Journal of Global Optimization
https://doi.org/10.1007/s10898-026-01642-5

Projected subgradient methods for paraconvex optimization: Application to robust low-rank matrix recovery

Morteza Rahimi[1] · **Susan Ghaderi**[2] · **Yves Moreau**[2] · **Masoud Ahookhosh**[1]

Received: 23 December 2024 / Accepted: 22 August 2026
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract
This paper is devoted to the class of paraconvex functions and presents some of its fundamental properties, characterization, and examples that can be used for their recognition and optimization. Next, the convergence analysis of the projected subgradient methods with several step-sizes (i.e., constant, nonsummable, square-summable but not summable, geometrically decaying, and Scaled Polyak's step-sizes) to global minima for this class of functions is studied. In particular, the convergence rate of the proposed methods is investigated under paraconvexity and the Hölderian error bound condition, where the latter is an extension of the classical error bound condition. The preliminary numerical experiments on several robust low-rank matrix recovery problems (i.e., robust matrix completion, robust matrix compression, and robust image deblurring) indicate promising behavior for these projected subgradient methods, validating our theoretical foundations.

Keywords Nonsmooth nonconvex optimization · Paraconvex functions · Projected subgradient methods · Hölderian error bound · Linear convergenc · Robust low-rank matrix recovery

Mathematics Subject Classification 90C06, 90C25, 90C26, 49J52, 49J53

MR and MA acknowledge the support by the *Research Foundation Flanders (FWO)* research project G081222N and *UA BOF DocPRO4 projects* with ID 46929 and 48996; SG also acknowledges the support of the *Research Foundation Flanders (FWO)* research project 12AK924N.

✉ Morteza Rahimi
morteza.rahimi@uantwerpen.be

Susan Ghaderi
susan.ghaderi@esat.kuleuven.be

Yves Moreau
Yves.Moreau@kuleuven.be

Masoud Ahookhosh
masoud.ahookhosh@uantwerpen.be

[1] Department of Mathematics, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium

[2] Department of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven KU Leuven, Belgium

Published online: 07 September 2026

1 Introduction

Subgradient methods and their projected variants were originally introduced for solving convex nonsmooth problems in the 1960s; see, e.g., [24, 54, 56, 65]. Due to their simple structure and low memory requirement, these methods have been widely employed [47, 48, 50], especially when the objective function does not admit a simple structure for splitting algorithms; cf. [1, 3, 4, 51]. Recent trends in application domains such as signal and image processing, machine learning, data science, statistics, and sparse optimization often involve nonsmooth and nonconvex functions. Representative examples include image deblurring [43], image inpainting [77], compressed sensing [14], sparse recovery [72], matrix factorization [16], nonnegative matrix factorization [27], matrix completion [66], tensor factorization [71], phase retrieval [23], and deep neural networks [31]. This motivates the development of efficient subgradient methods for large-scale nonconvex problems, which is the primary goal of this work.

In the current paper, we are interested in solving the nonsmooth and nonconvex optimization problem

$$\min_{x\in X} f(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, nonsmooth, and nonconvex, and $X \subseteq \mathrm{dom}(f)$ is nonempty, closed, and convex. We assume the set of optimal solutions $\mathcal{X}^*$ is nonempty and $f^* > -\infty$ is the optimal value. Compared to convex problems, nonsmooth and nonconvex problems are challenging due to (i) possible nondifferentiability of f, requiring nonempty Fréchet/Mordukhovich/Clarke subdifferentials [17], (ii) complex landscapes with minima, maxima, and saddle points, and (iii) local minimizers not necessarily being global.

