}$ and $\hat{m} = \inf\{P_1(x) : x \in C\}$, (b) holds because of the Lipschitz continuity of ∇g_i, (c) follows from $L_g = \max\{L_{g_i} : i = 1, \ldots, m\}$, (d) holds upon invoking (4.21) with $x = x^k$ (as $x^k \in C$), (e) follows from the convexity of g_i, and the last equality follows from the definition of $E(x^k, x^{k-1}, \theta_k)$. The desired inequality now follows immediately from the above display and the definition of $E(x^{k+1}, x^k, \theta_{k+1})$.

(ii): Using similar arguments as Lemma 4.1 (but using item (i) in place of Remark 4.2, and noting that Assumption 4.1 holds according to Remark 4.3), one can show that (ii) holds. We omit its proof for brevity.

(iii): Let Λ be the set of accumulation points of $\{x^k\}$ for notational simplicity. From Remark 4.3, Theorem 4.3 and the formula for the subdifferential of $\max_{i=1,\ldots,m}[g_i(\cdot)]_+$ (see [35, Exercise 8.31]), we deduce that

$$\emptyset \ne \Lambda \subseteq \operatorname{Argmin} F_{\hat{\theta}} =: S. \tag{4.22}$$

Now, write $E_\theta(x, y) := E(x, y, \theta)$ for notational simplicity. By the definitions of F_η in (4.20) and $E(x, y, \theta)$ in item (i), we see that $E_{\hat{\theta}}(x, y) = F_{\hat{\theta}}(x) + \frac{L_g}{2}\|x - y\|^2$, where $\hat{\theta}$ is as in (4.16). From Remark 4.3, Theorem 4.2 and item (i), we have that for any $k \ge N_0$, it holds that $\theta_k = \hat{\theta}$ and

$$E_{\hat{\theta}}(x^{k+1}, x^k) \le E_{\hat{\theta}}(x^k, x^{k-1}) - \frac{L_g(1-\bar{\beta}^2)}{2}\|x^k - x^{k-1}\|^2, \tag{4.23}$$

where $\bar{\beta} = \sup_k \beta_k < 1$ (recall that $\ell_g = 0$ under Assumption 4.3).

Let $\widetilde{S} = \{(x^*, x^*) : x^* \in S\}$ and $\widetilde{\Lambda} = \{(x^*, x^*) : x^* \in \Lambda\}$. In view of (4.22), we have $F_{\hat{\theta}}(\bar{x}) = \inf F_{\hat{\theta}}$ for any $\bar{x} \in S$. Using this together with item (ii) and the definition of $E_{\hat{\theta}}$, one can show readily that whenever $\bar{x} \in S$

$$\bar{\omega} = E_{\hat{\theta}}(\bar{x}, \bar{x}) = F_{\hat{\theta}}(\bar{x}) = \inf_x F_{\hat{\theta}}(x) = \inf_{x,y} E_{\hat{\theta}}(x, y). \tag{4.24}$$

Moreover, in view of (4.22) and the definition of $E_{\hat{\theta}}$, we have

$$\emptyset \ne \widetilde{\Lambda} \subseteq \widetilde{S} = \operatorname*{Argmin}_{x,y} E_{\hat{\theta}}(x, y). \tag{4.25}$$

Furthermore, since $F_{\hat{\theta}}$ is a KL function with exponent $\frac{1}{2}$, we conclude from [23, Theorem 3.6] that $E_{\hat{\theta}}$ is a KL function with exponent $\frac{1}{2}$. Using this together with (4.24), (4.25) and Lemma 2.1, we deduce that there exist $\epsilon_0 > 0$, $r_0 > 0$, and $c_0 > 0$ such that

$$\operatorname{dist}((x, y), \widetilde{S})^2 \le c_0(E_{\hat{\theta}}(x, y) - \bar{\omega}), \tag{4.26}$$

for any $(x, y) \in \operatorname{dom} \partial E_{\hat{\theta}}$ satisfying $\operatorname{dist}((x, y), \widetilde{S}) \le \epsilon_0$ and $\bar{\omega} \le E_{\hat{\theta}}(x, y) < \bar{\omega} + r_0$.

Next, notice that $\{(x^k, x^{k-1})\} \subseteq C \times C \subset \operatorname{dom} \partial E_{\hat\theta} = C \times \mathbb{R}^n$, and we have from Theorem 4.1(iii) that $\widetilde\Lambda$ is the set of accumulation points of the bounded sequence $\{(x^k, x^{k-1})\}$. Using this and (4.25), we deduce that there exists $k_1 \in \mathbb{N}$ such that

$$\operatorname{dist}((x^k, x^{k-1}), \widetilde S) \le \operatorname{dist}((x^k, x^{k-1}), \widetilde\Lambda) \le \epsilon_0 \quad \forall k \ge k_1. \tag{4.27}$$

On the other hand, from Remark 4.3, Theorem 4.2 and item (ii), we deduce the existence of $k_2 \in \mathbb{N}$ such that

$$\bar\omega \le E_{\hat\theta}(x^k, x^{k-1}) < \bar\omega + r_0 \quad \forall k \ge k_2. \tag{4.28}$$

Combining (4.26), (4.27) and (4.28), we conclude that for any $k \ge k_3 := \max\{k_1, k_2\}$,

$$\operatorname{dist}(x^k, S)^2 \le \operatorname{dist}((x^k, x^{k-1}), \widetilde S)^2 \le c_0(E_{\hat\theta}(x^k, x^{k-1}) - \bar\omega). \tag{4.29}$$

Next, let $\bar x^k \in S$ satisfy $\|x^k - \bar x^k\| = \operatorname{dist}(x^k, S)$. Then for any $k \ge N_0$ (note that N_0 is defined in Theorem 4.2) and $\gamma \in (\frac{L_g c_0}{1+L_g c_0}, 1)$, we have

$$\begin{aligned}
F_{\hat\theta}(x^{k+1}) &= \frac{1}{\hat\theta}\left(P_1(x^{k+1}) - \hat m + \hat\theta \max_{i=1,\cdots,m}\left[g_i(x^{k+1})\right]_+\right)\\
&\overset{\text{(a)}}{\le} \frac{1}{\hat\theta}\left(P_1(x^{k+1}) - \hat m + \hat\theta \max_{i=1,\cdots,m}\left[\operatorname{lin}_{g_i}(x^{k+1}, y^k) + \frac{L_{g_i}}{2}\|x^{k+1} - y^k\|^2\right]_+\right)\\
&\overset{\text{(b)}}{\le} \frac{1}{\hat\theta}\left(P_1(x^{k+1}) - \hat m + \hat\theta \max_{i=1,\cdots,m}\left[\operatorname{lin}_{g_i}(x^{k+1}, y^k)\right]_+ + \frac{\hat\theta L_g}{2}\|x^{k+1} - y^k\|^2\right)\\
&\overset{\text{(c)}}{\le} \frac{1}{\hat\theta}\left(P_1(\bar x^k) - \hat m\right) + \max_{i=1,\cdots,m}\left[\operatorname{lin}_{g_i}(\bar x^k, y^k)\right]_+ + \frac{L_g}{2}\|\bar x^k - y^k\|^2 - \frac{L_g}{2}\|\bar x^k - x^{k+1}\|^2\\
&\overset{\text{(d)}}{\le} F_{\hat\theta}(\bar x^k) + \frac{L_g}{2}\|\bar x^k - y^k\|^2 - \frac{L_g}{2}\|\bar x^k - x^{k+1}\|^2,\\
&\overset{\text{(e)}}{\le} F_{\hat\theta}(\bar x^k) + \frac{L_g}{2}\left(\|\bar x^k - x^k\| + \|x^k - y^k\|\right)^2 - \frac{L_g}{2}\|\bar x^k - x^{k+1}\|^2,\\
&\overset{\text{(f)}}{\le} F_{\hat\theta}(\bar x^k) + \frac{L_g}{2\gamma}\|\bar x^k - x^k\|^2 + \frac{L_g}{2(1-\gamma)}\|x^k - y^k\|^2 - \frac{L_g}{2}\|\bar x^k - x^{k+1}\|^2\\
&\overset{\text{(g)}}{\le} \bar\omega + \frac{L_g}{2\gamma}\operatorname{dist}(x^k, S)^2 + \frac{L_g}{2(1-\gamma)}\|x^k - y^k\|^2 - \frac{L_g}{2}\operatorname{dist}(x^{k+1}, S)^2,
\end{aligned} \tag{4.30}$$

where (a) holds because of the Lipschitz continuity of ∇g_i, (b) holds because $L_g = \max_{i=1,\cdots,m}\{L_{g_i}\}$, (c) follows from (4.21) with $x = \bar x^k$ (thanks to $\bar x^k \in S \subseteq C$ and the fact that $\theta_k = \hat\theta$ for $k \ge N_0$), (d) follows from the convexity of g_i so that $\operatorname{lin}_{g_i}(\bar x^k, y^k) \le g_i(\bar x^k)$ for all i, (e) follows from the triangle inequality, (f) follows from the fact that $(a+b)^2 = (\gamma\frac{a}{\gamma} + (1-\gamma)\frac{b}{(1-\gamma)})^2 \le \frac{a^2}{\gamma} + \frac{b^2}{(1-\gamma)}$ as $\gamma \in (0, 1)$, and (g) holds thanks to (4.24) and the definition of $\bar x^k$.