First-order methods are suitable for problem (1.1), as they rely only on function values and (sub)gradients [7, 49, 62]. Projected Subgradient Methods (PSMs), in particular, iterate as

$$\begin{cases} \text{Select } \zeta_k \in \partial f(x_k), \\ x_{k+1} = \mathrm{proj}_X\left(x_k + \alpha_k \frac{\zeta_k}{\|\zeta_k\|}\right), \end{cases}$$

where $\alpha_k > 0$ is a predetermined step-size and proj_X is the projection onto X. This approach avoids function evaluations and has the step-size predefined before its implementation, making it suitable for large-to huge-scale problems. In the nonconvex setting, most studies of subgradient methods have only been carried out for limited classes of functions, such as quasiconvex functions [29, 30, 38, 57] and weakly convex functions [19, 41], along with references therein. This is our primary motivation to explore the convergence analysis of subgradient methods for larger classes of functions.

We here focus on the class of ν-paraconvex functions with $\nu \in (0, 1]$. A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be ν-paraconvex if there exists $\rho \geq 0$ such that

$$h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y) + \rho \min\{\lambda, 1-\lambda\}\|x-y\|^{1+\nu}, \tag{1.2}$$

for each $x, y \in X$ and $\lambda \in [0, 1]$. The class of ν-paraconvex functions generalizes weakly convex functions (1-paraconvex functions) [59, 60, 69], with additional examples not satisfying weak convexity. The class of 1-paraconvex functions arises in applications including optimal control, image processing, and machine learning [13, 19, 41]. In particular, Rolewicz [60] demonstrated that 1-paraconvexity of h is equivalent to the convexity of the perturbed function $x \mapsto h(x)+\rho\|x\|^2$ for some $\rho \geq 0$, while this property does not hold for $\nu \in (0, 1)$, providing a counterexample in [59]. Nevertheless, we establish that if the function $x \mapsto f(x)+\rho\|x\|^{1+\nu}$ is convex for some $\rho \geq 0$, then f qualifies as ν-paraconvex. These functions admit useful

properties such as local Lipschitzness, subdifferentiability, and subdifferential monotonicity [33, 34, 67].

In absence of the strong convexity of the objective function, studying convergence under *error bound* conditions has a long history [9–12, 19, 25, 32, 40, 45, 53, 74, 79, 80]. In particular, the *Hölderian error bound* with order $\delta \in (0, 1]$ provides an upper bound on the distance to the solution set in terms of the residual, with special cases including sharpness error bound ($\delta = 1$), [11, 12, 19, 55], and quadratic growth condition ($\delta = \frac{1}{2}$), [37, 79]. Under these assumptions, PSMs with carefully chosen step-sizes can achieve linear convergence in certain settings [19]. Our work studies convergence rates of PSMs (with several different step-sizes) for ν-paraconvex functions under Hölderian error bounds.

1.1 Contribution

Our primary goal is to extend the applicability of PSMs to a broader class of nonconvex, nonsmooth optimization problems, particularly focusing on paraconvex functions. We aim to provide comprehensive theoretical foundations to investigate convergence rates of the PSMs for such a class of functions and practical insights into this area. Our contribution has threefold, which we describe as follows:

(i) **Characterization of the class of paraconvex functions.** We first describe the class of paraconvex functions as a generalization of the class of weakly convex functions. Then, we verify that while preserving the concept of paraconvexity, the term $\min\{\lambda, 1-\lambda\}$ can be removed from (1.2). Next, it is shown that the continuous midpoint ν-paraconvexity implies ν-paraconvexity, facilitating the assessment of paraconvexity. We particularly demonstrate a connection to a local version of ν-paraconvex functions, known as ν-weakly convex functions (see [18]), coincide with ν-paraconvex functions on compact convex sets. In addition, we establish the local Lipschitzness of such functions, ensuring the nonemptiness of their Clarke subdifferential. As a wider class than weakly convex functions, they accept saddle points. As such, under Hölderian error bound, we identify a region around the optimal set containing no saddle points.

(ii) **Convergence rate of PSMs with several step-sizes.** We describe a generic framework for PSMs for paraconvex functions under the Hölderian error bound condition. In this setting, we study the convergence behavior of this subgradient method for several different step-sizes, such as constant, nonsummable diminishing, square-summable but not summable, geometrically decaying, and Scaled Polyak's step-sizes. We demonstrate recurrences to establish upper bounds for the sequence of iterations (i.e., $\{\text{dist}(x^k, \mathcal{X}^*)\}_{k\in\mathbb{N}}$) and function values (i.e., $\{f(x^k) - f^*\}_{k\in\mathbb{N}}$) residuals. It is shown that: (i) for the PSM with *constant step-size*, $\{\text{dist}(x^k, \mathcal{X}^*)\}_{k\in\mathbb{N}}$ is linearly convergent up to a specified tolerance; (ii) for the PSMs with nonsummable diminishing and square-summable but not summable step-sizes, the subsequential and global convergence are established, respectively; (iii) for the PSMs with geometrically decaying and Scaled Polyak's step-sizes, $\{\text{dist}(x^k, \mathcal{X}^*)\}_{k\in\mathbb{N}}$ is linearly convergent.

(iii) **Applications to low-rank robust matrix recovery.** We introduce a class of robust low-rank matrix recovery appearing in several applications such as robust matrix completion, robust matrix compression, and robust image deblurring. We apply the proposed PSM to these applications, and the results indicate the promising behavior of these algorithms. To the best of our knowledge, this is the first theoretical and numerical appearance of the PSM with Scaled Polyak's step-size in the literature.

1.2 Organization

The remainder of the paper is organized as follows. Section 2 introduces the notations and reviews foundational concepts. In Section 3, we analyze the class of ν-paraconvex functions and explore their fundamental properties. Section 4 focuses on PSMs for ν-paraconvex functions satisfying the Hölderian error bound condition, deriving key inequalities. In Section 4.1, we investigate PSM with constant step-size. Section 4.2 discusses PSMs with diminishing step-size strategies. Section 4.3 examines PSMs using Scaled Polyak's step-size. In Section 5, we present preliminary numerical results on robust low-rank matrix recovery validating our theoretical foundations. Finally, Section 6 delivers our concluding remarks.

2 Preliminaries

2.1 Notation

In this paper, we denote the standard *inner product* and the *Euclidean norm* in n-dimensional real *Euclidean space* $\mathbb{R}^n$ by $\langle \cdot, \cdot \rangle$ and $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$, respectively. The set of *natural numbers* is denoted by $\mathbb{N}$. The notion x^T represents the *transpose* of a vector $x \in \mathbb{R}^n$. For a real $m \times n$ matrix $C = [c_{ij}] \in \mathbb{R}^{m\times n}$ the *Frobenius norm* is $\|C\| = \sqrt{\sum_{i,j} |c_{ij}|^2}$. The open ball with center $x \in \mathbb{R}^n$ and radius $r > 0$ is expressed as $\mathbb{B}(x; r)$. The *interior* and *closure* of a set $S \subseteq \mathbb{R}^n$ are indicated by $\operatorname{int} S$ and $cl\ S$, respectively. The *Euclidean distance* from a point $x \in \mathbb{R}^n$ to a nonempty set $S \subseteq \mathbb{R}^n$ is defined as $\operatorname{dist}(x; S) = \inf_{z\in S} \|z - x\|$. Moreover, the *Euclidean projection* of the point x onto S is given by $\operatorname{proj}_S(x) := \arg\min_{z\in S} \|z - x\|$. For a given function $h : \mathbb{R}^n \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$, the *effective domain* of h is defined as $\operatorname{dom}(h) := \{x \in \mathbb{R}^n : h(x) < +\infty\}$. The function h is said to be *proper* if $\operatorname{dom}(h) \neq \emptyset$. The *indicator* function of a nonempty set $S \subseteq \mathbb{R}^n$, $\delta_S : \mathbb{R}^n \to \overline{\mathbb{R}}$, is defined as $\delta_S(x) = 0$ if $x \in S$, and $\delta_S(x) = +\infty$ otherwise.

Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function that is locally Lipschitz at $\bar{x} \in \operatorname{int}\operatorname{dom}(h)$. The *generalized directional derivative* of h at $\bar{x}$ in the direction d is given by $h^\circ(\bar{x}; d) := \limsup_{\substack{x\to\bar{x} \\ t\downarrow 0}} \frac{h(x+td)-h(x)}{t}$. Moreover, the *Clarke's subdifferential* of h at $\bar{x}$ [17] is defined as $\partial h(\bar{x}) := \{\zeta \in \mathbb{R}^n : \langle \zeta, d\rangle \leq h^\circ(\bar{x}; d), \forall d \in \mathbb{R}^n\}$. Given a nonempty set $S \subseteq \operatorname{int}\operatorname{dom}(h)$, a point $\bar{x} \in S$ is said to be a *stationary* point of the problem $\min_{x\in S} h(x)$ if $0 \in \partial(h + \delta_S)(\bar{x})$.

To derive the convergence rate bounds of our algorithms, we use the following fact from [6].

Fact 2.1 *Let s_k be a nonnegative real sequence and let ς, ϖ be some positive constants. Suppose that $s_k \to 0$ and that the sequence satisfies $s_k^\varsigma \leq \varpi(s_k - s_{k+1})$, for all k sufficiently large. Then, the following assertions hold:*

(a) If $\varsigma = 0$, the sequence s_k converges to 0 in a finite number of steps;
(b) If $\varsigma \in (0, 1]$, the sequence s_k converges linearly to 0 with rate $1 - 1/\varpi$;
(c) If $\varsigma > 1$, there exists $\eta > 0$ such that $s_k \leq \eta k^{-\frac{1}{\varsigma-1}}$, for all $k \in \mathbb{N}$ sufficiently large.

2.2 Hölderian error bound condition

An error bound is a crucial tool in optimization to achieve faster convergence rates in algorithms. In this study, we focus on a specific type of error bound known as the Hölderian error bound.

Definition 2.2 A proper function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to admit a Hölderian error bound (HEB) of order $\delta \in (0, 1]$ with constant $\mu > 0$ on a nonempty set $S \subseteq \mathrm{dom}(h)$ if

$$\mu \, \mathrm{dist}^{1/\delta}(x; \mathcal{X}^*) \leq h(x) - h^*, \quad \forall x \in S, \tag{2.1}$$

where $h^* := \inf_{x \in S} h(x)$ and the set of minimizers $\mathcal{X}^* := \{x \in S : h(x) = h^*\}$ is nonempty.

In the literature, this property is also known by other names, such as the Hölderian growth condition, the Hölder growth bound, the Hölder error bound and the global Hölderian error bound; see, e.g., [28, 32]. The HEB is instrumental in ensuring that PSMs with appropriate step-sizes achieve linear convergence, even for nonconvex problems [19]. Two notable cases of the HEB arise when $\delta = 1$ and $\delta = \frac{1}{2}$. For $\delta = 1$, the condition is known as the sharpness error bound or the weak sharpness condition [12, 55], while for $\delta = \frac{1}{2}$, it is known as the quadratic growth condition [37, 79].

Beyond the HEB (including the sharpness error bound and quadratic growth condition), other prominent properties have been studied, such as the Łojasiewicz inequality [9], the Kurdyka-Łojasiewicz (KL) inequality [10], and (Hölder) metric subregularity/regularity [39]. The Łojasiewicz inequality can be considered a version of the HEB, where the inequality (2.1) holds on each compact subset of an open set S. Furthermore, it has been established that the HEB and the KL inequality are equivalent for convex, closed, and proper functions [9, 10]. On the other hand, Hölder metric subregularity represents a localized version of the HEB. Additionally, [26, Proposition 3.3] showed that the HEB of order δ implies Hölder metric subregularity of ∂f in the convex setting.

3 Fundamental properties of paraconvex functions and optimization

This section introduces the class of ν-paraconvex functions and describes some of their key properties, which are necessary for the coming sections.