Then, we have for any $k \ge k_4 := \max\{k_3, N_0\}$ that

$$E_{\hat\theta}(x^{k+1}, x^k) - \bar\omega = F_{\hat\theta}(x^{k+1}) - \bar\omega + \frac{L_g}{2}\|x^{k+1} - x^k\|^2$$

$$
\begin{aligned}
&\overset{\text{(a)}}{\leq} \frac{L_g}{2\gamma}\operatorname{dist}(x^k,S)^2 + \frac{L_g}{2(1-\gamma)}\|x^k-y^k\|^2 - \frac{L_g}{2\gamma}\operatorname{dist}(x^{k+1},S)^2 \\
&\quad + \frac{L_g}{2}\left(\frac{1}{\gamma}-1\right)\operatorname{dist}(x^{k+1},S)^2 + \frac{L_g}{2}\|x^{k+1}-x^k\|^2 \\
&\overset{\text{(b)}}{\leq} \left(\frac{L_g}{2\gamma}\operatorname{dist}(x^k,S)^2 - \frac{L_g}{2}\|x^k-x^{k-1}\|^2\right) - \left(\frac{L_g}{2\gamma}\operatorname{dist}(x^{k+1},S)^2 - \frac{L_g}{2}\|x^{k+1}-x^k\|^2\right) \\
&\quad + \frac{L_g\bar{\beta}^2}{2(1-\gamma)}\|x^k-x^{k-1}\|^2 + \frac{L_g}{2}\|x^k-x^{k-1}\|^2 + \frac{L_g}{2}\left(\frac{1}{\gamma}-1\right)c_0(E_{\hat{\theta}}(x^{k+1},x^k)-\bar{\omega}),
\end{aligned}
$$

where (a) holds because of (4.30), and (b) follows from (4.29), $y^k = x^k + \beta_k(x^k - x^{k-1})$ and $\bar{\beta} = \sup_k \beta_k$.

Now, notice that $\gamma \in (\frac{L_g c_0}{1+L_g c_0}, 1)$ implies $\frac{L_g}{2}(\frac{1}{\gamma}-1)c_0 < \frac{1}{2}$. Letting $\vartheta := 1 - \frac{L_g}{2}(\frac{1}{\gamma}-1)c_0$, then we known that $\vartheta > \frac{1}{2}$. Rearranging terms in the above display inequality, we have that for any $k \geq k_4$,

$$
\begin{aligned}
&\vartheta\left(E_{\hat{\theta}}(x^{k+1},x^k)-\bar{\omega}\right) \\
&\leq \frac{L_g}{2}\left(\frac{1}{\gamma}\operatorname{dist}(x^k,S)^2 - \|x^k-x^{k-1}\|^2\right) - \frac{L_g}{2}\left(\frac{1}{\gamma}\operatorname{dist}(x^{k+1},S)^2 - \|x^{k+1}-x^k\|^2\right) \\
&\quad + \frac{L_g(1-\gamma+\bar{\beta}^2)}{2(1-\gamma)}\|x^k-x^{k-1}\|^2 \\
&\overset{\text{(a)}}{\leq} \frac{L_g}{2}\left(\frac{1}{\gamma}\operatorname{dist}(x^k,S)^2 - \|x^k-x^{k-1}\|^2\right) - \frac{L_g}{2}\left(\frac{1}{\gamma}\operatorname{dist}(x^{k+1},S)^2 - \|x^{k+1}-x^k\|^2\right) \\
&\quad + \frac{L_g(1-\gamma+\bar{\beta}^2)}{2(1-\gamma)}\cdot\frac{2}{L_g(1-\bar{\beta}^2)}\left(E_{\hat{\theta}}(x^k,x^{k-1}) - E_{\hat{\theta}}(x^{k+1},x^k)\right),
\end{aligned}
$$

where (a) follows from (4.23).

Denote $\zeta := \frac{1+\bar{\beta}^2-\gamma}{(1-\gamma)(1-\bar{\beta}^2)} > 1$ and $A_k := \frac{L_g}{2}(\frac{1}{\gamma}\operatorname{dist}(x^k,S)^2 - \|x^k-x^{k-1}\|^2)$. Rearranging terms in the above inequality, we obtain that for any $k \geq k_4$,

$$
(\vartheta+\zeta)\left(E_{\hat{\theta}}(x^{k+1},x^k)-\bar{\omega}\right) \leq A_k - A_{k+1} + \zeta\left(E_{\hat{\theta}}(x^k,x^{k-1})-\bar{\omega}\right).
$$

Dividing $\vartheta+\zeta$ on both sides in the above inequality, we see that for any $k \geq k_4$,

$$
E_{\hat{\theta}}(x^{k+1},x^k)-\bar{\omega} \leq \frac{\zeta}{\vartheta+\zeta}\left(E_{\hat{\theta}}(x^k,x^{k-1})-\bar{\omega}\right) + \frac{1}{\vartheta+\zeta}A_k - \frac{1}{\vartheta+\zeta}A_{k+1}. \tag{4.31}
$$

Since $\vartheta > \frac{1}{2}$ and $\zeta > 1$, we have that for any $k \geq k_4$,

$$
\begin{aligned}
\left|\frac{A_k}{\vartheta+\zeta}\right| &\leq |A_k| \leq \frac{L_g}{2}\left(\frac{1}{\gamma}\operatorname{dist}(x^k,S)^2 + \|x^k-x^{k-1}\|^2\right) \\
&\overset{\text{(a)}}{\leq} \frac{L_g c_0}{2\gamma}\left(E_{\hat{\theta}}(x^k,x^{k-1})-\bar{\omega}\right) + \frac{L_g}{2}\|x^k-x^{k-1}\|^2
\end{aligned}
$$

$$\overset{(b)}{\leq} \frac{L_g c_0}{2\gamma}\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega}\right) + \frac{1}{(1-\bar{\beta}^2)}\left(E_{\hat{\theta}}(x^k, x^{k-1}) - E_{\hat{\theta}}(x^{k+1}, x^k)\right)$$
$$\overset{(c)}{\leq} c_1\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega}\right), \tag{4.32}$$

where (a) holds thanks to (4.29), (b) holds because of (4.23), and (c) follows from $E_{\hat{\theta}}(x^{k+1}, x^k) \geq \bar{\omega}$ (see (4.24)) with $c_1 := \frac{L_g c_0}{2\gamma} + \frac{1}{(1-\bar{\beta}^2)}$.

Let $\varrho = \frac{c_1 + \frac{\zeta}{\vartheta+\zeta}}{c_1+1} \in (0, 1)$. Then one can see that

$$\frac{\zeta}{\vartheta+\zeta} + (1-\varrho)c_1 = \varrho. \tag{4.33}$$

Then, from (4.31), we obtain that for any $k \geq k_4$,

$$E_{\hat{\theta}}(x^{k+1}, x^k) - \bar{\omega} + \frac{1}{\vartheta+\zeta}A_{k+1} \leq \frac{\zeta}{\vartheta+\zeta}\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega}\right) + \frac{1}{\vartheta+\zeta}A_k$$
$$\overset{(a)}{\leq} \frac{\zeta}{\vartheta+\zeta}\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega}\right) + \frac{\varrho}{\vartheta+\zeta}A_k + (1-\varrho)\left|\frac{A_k}{\vartheta+\zeta}\right|$$
$$\overset{(b)}{\leq} \left(\frac{\zeta}{\vartheta+\zeta} + (1-\varrho)c_1\right)\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega}\right) + \frac{\varrho}{\vartheta+\zeta}A_k$$
$$\overset{(c)}{=} \varrho\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega} + \frac{1}{\vartheta+\zeta}A_k\right),$$

where (a) holds as $\varrho \in (0, 1)$, (b) follows from (4.32), and (c) holds because of (4.33).

Inductively, since $\varrho > 0$, we see that for any $k \geq k_4$,

$$E_{\hat{\theta}}(x^{k+1}, x^k) - \bar{\omega} + \frac{1}{\vartheta+\zeta}A_{k+1} \leq \varrho^{k-k_4+1}\left(E_{\hat{\theta}}(x^{k_4}, x^{k_4-1}) - \bar{\omega} + \frac{1}{\vartheta+\zeta}A_{k_4}\right),$$

which means, there exists $M > 0$ such that, for any $k \geq k_4$,

$$0 \leq E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega} \leq M\varrho^k - \frac{1}{\vartheta+\zeta}A_k$$
$$\overset{(a)}{=} M\varrho^k - \frac{L_g}{2(\vartheta+\zeta)}\left(\frac{1}{\gamma}\mathrm{dist}(x^k, S)^2 - \|x^k - x^{k-1}\|^2\right) \leq M\varrho^k + \frac{L_g}{2(\vartheta+\zeta)}\|x^k - x^{k-1}\|^2$$
$$\overset{(b)}{\leq} M\varrho^k + \frac{1}{(\vartheta+\zeta)(1-\bar{\beta}^2)}\left(E_{\hat{\theta}}(x^k, x^{k-1}) - E_{\hat{\theta}}(x^{k+1}, x^k)\right), \tag{4.34}$$

where (a) follows from the definition of A_k, and (b) holds because of (4.23).

Taking $\mu > \max\{\frac{1}{(\vartheta+\zeta)(1-\bar{\beta}^2)}, \frac{1}{1-\varrho}\}$. From $\varrho \in (0, 1)$, we see that

$$\mu > 1 \text{ and } 1 - \mu^{-1} > \varrho, \tag{4.35}$$

and from (4.34) (and (4.23), which asserts the nonnegativity of the difference $E_{\hat{\theta}}(x^k, x^{k-1}) - E_{\hat{\theta}}(x^{k+1}, x^k)$), we have that

$$E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega} \le M\varrho^k + \mu(E_{\hat{\theta}}(x^k, x^{k-1}) - E_{\hat{\theta}}(x^{k+1}, x^k)),$$

which implies

$$\mu(E_{\hat{\theta}}(x^{k+1}, x^k) - \bar{\omega}) \le (\mu - 1)\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega}\right) + M\varrho^k.$$

Dividing $\mu > 0$ on the both sides in the above display, we see that for any $k \ge k_4$,

$$\begin{aligned}
&E_{\hat{\theta}}(x^{k+1}, x^k) - \bar{\omega} \le \left(1 - \mu^{-1}\right)\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega}\right) + \frac{M}{\mu}\varrho^k \\
&= \left(1 - \mu^{-1}\right)\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega}\right) + \frac{M}{\mu}\left(\frac{1 - \mu^{-1}}{1 - \mu^{-1} - \varrho} - \frac{\varrho}{1 - \mu^{-1} - \varrho}\right)\varrho^k \\
&= \left(1 - \mu^{-1}\right)\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega} + \frac{M}{\mu(1 - \mu^{-1} - \varrho)}\varrho^k\right) - \frac{M}{\mu(1 - \mu^{-1} - \varrho)}\varrho^{k+1},
\end{aligned}$$

where the division by $1 - \mu^{-1} - \varrho$ is valid thanks to (4.35). Rearranging terms in the above display inequality, we have that for any $k \ge k_4$,

$$\begin{aligned}
&E_{\hat{\theta}}(x^{k+1}, x^k) - \bar{\omega} + \frac{M\varrho^{k+1}}{\mu(1 - \mu^{-1} - \varrho)} \\
&\quad \le \left(1 - \mu^{-1}\right)\left(E_{\hat{\theta}}(x^k, x^{k-1}) - \bar{\omega} + \frac{M\varrho^k}{\mu(1 - \mu^{-1} - \varrho)}\right).
\end{aligned}$$

Inductively, since $1 - \mu^{-1} > 0$ thanks to (4.35), we see that for any $k \ge k_4$,

$$\begin{aligned}
&E_{\hat{\theta}}(x^{k+1}, x^k) - \bar{\omega} \le E_{\hat{\theta}}(x^{k+1}, x^k) - \bar{\omega} + \frac{M\varrho^{k+1}}{\mu(1 - \mu^{-1} - \varrho)} \\
&\le \left(1 - \mu^{-1}\right)^{k - k_4 + 1}\left(E_{\hat{\theta}}(x^{k_4}, x^{k_4 - 1}) - \bar{\omega} + \frac{M\varrho^{k_4}}{\mu(1 - \mu^{-1} - \varrho)}\right) \overset{\text{(a)}}{=} c_2\left(1 - \mu^{-1}\right)^{k+1},
\end{aligned} \tag{4.36}$$

where (a) holds with $c_2 := \left(1 - \mu^{-1}\right)^{-k_4}\left(E_{\hat{\theta}}(x^{k_4}, x^{k_4 - 1}) - \bar{\omega} + \frac{M}{\mu(1 - \mu^{-1} - \varrho)}\varrho^{k_4}\right) > 0$.

Finally, we obtain that for any $k \ge k_4$,

$$\begin{aligned}
\|x^{k+1} - x^k\|^2 &\overset{\text{(a)}}{\le} \frac{2}{L_g(1 - \bar{\beta}^2)}\left(E_{\hat{\theta}}(x^{k+1}, x^k) - E_{\hat{\theta}}(x^{k+2}, x^{k+1})\right) \\
&\overset{\text{(b)}}{\le} \frac{2}{L_g(1 - \bar{\beta}^2)}\left(E_{\hat{\theta}}(x^{k+1}, x^k) - \bar{\omega}\right) \overset{\text{(c)}}{\le} \frac{2c_2}{L_g(1 - \bar{\beta}^2)}\left(1 - \mu^{-1}\right)^{k+1},
\end{aligned}$$

where (a) holds because of (4.23), (b) follows from $E_{\hat{\theta}}(x^{k+1}, x^k) \geq \bar{\omega}$ (see (4.24)), and (c) holds because of (4.36). Consequently, for any $j \geq k \geq k_4$,

$$\sum_{i=k}^{j} \|x^{i+1} - x^i\| \leq \sum_{i=k}^{\infty} \sqrt{\frac{2c_2}{L_g(1-\bar{\beta}^2)}} \left(\sqrt{1-\mu^{-1}}\right)^{i+1} = c_3 \left(\sqrt{1-\mu^{-1}}\right)^{k+1}, \tag{4.37}$$

with $c_3 := \sqrt{\frac{2c_2}{L_g(1-\bar{\beta}^2)}} \cdot \frac{1}{1-\sqrt{1-\mu^{-1}}} > 0$, which implies that $\{x^k\}$ is a Cauchy sequence. Combining this with Remark 4.3 and Theorem 4.3, we see that $\{x^k\}$ converges to a minimizer x^* of (1.1). The claimed linear rate of convergence also follows immediately from (4.37). □

5 KL exponent and exact penalty

In Sect. 4.2, the KL exponent of the $F_{\hat{\theta}}$ in Theorem 4.5(iii) was used for establishing the convergence rate of the $\{x^k\}$ generated by Algorithm 1 in the convex setting. In this section, we examine how the KL exponent of functions of the form (4.20) can be deduced from the corresponding problem (1.1).

Specifically, we consider the following optimization problem:

$$\min_{x \in \mathbb{R}^n} \hat{F}(x) := P_1(x) + \delta_C(x) + \delta_{\mathcal{F}}(x), \tag{5.1}$$

where $P_1 : \mathbb{R}^n \to \mathbb{R}$ is convex, C is compact and convex, $\mathcal{F} := \{x \in \mathbb{R}^n : g_i(x) \leq 0, i = 1, \ldots, m\}$ with each $g_i : \mathbb{R}^n \to \mathbb{R}$ being convex, and $C \cap \{x \in \mathbb{R}^n : \max_{i=1,\ldots,m}\{g_i(x)\} < 0\} \neq \emptyset$; we also consider the associated penalty function

$$\hat{F}_\eta(x) := P_1(x) + \delta_C(x) + \eta \max_{i=1,\ldots,m} [g_i(x)]_+, \tag{5.2}$$

where $\eta > 0$. Notice that for (1.1), the KL property of the corresponding $\hat{F}_{\hat{\theta}}$ was the key for establishing the convergence rate of the $\{x^k\}$ generated by Algorithm 1; see Theorem 4.5(iii).

We next recall the definition of exact penalty parameter.

Definition 5.1 [Exact penalty parameter] Consider (5.1) and (5.2). If there exists $\bar{\eta} > 0$ such that for all $\eta \geq \bar{\eta}$,

$$\operatorname*{Argmin}_{x \in \mathbb{R}^n} \hat{F}_\eta(x) = \operatorname*{Argmin}_{x \in \mathbb{R}^n} \hat{F}(x),$$

then $\bar{\eta}$ is called an exact penalty parameter of (5.1).

We will argue that the set of exact penalty parameters of (5.1) is nonvoid. We start by recalling the following well-known result, whose short proof is included for the convenience of the readers.

Lemma 5.1 *Let C and $\mathcal{F}$ be as in (5.1). Then there exist $\kappa > 0$ and $\tau > 0$ such that*

$$\operatorname{dist}(x, C \cap \mathcal{F}) \leq \kappa \operatorname{dist}(x, \mathcal{F}) \leq \tau \max_{i=1,\ldots,m} [g_i(x)]_+ \quad \forall x \in C. \tag{5.3}$$

Proof First, since $C \cap \{x \in \mathbb{R}^n : \max_{i=1,\ldots,m}\{g_i(x)\} < 0\} \neq \emptyset$ (say, it contains $\hat{x}$), we deduce from [6, Corollary 3] that there exists $\kappa > 0$ such that

$$\operatorname{dist}(x, C \cap \mathcal{F}) \leq \kappa \operatorname{dist}(x, \mathcal{F}) \quad \forall x \in C. \tag{5.4}$$

We then apply Lemma 2.2 with $\Omega := \mathcal{F}$, $h(x) = (g_1(x), g_2(x), \ldots, g_m(x))$, $x^s = \hat{x}$ and $\delta_0 = \left|\max_{i=1,\ldots,m}\{g_i(\hat{x})\}\right|$ to obtain

$$\operatorname{dist}(x, \mathcal{F}) \leq \frac{\|x - \hat{x}\|}{\left|\max_{i=1,\ldots,m}\{g_i(\hat{x})\}\right|} \operatorname{dist}(0, g(x) + \mathbb{R}^m_+) \quad \forall x \in \mathbb{R}^n.$$

Since C is compact, we deduce further that there exists $M_1 > 0$ such that

$$\operatorname{dist}(x, \mathcal{F}) \leq M_1 \max_{i=1,\cdots,m} [g_i(x)]_+ \;\; \forall x \in C. \tag{5.5}$$

The desired conclusion now follows upon combining (5.4) and (5.5). □

Remark 5.1 [Nonemptiness of the set of exact penalty parameters] Consider (5.1) and (5.2). Since $C \cap \mathcal{F}$ is compact and P_1 is continuous, we see that $\operatorname{Argmin} \hat{F} \neq \emptyset$. Using this together with (5.3), we can now deduce from [17, Lemma 3.1] that any $\eta > \bar{L}_{P_1}\tau$ is an exact penalty parameter of (5.1), where $\bar{L}_{P_1}$ is a Lipschitz continuity modulus for P_1 on the compact convex set C.

Now, we show that if the $\hat{F}$ in (5.1) is a KL function with exponent $\alpha \in (0, 1)$, then for any $\eta > \bar{\eta}$, the $\hat{F}_\eta$ in (5.2) is a KL function with the same exponent, where $\bar{\eta}$ is an exact penalty parameter of (5.1).

Theorem 5.1 *[KL exponent of $\hat{F}_\eta$ from that of $\hat{F}$] Let $\hat{F}$ be as in (5.1), $\bar{x} \in \operatorname{Argmin} \hat{F}$ and $\bar{\eta}$ be an exact penalty parameter of (5.1). If $\hat{F}$ satisfies the KL property with exponent $\alpha \in (0, 1)$ at $\bar{x}$, then for any $\eta > \bar{\eta}$, the $\hat{F}_\eta$ defined in (5.2) satisfies the KL property with exponent α at $\bar{x}$.*

Proof Fix any $\eta > \bar{\eta}$. Since $\bar{\eta}$ is an exact penalty parameter of (5.1), we see that $\operatorname{Argmin} \hat{F} = \operatorname{Argmin} \hat{F}_\eta$; also, note that $\operatorname{dom} \partial \hat{F}_\eta = C$ and $\operatorname{dom} \partial \hat{F} = C \cap \mathcal{F}$.