3.1 Fundamental properties of paraconvex functions

Let us begin with the definition of ν-paraconvexity.

Definition 3.1 A proper function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be ν-paraconvex function on a convex set $S \subseteq \mathrm{dom}(h)$ for some $0 < \nu \leq 1$ if there exists $\rho > 0$ such that for any $x, y \in S$ and $\lambda \in [0, 1]$,

$$h(\lambda x + (1 - \lambda)y) \leq \lambda h(x) + (1 - \lambda)h(y) + \rho \min\{\lambda, 1 - \lambda\}\|x - y\|^{1+\nu}. \tag{3.1}$$

The following result shows that the notion of ν-paraconvexity is unchanged if the factor $\min\{\lambda, 1 - \lambda\}$ is omitted from the inequality in (3.1). Nevertheless, this factor plays a useful role in (3.1), as it ensures equality at the endpoints $\lambda = 0$ and $\lambda = 1$.

Proposition 3.2 (Equivalent definitions of ν-paraconvexity) *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper function, $S \subseteq \mathrm{dom}(h)$ be a convex set, and $\nu \in (0, 1]$. The following assertions are equivalent:*

(a) There exists a constant $C > 0$ such that for any $x, y \in S$ and $\lambda \in [0, 1]$,

$$h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y) + C\lambda(1-\lambda)\|x-y\|^{1+\nu}; \tag{3.2}$$

(b) The function h is ν-paraconvex on S;

(c) There exists a constant $C > 0$ such that for any $x, y \in S$ and $\lambda \in [0, 1]$,

$$h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y) + C\|x-y\|^{1+\nu}. \tag{3.3}$$

Proof The implication (a) $\Rightarrow$ (b) $\Rightarrow$ (c) follows directly from the definition of ν-paraconvexity and the fact that $\lambda(1-\lambda) \leq \min\{\lambda, 1-\lambda\} \leq 1$. We next show that (c) implies (a). Assume that (c) holds. Fix arbitrary points $x, y \in S$ and $\lambda \in [0, 1]$. Without loss of generality, we suppose that $0 \leq \lambda \leq \frac{1}{2}$. Let us define the function $\varphi : [0, 1] \to \mathbb{R}$ by $\varphi(t) := h(z_t) - th(x) - (1-t)h(y)$ in which $z_t = tx + (1-t)y$. Then, the argument reduces to showing that there exists some $\overline{C} > 0$ such that $\varphi(\lambda) \leq \overline{C}\lambda(1-\lambda)\|x-y\|^{1+\nu}$. By (3.3), $\varphi(0) = 0$ and $\varphi(\frac{1}{2}) \leq C\|x-y\|^{1+\nu} \leq \frac{\overline{C}}{4}\|x-y\|^{1+\nu}$ for some $4C \leq \overline{C}$. Assume that $\lambda \in (0, \frac{1}{2}]$. Then, there exists $s \in \mathbb{N}$ such that $\lambda \in [2^{-s}, 2^{-s+1})$ having the expression $\lambda = \mu 2^{-s} + (1-\mu)2^{-s+1}$ for some $\mu \in (0, 1]$. We claim that, for all $k \in \mathbb{N}$,

$$\varphi(\tfrac{1}{2^k}) \leq \tfrac{2^{1+\nu}}{2^k} C\|x-y\|^{1+\nu} \sum_{i=1}^{k} (\tfrac{1}{2^\nu})^i. \tag{3.4}$$

The above inequality is valid for $k = 1$. By induction, we assume that it holds for $k \in \mathbb{N}$. Using (3.3), for each $t \in [0, 1]$, we get

$$\begin{aligned}
\varphi(\tfrac{t}{2}) &= h\left(\tfrac{z_t + y}{2}\right) - \tfrac{t}{2}h(x) - (1-\tfrac{t}{2})h(y) \\
&\leq \tfrac{1}{2}h(z_t) + \tfrac{1}{2}h(y) + Ct^{1+\nu}\|x-y\|^{1+\nu} - \tfrac{t}{2}h(x) - (1-\tfrac{t}{2})h(y) \\
&= \tfrac{1}{2}\varphi(t) + Ct^{1+\nu}\|x-y\|^{1+\nu}.
\end{aligned}$$