Since $\hat{F}$ satisfies the KL property with exponent α at $\bar{x}$, in view of [11, Theorem 5], there exist $c > 0$ and $a, \epsilon \in (0, 1)$ such that

$$\operatorname{dist}(x, \operatorname{Argmin} \hat{F}) \leq c(\hat{F}(x) - \hat{F}(\bar{x}))^{1-\alpha}, \tag{5.6}$$

whenever $x \in \operatorname{dom} \partial \hat{F} = C \cap \mathcal{F}$ satisfies $\|x - \bar{x}\| \leq \epsilon$ and $\hat{F}(\bar{x}) \leq \hat{F}(x) \leq \hat{F}(\bar{x}) + a$. Since $\hat{F}$ is continuous on its domain, by shrinking ϵ further if necessary, we assume that (5.6) holds whenever $x \in \operatorname{dom} \partial \hat{F} = C \cap \mathcal{F}$ satisfies $\|x - \bar{x}\| \leq \epsilon$.

Next, since $P_1 : \mathbb{R}^n \to \mathbb{R}$ is convex, we know that P_1 is locally Lipschitz continuous at $\bar{x}$. Hence, there exist $\bar{\epsilon} > 0$ and $\hat{L}_{P_1} > 0$ such that

$$|P_1(x) - P_1(y)| \le \hat{L}_{P_1}\|x - y\| \quad \forall x, y \in \{u \in \mathbb{R}^n : \|u - \bar{x}\| \le \bar{\epsilon}\}. \tag{5.7}$$

Now, take $\epsilon_0 := \min\{\epsilon, \bar{\epsilon}\}$. Then for any $x \in C = \operatorname{dom} \partial \hat{F}_\eta$ satisfying $\|x - \bar{x}\| \le \epsilon_0$, we have upon letting $\Pi_{C\cap\mathcal{F}}(x)$ denote the orthogonal projection of x onto $C \cap \mathcal{F}$ that

$$\begin{aligned}
&\operatorname{dist}(x, \operatorname{Argmin} \hat{F}_\eta) \le \operatorname{dist}(\Pi_{C\cap\mathcal{F}}(x), \operatorname{Argmin} \hat{F}_\eta) + \operatorname{dist}(x, C \cap \mathcal{F})\\
&\overset{\text{(a)}}{=} \operatorname{dist}(\Pi_{C\cap\mathcal{F}}(x), \operatorname{Argmin} \hat{F}) + \operatorname{dist}(x, C \cap \mathcal{F})\\
&\overset{\text{(b)}}{\le} c(\hat{F}(\Pi_{C\cap\mathcal{F}}(x)) - \hat{F}(\bar{x}))^{1-\alpha} + \kappa \operatorname{dist}(x, \mathcal{F}) = c(P_1(\Pi_{C\cap\mathcal{F}}(x)) - P_1(\bar{x}))^{1-\alpha} + \kappa \operatorname{dist}(x, \mathcal{F})\\
&\overset{\text{(c)}}{\le} c(P_1(x) - P_1(\bar{x}) + \hat{L}_{P_1} \operatorname{dist}(x, C \cap \mathcal{F}))^{1-\alpha} + \kappa \operatorname{dist}(x, \mathcal{F})\\
&\overset{\text{(d)}}{\le} c(P_1(x) - P_1(\bar{x}) + \hat{L}_{P_1} \kappa \operatorname{dist}(x, \mathcal{F}))^{1-\alpha} + \kappa \operatorname{dist}(x, \mathcal{F})^{1-\alpha}\\
&\overset{\text{(e)}}{\le} \hat{c}\big[(P_1(x) - P_1(\bar{x}) + \hat{L}_{P_1} \kappa \operatorname{dist}(x, \mathcal{F}))^{1-\alpha} + [\kappa^{1/(1-\alpha)} \operatorname{dist}(x, \mathcal{F})]^{1-\alpha}\big]\\
&\overset{\text{(f)}}{\le} \bar{c}(P_1(x) - P_1(\bar{x}) + \kappa_1 \operatorname{dist}(x, \mathcal{F}))^{1-\alpha} \overset{\text{(g)}}{\le} \bar{c}\big(P_1(x) - P_1(\bar{x}) + \kappa_2 \max_{i=1,\cdots,m} [g_i(x)]_+\big)^{1-\alpha},
\end{aligned} \tag{5.8}$$

where (a) holds because $\operatorname{Argmin} \hat{F} = \operatorname{Argmin} \hat{F}_\eta$, (b) holds because of (5.3), (5.6) and the fact that $\|\Pi_{C\cap\mathcal{F}}(x) - \bar{x}\| \le \|x - \bar{x}\| \le \epsilon_0 \le \epsilon$ (thanks to $\bar{x} \in C \cap \mathcal{F}$ and the projection mapping being nonexpansive), (c) follows from (5.7) and the fact that $\epsilon_0 \le \bar{\epsilon}$, (d) follows from (5.3) and the facts that $\operatorname{dist}(x, \mathcal{F}) \le \|x - \bar{x}\| \le \epsilon_0 \le \epsilon < 1$ and $\alpha \in (0, 1)$, (e) holds with $\hat{c} = \max\{c, 1\}$, (f) holds with $\bar{c} = 2^\alpha \hat{c}$ and $\kappa_1 = \hat{L}_{P_1}\kappa + \kappa^{\frac{1}{1-\alpha}}$ thanks to the fact that $a^{1-\alpha} + b^{1-\alpha} \le 2^\alpha (a + b)^{1-\alpha}$ for any $a, b \ge 0$ and $\alpha \in (0, 1)$, and (g) holds with $\kappa_2 = \kappa_1 \tau/\kappa$ thanks to (5.3).

Now, if $\eta \ge \kappa_2$, then, from (5.8), we have that

$$\operatorname{dist}(x, \operatorname{Argmin} \hat{F}_\eta) \le \bar{c}\left(P_1(x) - P_1(\bar{x}) + \eta \max_{i=1,\cdots,m} [g_i(x)]_+\right)^{1-\alpha} = \bar{c}(\hat{F}_\eta(x) - \hat{F}_\eta(\bar{x}))^{1-\alpha}.$$

On the other hand, if $\kappa_2 > \eta > \bar{\eta}$, then, from (5.8), we obtain that

$$\begin{aligned}
&\operatorname{dist}(x, \operatorname{Argmin} \hat{F}_\eta) \le \bar{c}\left(P_1(x) - P_1(\bar{x}) + \bar{\eta} \max_{i=1,\cdots,m} [g_i(x)]_+ + (\kappa_2 - \bar{\eta}) \max_{i=1,\cdots,m} [g_i(x)]_+\right)^{1-\alpha}\\
&\overset{\text{(a)}}{\le} \bar{c}\left(\frac{\kappa_2 - \bar{\eta}}{\eta - \bar{\eta}}\right)^{1-\alpha}\left(P_1(x) - P_1(\bar{x}) + \bar{\eta} \max_{i=1,\cdots,m} [g_i(x)]_+ + (\eta - \bar{\eta}) \max_{i=1,\cdots,m} [g_i(x)]_+\right)^{1-\alpha}\\
&= \bar{c}\left(\frac{\kappa_2 - \bar{\eta}}{\eta - \bar{\eta}}\right)^{1-\alpha}\left(\hat{F}_\eta(x) - \hat{F}_\eta(\bar{x})\right)^{1-\alpha},
\end{aligned}$$

where (a) holds because $a + b \le \frac{1}{\epsilon}(a + \epsilon b)$ for any $a \ge 0$, $b \ge 0$ and $0 < \epsilon \le 1$.[6] The desired conclusion now follows immediately upon invoking [11, Theorem 5]. $\square$

We next comment on how Theorem 5.1 can be applied to find the KL exponent of $F_{\hat{\theta}}$ in Theorem 4.5(iii); specifically, we will comment on the condition $\eta > \bar{\eta}$ in Theorem 5.1. We first recall the following well-known result concerning exact penalty parameters.

Lemma 5.2 *Consider (5.1) and (5.2). If $\widetilde{\eta} > 0$ is such that* $\operatorname{Argmin} \hat{F}_{\widetilde{\eta}} \cap \operatorname{Argmin}_{x \in C \cap \mathcal{F}} P_1(x) \neq \emptyset$, *then* $\operatorname{Argmin} \hat{F}_\eta = \operatorname{Argmin}_{x \in C \cap \mathcal{F}} P_1(x)$ *whenever* $\eta > \widetilde{\eta}$.

Proof Fix any $\eta > \widetilde{\eta}$ and let $\hat{x} \in \operatorname{Argmin} \hat{F}_{\widetilde{\eta}} \cap \operatorname{Argmin}_{x \in C \cap \mathcal{F}} P_1(x)$. We first argue that $\operatorname{Argmin}_{x \in C \cap \mathcal{F}} P_1(x) = \operatorname{Argmin} \hat{F}_\eta \cap \mathcal{F}$. Indeed, if $\tilde{x} \in \operatorname{Argmin}_{x \in C \cap \mathcal{F}} P_1(x)$, then $\tilde{x} \in C \cap \mathcal{F} \subseteq \mathcal{F}$ and hence $\max_{i=1,\ldots,m}[g_i(\tilde{x})]_+ = 0$. Moreover, it holds that

$$\hat{F}_\eta(\tilde{x}) = P_1(\tilde{x}) \overset{\text{(a)}}{=} P_1(\hat{x}) = \hat{F}_{\widetilde{\eta}}(\hat{x}) \overset{\text{(b)}}{\le} \hat{F}_{\widetilde{\eta}}(x) \overset{\text{(c)}}{\le} \hat{F}_\eta(x)$$

for any $x \in C$, where (a) holds because both $\hat{x}$ and $\tilde{x}$ minimize P_1 over $C \cap \mathcal{F}$, (b) holds because $\hat{x}$ also minimizes $\hat{F}_{\widetilde{\eta}}$, and (c) holds because $\eta > \widetilde{\eta}$. As for the converse inclusion, let $\tilde{x} \in \operatorname{Argmin} \hat{F}_\eta \cap \mathcal{F}$. Then for any $x \in C \cap \mathcal{F}$, we have

$$P_1(\tilde{x}) = \hat{F}_\eta(\tilde{x}) \le \hat{F}_\eta(x) = P_1(x),$$

where the equalities hold because $u \in \mathcal{F}$ implies $\max_{i=1,\ldots,m}[g_i(u)]_+ = 0$. The above arguments establish $\operatorname{Argmin}_{x \in C \cap \mathcal{F}} P_1(x) = \operatorname{Argmin} \hat{F}_\eta \cap \mathcal{F}$.