Applying this inequality with $t = 2^{-k}$ and invoking the induction hypothesis yields

$$\begin{aligned}
\varphi(\tfrac{1}{2^{k+1}}) &= \varphi\left(\frac{\frac{1}{2^k}}{2}\right) \leq \tfrac{1}{2}\varphi(\tfrac{1}{2^k}) + \frac{C}{2^{k(1+\nu)}}\|x-y\|^{1+\nu} \\
&\leq \left(\tfrac{2^{1+\nu}}{2^{k+1}} \sum_{i=1}^{k} (\tfrac{1}{2^\nu})^i + \frac{1}{2^{k(1+\nu)}}\right) C\|x-y\|^{1+\nu} = \tfrac{2^{1+\nu}}{2^{k+1}} C\|x-y\|^{1+\nu} \sum_{i=1}^{k+1} (\tfrac{1}{2^\nu})^i,
\end{aligned}$$

which completes the induction.

Finally, using the inequality (3.3) and the above bound, we obtain

$$\begin{aligned}
\varphi(\lambda) &= \varphi\big(\mu\tfrac{1}{2^s} + (1-\mu)\tfrac{1}{2^{s-1}}\big) = h\big(\mu z_{2^{-s}} + (1-\mu)z_{2^{1-s}}\big) - \lambda h(x) - (1-\lambda)h(y) \\
&\leq \mu h(z_{2^{-s}}) + (1-\mu)h(z_{2^{1-s}}) + C\|z_{2^{-s}} - z_{2^{1-s}}\|^{1+\nu} - \lambda h(x) - (1-\lambda)h(y) \\
&= \mu\varphi(\tfrac{1}{2^s}) + (1-\mu)\varphi(\tfrac{1}{2^{s-1}}) + \tfrac{C}{2^{s(1+\nu)}}\|x-y\|^{1+\nu} \\
&\leq \left(\mu\tfrac{2^{1+\nu}}{2^s} \sum_{i=1}^{s} (\tfrac{1}{2^\nu})^i + (1-\mu)\tfrac{2^{1+\nu}}{2^{s-1}} \sum_{i=1}^{s-1} (\tfrac{1}{2^\nu})^i + \tfrac{1}{2^{s(1+\nu)}}\right) C\|x-y\|^{1+\nu}
\end{aligned}$$

$$\leq \left(\mu 2^{1+\nu}\sum_{i=1}^{s}(\tfrac{1}{2^\nu})^i + (1-\mu)2^{2+\nu}\sum_{i=1}^{s-1}(\tfrac{1}{2^\nu})^i + \tfrac{1}{2^{s\nu}}\right)\lambda C\|x-y\|^{1+\nu}$$
$$\leq\leq \overline{C}\lambda(1-\lambda)\|x-y\|^{1+\nu},$$

where $\overline{C} := \left(\frac{2^{3+\nu}}{2^\nu-1} + 2^{1-\nu}\right)C$. This verifies (3.2) and completes the proof. □

Remark 3.3 To the best of our knowledge, Rolewicz [59, 60] pioneered the study of paraconvexity, introducing this concept for the class of multi-valued mappings. This property reduces to convexity when $\nu > 1$, due to [61, Proposition 7]. Furthermore, several important properties and characterizations, discussed later in this section, hold when $0 < \nu \leq 1$.

The next proposition states that continuous midpoint ν-paraconvex functions are ν-paraconvex, which result simplifies the verification of paraconvexity.