To complete the proof, it now suffices to show that $\operatorname{Argmin} \hat{F}_\eta \subseteq \mathcal{F}$. To this end, let $\tilde{x} \in \operatorname{Argmin} \hat{F}_\eta$. Then we have

$$\begin{aligned} P_1(\tilde{x}) + \eta \max_{i=1,\ldots,m}[g_i(\tilde{x})]_+ &= \hat{F}_\eta(\tilde{x}) \le \hat{F}_\eta(\hat{x}) = P_1(\hat{x}) + \eta \max_{i=1,\ldots,m}[g_i(\hat{x})]_+ \\ &\overset{\text{(a)}}{=} P_1(\hat{x}) + \widetilde{\eta} \max_{i=1,\ldots,m}[g_i(\hat{x})]_+ = \hat{F}_{\widetilde{\eta}}(\hat{x}) \overset{\text{(b)}}{\le} \hat{F}_{\widetilde{\eta}}(\tilde{x}) = P_1(\tilde{x}) + \widetilde{\eta} \max_{i=1,\ldots,m}[g_i(\tilde{x})]_+, \end{aligned}$$

where (a) holds because $\hat{x} \in C \cap \mathcal{F}$ (hence $\max_{i=1,\ldots,m}[g_i(\hat{x})]_+ = 0$) and (b) holds because $\hat{x}$ minimizes $\hat{F}_{\widetilde{\eta}}$. Rearranging terms in the above inequality, we obtain $(\eta - \widetilde{\eta}) \max_{i=1,\ldots,m}[g_i(\tilde{x})]_+ = 0$, which means $\tilde{x} \in \mathcal{F}$. $\square$

Remark 5.2 [On the condition $\eta > \bar{\eta}$ in Theorem 5.1] We comment on the applicability of Theorem 5.1, which only infers the KL exponent of $\hat{F}_\eta$ when $\eta > \bar{\eta}$ for some exact penalty parameter $\bar{\eta}$.

Particularly, we consider (1.1). Suppose that Assumptions 4.3 and 4.4 hold and let $\{(x^k, \theta_k)\}$ be generated by Algorithm 1. Using Theorem 4.3 (see also Remark 4.3) and

[6] We apply this relation to $\epsilon := (\eta - \bar{\eta})/(\kappa_2 - \bar{\eta}) \in (0, 1)$, $b := (\kappa_2 - \bar{\eta}) \max_{i=1,\cdots,m}[g_i(x)]_+ \ge 0$, and $a := P_1(x) - P_1(\bar{x}) + \bar{\eta} \max_{i=1,\cdots,m}[g_i(x)]_+$, which is nonnegative because $\bar{\eta}$ is an exact penalty parameter, $\bar{x} \in \operatorname{Argmin} \hat{F} = \operatorname{Argmin} \hat{F}_{\bar{\eta}}$ and $x \in C$.

the formula for the subdifferential of $\max_{i=1,\ldots,m}[g_i(\cdot)]_+$ (see [35, Exercise 8.31]), we deduce from the definition of F_η in (4.20) that

$$\emptyset \neq \Lambda \subseteq \operatorname{Argmin} F_{\hat{\theta}} \cap \operatorname*{Argmin}_{x\in C\cap\mathscr{F}} P_1(x). \tag{5.9}$$

where $C \cap \mathscr{F}$ is the feasible set of (1.1) and Λ is the set of accumulation points of $\{x^k\}$. Combining (5.9) with Lemma 5.2, we deduce that the set of exact penalty parameters is nonempty; indeed, it contains the interval $(\hat{\theta}, \infty)$. Hence, if we let $\tilde{\eta}$ denote the infimum of the set of exact penalty parameters, then $\hat{\theta} \geq \tilde{\eta}$.

Now, note that we have $\theta_k \equiv \hat{\theta}$ whenever $k \geq N_0$ (where N_0 is defined in Theorem 4.2) and θ_k is nondecreasing. Intuitively, it is likely that the update rule of θ_k will result in $\hat{\theta} > \tilde{\eta}$. In this case, Theorem 5.1 asserts that the KL property required in Theorem 4.5(iii) can be inferred from that of $P_1 + \delta_C + \delta_{\mathscr{F}}$ in (1.1). On the other hand, in the case $\hat{\theta} = \tilde{\eta}$, Theorem 5.1 is not applicable for connecting the KL property of $F_{\hat{\theta}}$ to that of $P_1 + \delta_C + \delta_{\mathscr{F}}$.

Example 5.1 Suppose that in (5.1), $P_1 = \|\cdot\|_1$, C is a polytope containing the origin, $m = 1$, and $g_1 = q_1 \circ A_1$ for some matrix $A_1 \in \mathbb{R}^{s_1\times n}$ and $q_1 : \mathbb{R}^{s_1} \to \mathbb{R}$ taking one of the following forms with $b \in \mathbb{R}^{s_1}$ and $\sigma > 0$ chosen so that the origin is not feasible and that $\inf_{x\in C} g_1(x) < 0$:

(i) (Basis pursuit denoising [15]) $q_1(z) = \frac{1}{2}\|z - b\|^2 - \sigma$.
(ii) (Logistic loss [22]) $q_1(z) = \sum_{i=1}^{s_1} \log(1+\exp(b_i z_i)) - \sigma$ for some $b \in \{-1, 1\}^{s_1}$.

Let $\bar{\eta}$ be the exact penalty parameter of (5.1). We deduce from [41, Section 5.1] and Theorem 5.1 that, for any $\eta > \bar{\eta}$, the KL exponent of the corresponding $\hat{F}_\eta$ in (5.2) (and hence the corresponding F_η in (4.20)) is $\frac{1}{2}$.

6 Numerical experiments

In this section, we perform numerical experiments to illustrate the performance of Algorithm 1. Particularly, motivated by the use of the difference of ℓ_1 and ℓ_2 norms (ℓ_{1-2}), which was first introduced in [19] and further studied in [26, 39] for sparse signal recovery, we consider the following model for compressed sensing:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|_1 - \mu\|x\| \\ \text{s.t.} \quad & h(Ax - b) \leq \sigma, \end{aligned} \tag{6.1}$$

where $\mu \in [0, 1)$, $A \in \mathbb{R}^{q\times n}$ has full row rank, $b \in \mathbb{R}^q$, $h : \mathbb{R}^q \to \mathbb{R}_+$ is an analytic function whose gradient is Lipschitz continuous with modulus L_h and satisfies $h(0) = 0$, and $\sigma \in (0, h(-b))$.[7]

[7] On passing, we would like to point out that projecting onto the constraint set of (6.1) is in general not easy, even though it only involves a single constraint function. For example, when $h(\cdot) = \frac{1}{2}\|\cdot\|^2$, the projection problem becomes a generalized trust region subproblem, for which state-of-the-art solvers would require

Although the feasible region of (6.1) is unbounded and Algorithm 1 cannot be directly applied to solving (6.1), one can argue as in the discussion following [41, Eq. (6.2)] that (6.1) is equivalent to the following model:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|_1 - \mu\|x\| \\ \text{s.t.} \quad & h(Ax-b) \le \sigma, \\ & \|x\|_\infty \le M, \end{aligned} \tag{6.2}$$

where $M := (1-\mu)^{-1}\Big(\|A^\dagger b\|_1 - \mu\|A^\dagger b\|\Big)$.[8] Notice that the equivalent problem (6.2) is a special case of (1.1) with $P_1(x) = \|x\|_1$, $P_2(x) = \mu\|x\|$, $m = 1$, $g_1(x) = h(Ax-b)-\sigma$ and $C = \{x : \|x\|_\infty \le M\}$; since A has full row rank and $h(0) = 0 < \sigma$, we see that $A^\dagger b \in C \cap \{x : g_1(x) < 0\} \neq \emptyset$.

Next, we will focus on (6.2) and consider two specific choices of h. All numerical experiments are performed in MATLAB R2022a on a 64-bit PC with an Intel(R) Core(TM) i7-10710U CPU (@1.10GHz, 1.61GHz) and 16GB of RAM.

6.1 $h(\cdot) = \frac{1}{2}\|\cdot\|^2$

In this subsection, we take $h(\cdot) = \frac{1}{2}\|\cdot\|^2$, then (6.2) becomes

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|_1 - \mu\|x\| \\ \text{s.t.} \quad & 0.5\cdot\|Ax-b\|^2 \le \sigma, \\ & \|x\|_\infty \le M. \end{aligned} \tag{6.3}$$

Notice that h is convex, the Slater condition holds for the feasible region of (6.3), and the origin is not feasible as $\sigma \in (0, \frac{1}{2}\|b\|^2)$. These together with Remark 4.3 imply that Assumptions 4.1 and 4.2 hold. Since the H in (4.16) corresponding to (6.3) is clearly semi-algebraic and hence a KL function, one can then apply Theorem 4.4 with $\ell_g = 0$ to deduce the convergence of the (whole) sequence $\{x^k\}$ generated by Algorithm 1 with $\sup_k \beta_k < 1$ for solving (6.3).[9]

We compare $\mathrm{SCP_{ls}}$ in [40], $\mathrm{ESQM_e}$ (Algorithm 1 with $\{\beta_k\}$ specified below) and $\mathrm{ESQM_b}$ (this is a basic version of ESQM obtained by setting $\beta_k \equiv 0$ in Algorithm 1). We use the same parameter settings for $\mathrm{SCP_{ls}}$ in [40], and the initial point of $\mathrm{SCP_{ls}}$ is

---

computing the matrix vector products $A^T u$ and Av (possibly multiple times); see [1, 24, 27]. As another example, when h is the Lorentzian norm [16], the constraint set in (6.1) is *nonconvex* and it is unclear how the projection onto this set can be computed efficiently.

8 We also recall from [41, Section 6.1] that $\|A^\dagger b\|_\infty \le M$ by the construction of M.

9 Though we are not considering $\mu = 0$ in our experiments below, we also point out that when $\mu = 0$ in (6.3), thanks to $\sigma \in (0, \frac{1}{2}\|b\|^2)$ and the fact that $A^\dagger b \in C \cap \{x : g_1(x) < 0\}$, we can deduce from Example 5.1 that $x \to \|x\|_1 + \delta_C(x) + \eta[g_1(x)]_+$ is a KL function with exponent $\frac{1}{2}$ whenever η exceeds some exact penalty parameter. Therefore, according to Remark 5.2, for the sequence $\{(x^k, \theta_k)\}$ generated by Algorithm 1, if $\hat{\theta}$ exceeds some exact penalty parameter, then $\{x^k\}$ converges locally linearly thanks to Theorem 4.5(iii).

chosen as $x^0 = A^\dagger b$. For ESQM$_\mathrm{b}$ and ESQM$_\mathrm{e}$, we take $L_g = \|A\|^2$, $\ell_g = 0$, $d = 1$ and $\theta_0 = 1$, and their initial points are chosen as $x^0 = 0$. We terminate all algorithms when

$$\|x^{k+1} - x^k\| < \epsilon \cdot \max\{1, \|x^{k+1}\|\} \tag{6.4}$$

for some $\epsilon > 0$ specified below. The subproblems in these algorithms are solved according to the procedures described in the appendices of [40] and [41].

We use the same strategy of choosing β_k as in the FISTA with fixed and adaptive restart described in [18]. In more detail, we set the initial values $\vartheta_{-1} = \vartheta_0 = 1$ and define, for $k \geq 0$,

$$\beta_k = \frac{\vartheta_{k-1} - 1}{\vartheta_k} \quad \text{with} \quad \vartheta_{k+1} = \frac{1 + \sqrt{1 + 4\vartheta_k^2}}{2}, \tag{6.5}$$

and we reset $\vartheta_{k-1} = \vartheta_k = 1$ every $K = 200$ iterations or when $\langle y^{k-1} - x^k, x^k - x^{k-1}\rangle > 0$. One can show that $\{\beta_k\}$ generated this way satisfies $\{\beta_k\} \subseteq [0, 1)$ and $\sup_k \beta_k < 1$.

We perform tests on random instances of (6.3). Specifically, we generate an $A \in \mathbb{R}^{q\times n}$ with independent and identically distributed (i.i.d.) standard Gaussian entries, and then normalize this matrix so that each column of it has unit norm. Then we choose a subset T of size k uniformly at random from $\{1, 2, \cdots, n\}$ and a k-sparse vector x_{orig} having i.i.d. standard Gaussian entries on T is generated. We let $b = Ax_{\mathrm{orig}} + 0.01 \cdot \hat{n}$ with $\hat{n}$ being a random vector having i.i.d. standard Gaussian entries, and $\sigma = 0.5\sigma_1^2$ with $\sigma_1 = 1.1 \cdot \|0.01 \cdot \hat{n}\|$.

In our numerical tests, we let $\mu = 0.95$ in (6.3) and $(q, n, k) = (720i, 2560i, 160i)$ with $i \in \{2, 4, 6, 8, 10\}$. For each i, we generate 20 random instances as described above. We present the computational results when ϵ in (6.4) equals 10^{-4} and 10^{-6} in Tables 1 and 2, respectively, averaged over the 20 random instances. Here, we present the time for computing the QR decomposition of A^T (denoted by t_{QR}), the time for computing $\|A\|^2$ (denoted by $t_{\|A\|}$),[10] the time for computing $x^0 = A^\dagger b$ given the QR factorization of A^T (denoted by $t_{A^\dagger b}$),[11] the CPU times of the algorithms,[12] the number of iterations (denoted by Iter), the recovery errors RecErr $:= \frac{\|x^* - x_{\mathrm{orig}}\|}{\max\{1, \|x_{\mathrm{orig}}\|\}}$ and the residuals Residual $:= \frac{\|Ax^* - b\|^2 - \sigma_1^2}{\sigma_1^2}$, where x^* is the approximate solution returned by the respective algorithm.

From Tables 1 and 2, one can see that ESQM$_\mathrm{e}$ is the fastest algorithm, and the recovery errors of all three methods are comparable.

[10] The $\|A\|^2$ is computed via the Matlab code norm(A*A′) when $p \leq 2000$, and is computed using eigs(A*A′,1,′LM′) otherwise.

[11] Note that $A^\dagger b$ is used by SCP$_{\mathrm{ls}}$ as the initial point and for computing the M in (6.2) for ESQM$_\mathrm{b}$ and ESQM$_\mathrm{e}$, while $\|A\|$ is only used by ESQM$_\mathrm{b}$ and ESQM$_\mathrm{e}$.

[12] The CPU times do not include t_{QR}, $t_{\|A\|}$ and $t_{A^\dagger b}$.

Table 1 Computational results for problem (6.3) with $\epsilon = 10^{-4}$.

	Method	$i=2$	$i=4$	$i=6$	$i=8$	$i=10$
CPU time (sec)	t_{QR}	0.615	3.713	13.600	32.645	66.354
	$t_{A^\dagger b}$	0.006	0.024	0.059	0.113	0.176
	$t_{\|A\|}$	0.532	1.438	4.754	11.120	21.776
	$\mathrm{SCP}_{\mathrm{ls}}$	2.332	8.159	18.235	32.125	48.558
	$\mathrm{ESQM}_{\mathrm{b}}$	8.157	34.875	84.801	143.836	234.291
	$\mathrm{ESQM}_{\mathrm{e}}$	0.559	2.230	5.401	9.111	14.713
Iter	$\mathrm{SCP}_{\mathrm{ls}}$	208	213	211	212	212
	$\mathrm{ESQM}_{\mathrm{b}}$	1729	1781	1819	1768	1789
	$\mathrm{ESQM}_{\mathrm{e}}$	108	112	114	112	113
RecErr	$\mathrm{SCP}_{\mathrm{ls}}$	0.053	0.053	0.054	0.055	0.055
	$\mathrm{ESQM}_{\mathrm{b}}$	0.070	0.071	0.073	0.073	0.074
	$\mathrm{ESQM}_{\mathrm{e}}$	0.051	0.051	0.052	0.053	0.053
Residual	$\mathrm{SCP}_{\mathrm{ls}}$	−1.61e−05	−2.01e−05	−2.07e−05	−2.06e−05	−2.03e−05
	$\mathrm{ESQM}_{\mathrm{b}}$	6.36e−07	6.04e−07	5.42e−07	5.60e−07	5.38e−07
	$\mathrm{ESQM}_{\mathrm{e}}$	1.20e−07	1.11e−07	9.96e−08	9.71e−08	1.03e−07

Table 2 Computational results for problem (6.3) with $\epsilon = 10^{-6}$

	Method	$i = 2$	$i = 4$	$i = 6$	$i = 8$	$i = 10$
CPU time (sec)	t_{QR}	0.662	4.444	13.801	31.694	59.477
	$t_{A^{\dagger}b}$	0.007	0.029	0.060	0.104	0.160
	$t_{\|A\|}$	0.576	1.633	4.858	10.567	20.454
	$\mathrm{SCP_{ls}}$	2.919	10.359	22.412	38.074	58.432
	$\mathrm{ESQM_b}$	12.849	57.570	137.098	232.529	368.612
	$\mathrm{ESQM_e}$	0.936	4.470	10.606	18.551	29.806
Iter	$\mathrm{SCP_{ls}}$	251	257	257	258	259
	$\mathrm{ESQM_b}$	2756	2860	2954	2887	2924
	$\mathrm{ESQM_e}$	195	220	228	230	237
RecErr	$\mathrm{SCP_{ls}}$	0.051	0.051	0.052	0.053	0.053
	$\mathrm{ESQM_b}$	0.051	0.051	0.052	0.053	0.053
	$\mathrm{ESQM_e}$	0.051	0.051	0.052	0.052	0.053
Residual	$\mathrm{SCP_{ls}}$	−1.61e−09	−1.86e−09	−1.85e−09	−1.67e−09	−1.84e−09
	$\mathrm{ESQM_b}$	9.09e−11	9.04e−11	8.71e−11	8.74e−11	8.85e−11
	$\mathrm{ESQM_e}$	5.66e−11	1.00e−10	-4.51e−14	4.10e−11	−4.93e−13

6.2 When h is the Lorentzian norm

In this subsection, we consider h being the Lorentzian norm [16], which is defined as follows for any given $\gamma > 0$:

$$\|y\|_{LL_{2,\gamma}} := \sum_{i=1}^{q} \log\left(1 + \frac{y_i^2}{\gamma^2}\right).$$

Then, problem (6.2) becomes the following problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \|x\|_1 - \mu\|x\| \\ \text{s.t.} \quad & \|Ax - b\|_{LL_{2,\gamma}} \le \sigma, \\ & \|x\|_\infty \le M. \end{aligned} \tag{6.6}$$

We first argue that Assumption 4.1 holds for (6.6) under our assumptions on A and σ in (6.2). To this end, let $\hat{h}(y) := \|y\|_{LL_{2,\gamma}} - \sigma$ for notational simplicity. Then (6.6) is an instance of (1.1) with $g_1(x) = \hat{h}(Ax - b) - \sigma$ and $C := \{x : \|x\|_\infty \le M\}$. Next, recall that $A^\dagger b \in C$ by the construction of M. Moreover, observe that for any $x \in C$, we have

$$\begin{aligned} &\langle \nabla g_1(x), A^\dagger b - x\rangle = \langle A^T \nabla \hat{h}(Ax - b), A^\dagger b - x\rangle = \langle \nabla \hat{h}(Ax - b), b - Ax\rangle \\ &= 2\sum_{i=1}^{q} \frac{a_i^T x - b_i}{\gamma^2 + (a_i^T x - b_i)^2} \cdot (b_i - a_i^T x) = -2\sum_{i=1}^{q} \frac{(a_i^T x - b_i)^2}{\gamma^2 + (a_i^T x - b_i)^2}, \end{aligned} \tag{6.7}$$

where a_i^T is the i-th row of A. Now we consider two cases:

- $x \in C \setminus \mathscr{F}$. In this case, suppose that there exist u_i, $i \in I(x)$, such that (4.5) holds. Then, in particular, we must have $I(x)$ (defined in Assumption 4.1) being nonempty, which in turn means $I(x) = \{1\}$. In addition, (4.5) together with (6.7) implies that $a_i^T x = b_i$ for all i. But then $g_1(x) = \hat{h}(0) - \sigma = -\sigma < 0$, contradicting the fact that $I(x) = \{1\}$.
- $x \in C \cap \mathscr{F}$. In this case, we claim that $g_1(x) + \langle \nabla g_1(x), A^\dagger b - x\rangle < 0$. Suppose to the contrary that $g_1(x) + \langle \nabla g_1(x), A^\dagger b - x\rangle = 0$. Then using $x \in \mathscr{F}$ and (6.7), we deduce that $g_1(x) = \langle \nabla g_1(x), A^\dagger b - x\rangle = 0$. The second equality together with (6.7) implies that $a_i^T x = b_i$ for all i. But then we deduce from this and $g_1(x) = 0$ that

  $$0 = g_1(x) = \hat{h}(0) - \sigma = -\sigma < 0,$$

  which is a contradiction. Thus, we have shown that $RCQ(x)$ holds and one can actually choose $y = A^\dagger b$ there.

Consequently, Assumption 4.1 holds.

Next, observe that the $\hat{h}$ has Lipschitz continuous gradient with modulus $\frac{2}{\gamma^2}$. The following proposition shows that $\hat{h}$ can be represented as the difference of two convex functions $\hat{h}_1$ and $\hat{h}_2$ with Lipschitz continuous gradients, and the Lipschitz continuity modulus of $\nabla \hat{h}_1$ is $\frac{2}{\gamma^2}$ while that of $\nabla \hat{h}_2$ is $\frac{1}{4\gamma^2}$.

Proposition 6.1 *Let $\hat{h}(y) := \|y\|_{LL_{2,\gamma}} - \sigma$. Then there exist two convex functions $\hat{h}_1$ and $\hat{h}_2$ with Lipschitz continuous gradients such that $\hat{h}(y) = \hat{h}_1(y) - \hat{h}_2(y)$ and the Lipschitz continuity modulus of $\nabla \hat{h}_1$ is $\frac{2}{\gamma^2}$ while that of $\nabla \hat{h}_2$ is $\frac{1}{4\gamma^2}$.*

Proof First, notice that

$$\frac{d^2}{dt^2}\log(1+t^2) = \frac{2(1-t^2)}{(1+t^2)^2} = \left[\frac{2(1-t^2)}{(1+t^2)^2}\right]_+ - \left[\frac{2(1-t^2)}{(1+t^2)^2}\right]_- ,$$

where $s_+ := \max\{s, 0\} \geq 0$ and $s_- := -\min\{s, 0\} \geq 0$ for any $s \in \mathbb{R}$. Now, define, for each $t \in \mathbb{R}$,

$$r_1(t) = \int_0^t (t-s)\left[\frac{2(1-s^2)}{(1+s^2)^2}\right]_+ ds \ \text{ and } \ r_2(t) = \int_0^t (t-s)\left[\frac{2(1-s^2)}{(1+s^2)^2}\right]_- ds.$$

Then $r_1''(t) = \left[\frac{2(1-t^2)}{(1+t^2)^2}\right]_+$ and $r_2''(t) = \left[\frac{2(1-t^2)}{(1+t^2)^2}\right]_-$, showing that r_1 and r_2 are convex. Moreover, one can observe that $r_1(0) = r_2(0) = r_1'(0) = r_2'(0) = 0$, and a direct computation shows that $\log(1+t^2) = r_1(t) - r_2(t)$, $\sup_t |r_1''(t)| = 2$ and $\sup_t |r_2''(t)| = \frac{1}{4}$. Taking

$$\hat{h}_1(y) = \sum_{i=1}^m r_1(y_i/\gamma) - \sigma, \ \text{ and } \ \hat{h}_2(y) = \sum_{i=1}^m r_2(y_i/\gamma),$$

one can see that $\hat{h}_1$ and $\hat{h}_2$ are two convex functions with Lipschitz continuous gradients, and $\hat{h}(y) = \hat{h}_1(y) - \hat{h}_2(y)$. Furthermore, the Lipschitz continuity modulus of $\nabla \hat{h}_1$ and $\nabla \hat{h}_2$ are $\frac{2}{\gamma^2}$ and $\frac{1}{4\gamma^2}$, respectively. □

Recall that the origin is not feasible for (6.6) under our assumptions on A and σ in (6.2). In view of this, the above discussions, and the observation that the H in (4.16) corresponding to (6.3) is a subanalytic function that is continuous on its closed domain (and hence a KL function in view of [9, Theorem 3.1]), one can apply Theorem 4.4 with $L_g = \frac{2\|A\|^2}{\gamma^2}$ and $\ell_g = \frac{\|A\|^2}{4\gamma^2}$ to deduce the convergence of the $\{x^k\}$ generated by Algorithm 1 with $\sup_k \beta_k < \sqrt{\frac{L_g}{L_g+\ell_g}} = \sqrt{\frac{8}{9}}$ for solving (6.6).

As in the previous subsection, we compare $\mathrm{SCP_{ls}}$, $\mathrm{ESQM_b}$ and $\mathrm{ESQM_e}$. For $\mathrm{SCP_{ls}}$, we use the same parameter settings in [40], and initialize it at $x^0 = A^\dagger b$. For $\mathrm{ESQM_b}$ and $\mathrm{ESQM_e}$, we take $L_g = \frac{2\|A\|^2}{\gamma^2}$, $\ell_g = \frac{\|A\|^2}{4\gamma^2}$, $d = \frac{\gamma^2}{150\|A\|^2}$ and $\theta_0 = 1.1\gamma$, and they are initialized at $x^0 = 0$. We terminate all algorithms when (6.4) holds for some

$\epsilon > 0$ specified below. Furthermore, the subproblems in these algorithms are solved according to the procedures described in the appendices of [40] and [41].

We also choose $\{\beta_k\}$ as described in (6.5) but we set the fixed restart frequency as $K = 48$. This parameter will ensure that $\{\beta_k\}$ satisfies $\{\beta_k\} \subseteq \left[0, \sqrt{\frac{L_g}{L_g+\ell_g}}\right)$ and $\sup_k \beta_k < \sqrt{\frac{L_g}{L_g+\ell_g}}$.

We perform tests on random instances of (6.3). As in the previous section, we generate an $A \in \mathbb{R}^{q\times n}$ with i.i.d. standard Gaussian entries, and then normalize its columns. We then choose a subset T of size k uniformly at random from $\{1, 2, \cdots, n\}$ and generate a k-sparse vector $x_{\rm orig}$ with i.i.d. standard Gaussian entries on T. We let $b = Ax_{\rm orig} + 0.01 \cdot \bar{n}$ with $\bar{n}_i \sim \mathrm{Cauchy}(0, 1)$, specifically, we generate $\bar{n}_i$ as $\tan(\pi(\tilde{n}_i - \frac{1}{2}))$ with $\tilde{n}$ being a random vector with i.i.d. entries uniformly chosen in $[0, 1]$. We then set $\sigma = 1.05 \cdot \|0.01 \cdot \bar{n}\|_{LL_{2,\gamma}}$ with $\gamma = 0.08$.

In our numerical tests, we let $\mu = 0.95$ in (6.6) and $(q, n, k) = (720i, 2560i, 80i)$ with $i \in \{2, 4, 6, 8, 10\}$. For each i, we generate 20 random instances as described above. The computational results for ϵ in (6.4) being 10^{-4} and 10^{-6} are respectively presented in Tables 3 and 4, averaged over the 20 random instances. As before, we present the time for computing the QR decomposition of A^T (denoted by $t_{\rm QR}$), the time for computing $\|A\|^2$ (denoted by $t_{\|A\|}$), the time for computing $x^0 = A^\dagger b$ given the QR factorization of A^T (denoted by $t_{A^\dagger b}$), the CPU times of the algorithms,[13] the number of iterations (denoted by Iter), the recovery errors RecErr $:= \frac{\|x^*-x_{\rm orig}\|}{\max\{1,\|x_{\rm orig}\|\}}$ and the residuals Residual $:= \frac{\|Ax^*-b\|_{LL_{2,\gamma}}-\sigma}{\sigma}$, where x^* is the approximate solution returned by the respective algorithm.

From Tables 3 and 4, we observe a similar pattern as shown in Tables 1 and 2, i.e., $\mathrm{ESQM_e}$ is the fastest algorithm, and the recovery errors of all three methods are comparable.

7 Concluding remarks

In this paper, we developed a variant of the extended sequential quadratic method (ESQM) in [4] for (1.1), which we call ESQM with extrapolation ($\mathrm{ESQM_e}$), that incorporates Nesterov's extrapolation techniques for empirical acceleration. We established subsequential convergence and global convergence of the whole sequence under suitable assumptions. Our numerical experiments indicated that the extrapolation techniques are empirically effective in accelerating the convergence.

We conclude with several interesting future research directions. First of all, throughout the paper, we assumed that the set C in (1.1) is compact. While such an assumption is convenient for guaranteeing the boundedness of the sequence $\{x^k\}$ generated by our algorithm, it is conceivable that one may replace it with weaker assumptions such as the coercivity of $P + \delta_C$. In addition, it is also interesting to consider the case when $C = \mathbb{R}^n$ in (1.1) and P_1 is a proper closed convex function with dom P_1 being a *proper*

[13] The CPU times do not include $t_{\rm QR}$, $t_{\|A\|}$ and $t_{A^\dagger b}$.

Table 3 Computational results for problem (6.6) with $\varepsilon = 10^{-4}$

	Method	$i = 2$	$i = 4$	$i = 6$	$i = 8$	$i = 10$
CPU time (sec)	t_{QR}	0.577	4.143	11.775	27.407	50.278
	$t_{A^\dagger b}$	0.005	0.026	0.049	0.088	0.140
	$t_{\\|A\\|}$	0.467	1.548	4.432	9.593	17.722
	$\mathrm{SCP}_{\mathrm{ls}}$	1.186	6.915	8.429	45.353	29.767
	$\mathrm{ESQM}_{\mathrm{b}}$	2.587	11.985	26.931	47.601	75.835
	$\mathrm{ESQM}_{\mathrm{e}}$	0.561	2.557	5.635	9.984	15.804
Iter	$\mathrm{SCP}_{\mathrm{ls}}$	120	195	110	354	153
	$\mathrm{ESQM}_{\mathrm{b}}$	586	609	607	608	613
	$\mathrm{ESQM}_{\mathrm{e}}$	120	126	125	126	127
RecErr	$\mathrm{SCP}_{\mathrm{ls}}$	0.092	0.090	0.092	0.092	0.092
	$\mathrm{ESQM}_{\mathrm{b}}$	0.096	0.093	0.095	0.095	0.096
	$\mathrm{ESQM}_{\mathrm{e}}$	0.092	0.089	0.091	0.091	0.092
Residual	$\mathrm{SCP}_{\mathrm{ls}}$	−1.91e−07	−2.26e−07	−2.31e−07	−2.68e−07	−2.74e−07
	$\mathrm{ESQM}_{\mathrm{b}}$	8.81e−08	8.86e−08	8.58e−08	8.61e−08	8.51e−08
	$\mathrm{ESQM}_{\mathrm{e}}$	1.02e−08	1.23e−08	1.46e−08	5.75e−09	6.40e−09

Table 4 Computational results for problem (6.6) with $\varepsilon = 10^{-6}$

	Method	$i = 2$	$i = 4$	$i = 6$	$i = 8$	$i = 10$
CPU time (sec)	t_{QR}	0.558	4.093	12.902	31.823	61.492
	$t_{A^\dagger b}$	0.006	0.029	0.059	0.112	0.180
	$t_{\|A\|}$	0.466	1.546	4.660	10.864	21.334
	$\mathrm{SCP_{ls}}$	1.338	8.106	9.814	48.802	36.869
	$\mathrm{ESQM_b}$	4.006	19.608	41.110	72.591	120.954
	$\mathrm{ESQM_e}$	0.766	3.765	7.715	13.936	23.430
Iter	$\mathrm{SCP_{ls}}$	136	214	127	372	171
	$\mathrm{ESQM_b}$	882	914	914	914	919
	$\mathrm{ESQM_e}$	164	169	168	173	174
RecErr	$\mathrm{SCP_{ls}}$	0.092	0.089	0.091	0.091	0.092
	$\mathrm{ESQM_b}$	0.092	0.089	0.091	0.091	0.092
	$\mathrm{ESQM_e}$	0.092	0.089	0.091	0.091	0.092
Residual	$\mathrm{SCP_{ls}}$	−2.31e−11	−2.37e−11	−3.19e−11	−2.94e−11	−1.98e−11
	$\mathrm{ESQM_b}$	8.62e−12	8.68e−12	8.39e−12	8.44e−12	8.29e−12
	$\mathrm{ESQM_e}$	2.23e−12	3.59e−12	3.89e−12	7.46e−13	1.19e−12

subset of $\mathbb{R}^n$: this will necessitate the development of a variant of Assumption 4.1. These are several avenues for future research.

Acknowledgements The first author is supported in part by the National Natural Science Foundation of China (11901414) and (11871359). The second author is supported in part by the Hong Kong Research Grants Council PolyU153001/22p. The authors would also like to thank the referees for their comments, which helped improve the paper.

Funding Open access funding provided by The Hong Kong Polytechnic University

Data Availability The codes for generating the random data and implementing the algorithms in the numerical section are available from the second author upon request.

Declarations

Conflict of interest The second author is an editorial board member of this journal.

References

1. Adachi, S., Iwata, S., Nakatsukasa, Y., Takeda, A.: Solving the trust-region subproblem by a generalized eigenvalue problem. SIAM J. Optim. **27**, 269–291 (2017)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**, 438–457 (2010)
3. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math. Program. **137**, 91–129 (2013)
4. Auslender, A.: An extended sequential quadratically constrained quadratic programming algorithm for nonlinear, semidefinite, and second-order cone programming. J. Optim. Theory Appl. **156**, 183–212 (2013)
5. Auslender, A., Shefi, B., Teboulle, M.: A moving balls approximation method for a class of smooth constrained minimization problems. SIAM J. Optim. **20**, 3232–3259 (2010)
6. Bauschke, H.H., Borwein, J.M., Li, W.: Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization. Math. Program. **86**, 135–160 (1999)
7. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Process. **18**, 2419–2434 (2009)
8. Becker, S.R., Candès, E.J., Grant, M.C.: Templates for convex cone problems with applications to sparse signal recovery. Math. Program. Comput. **3**, 165–218 (2011)
9. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**, 1205–1223 (2007)
10. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**, 556–572 (2007)
11. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. **165**, 471–507 (2017)

12. Bolte, J., Pauwels, E.: Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. Math. Oper. Res. **41**, 442–465 (2016)
13. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2014)
14. Brezinski, C.: Convergence acceleration during the 20th century. J. Comput. Appl. Math. **122**, 1–21 (2000)
15. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. C.R. Math. **346**, 589–592 (2008)
16. Carrillo, R.E., Barner, K.E., Aysal, T.C.: Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise. IEEE J. Sel. Top. Signal Process. **4**, 392–408 (2010)
17. Chen, X., Lu, Z., Pong, T.K.: Penalty methods for a class of non-Lipschitz optimization problems. SIAM J. Optim. **26**, 1465–1492 (2016)
18. O'Donoghue, B., Candès, E.J.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**, 715–732 (2015)
19. Esser, E., Lou, Y., Xin, J.: A method for finding structured sparse solutions to non-negative least squares problems with applications. SIAM J. Imag. Sci. **6**, 2010–2046 (2013)
20. Gill, P.E., Wong, E.: Sequential quadratic programming methods. In: Lee, J., Leyffer, S. (eds.) Mixed Integer Nonlinear Programming, pp. 147–224. Springer, New York (2012)
21. Hadjidimos, A.: Successive overrelaxation (SOR) and related methods. J. Comput. Appl. Math. **123**, 177–199 (2000)
22. Hosmer, J.D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 3rd edn. Wiley, Hoboken (2013)
23. Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. **18**, 1199–1232 (2018)
24. Lieder, F.: Solving large scale cubic regularization by a generalized eigenvalue problem. SIAM J. Optim. **30**, 3345–3358 (2020)
25. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
26. Lou, Y., Yin, P., He, Q., Xin, J.: Computing sparse representation in a highly coherent dictionary based on difference of L1 and L2. J. Sci. Comput. **64**(1), 178–196 (2015)
27. Pong, T.K., Wolkowicz, H.: The generalized trust region subproblem. Comput. Optim. Appl. **58**, 273–322 (2014)
28. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence O $(\frac{1}{k^2})$. Doklady AN USSR **269**, 543–547 (1983)
29. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Boston (2004)
30. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**, 127–152 (2005)
31. Nesterov, Y.: Gradient methods for minimizing composite objective function. Math. Program. **140**, 125–161 (2013)
32. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
33. Robinson, S.M.: An application of error bounds for convex programming in a linear space. SIAM J. Control **13**, 271–273 (1975)
34. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
35. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1997)
36. Smith, D.A., Ford, W.F., Sidi, A.: Extrapolation methods for vector sequences. SIAM Rev. **29**, 199–233 (1987)
37. Wen, B., Chen, X., Pong, T.K.: Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. SIAM J. Optim. **27**, 124–145 (2017)
38. Wen, B., Chen, X., Pong, T.K.: A proximal difference-of-convex algorithm with extrapolation. Comput. Optim. Appl. **69**, 297–324 (2018)
39. Yin, P., Lou, Y., He, Q., Xin, J.: Minimization of ℓ_{1-2} for compressed sensing. SIAM J. Sci. Comput. **37**, A536–A563 (2015)
40. Yu, P., Pong, T.K., Lu, Z.: Convergence rate analysis of a sequential convex programming method with line search for a class of constrained difference-of-convex optimization problems. SIAM J. Optim. **31**, 2024–2054 (2021)

41. Zhang, Y., Li, G., Pong, T.K., Xu, S.: Retraction-based first-order feasible methods for difference-of-convex programs with smooth inequality and simple geometric constraints. Adv. Comput. Math. **49**, 8 (2023)