Proposition 3.4 (Midpoint paraconvexity). *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper continuous function on a nonempty convex set $S \subseteq \mathrm{dom}(h)$. Given some $\nu \in (0, 1]$, the function h is ν-paraconvex on S if and only if it is midpoint ν-paraconvex on S, that is, there exists the constant $\rho > 0$ such that*

$$h\left(\frac{x+y}{2}\right) \leq \frac{1}{2}h(x) + \frac{1}{2}h(y) + \frac{\rho}{2}\|x-y\|^{1+\nu}, \qquad \forall x, y \in S. \tag{3.5}$$

Proof The necessity follows directly from Definition 3.1. To justify sufficiency, we assume a constant $\rho > 0$ exists such that (3.5) holds. Let $x, y \in S$. It is clear that the set $\Lambda := \left\{\frac{p}{2^k} : k, p \in \mathbb{N}, p \leq 2^k\right\}$ is dense in $[0, 1]$. Thus, it suffices to verify that the inequality (3.1) holds for each $\lambda \in \Lambda$. In particular, we show by induction that for each $k \in \mathbb{N}$, it holds that

$$h\left(\frac{px+qy}{2^k}\right) \leq \frac{p}{2^k}h(x) + \frac{q}{2^k}h(y) + \rho\left(\sum_{i=1}^{k}\frac{1}{2^i}\right)\|x-y\|^{1+\nu}, \tag{3.6}$$

for all $p, q \in \mathbb{N}$ such that $p + q = 2^k$. For $k = 1$, this inequality holds thanks to (3.5). Suppose (3.6) is satisfied for some $k \in \mathbb{N}$. We show it for $k + 1$. Let $p, q \in \mathbb{N}$ such that $p + q = 2^{k+1}$. If $p = q = 2^k$, then there is nothing to prove. Without loss of generality, assume that $p > 2^k > q$, i.e.,

$$\begin{aligned} h\left(\frac{px+qy}{2^{k+1}}\right) &= h\left(\frac{\frac{(p-2^k)x+qy}{2^k} + x}{2}\right) \leq \frac{1}{2}h\left(\frac{(p-2^k)x+qy}{2^k}\right) \\ &\quad + \frac{1}{2}h(x) + \frac{\rho}{2}\left\|\frac{(p-2^k)x+qy}{2^k} - x\right\|^{1+\nu} \\ &\leq \frac{p}{2^{k+1}}h(x) + \frac{q}{2^{k+1}}h(y) + \rho\left(\sum_{i=1}^{k+1}\frac{1}{2^i}\right)\|x-y\|^{1+\nu}, \end{aligned}$$

which comes from the inductive assumption, completing the inductive step. Now, on the basis of the facts that Λ is dense in $[0, 1]$, $\sum_{i=1}^{\infty}\frac{1}{2^i} = 1$, and h is a continuous function, it is concluded that

$$h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y) + \rho\|x-y\|^{1+\nu},$$

for each $\lambda \in [0, 1]$. Thus, the function h is ν-paraconvex on S by Proposition 3.2, completing the proof. □

In the following, we provide some notable examples of the class of ν-paraconvex functions.

Example 3.5 *(Known paraconvex functions)* We next mention several classes of paraconvex functions.

(a) One notable example of ν-paraconvex functions lies in the class of composite optimization problems:

$$\min_{x\in S} h(x) := \varphi(g(x)), \tag{3.7}$$

where S is a convex set, $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a convex and Lipschitz function with constant $L_0 > 0$ on S, and the Jacobian of $g : \mathbb{R}^m \to \mathbb{R}^n$, $J(x)$, is Hölder continuous of order $\nu \in (0, 1]$ with constant $L_1 > 0$ on S:

$$\|J(x) - J(y)\| \le L\|x - y\|^{\nu}, \quad \forall x, y \in S.$$

It can be shown that the function $h = \varphi \circ g$ is ν-paraconvex with constant $\frac{2^{1-\nu}L_0L_1}{1+\nu}$ on S; see Proposition 3.11.

(b) Any smooth function $h : \mathbb{R}^n \to \mathbb{R}$ whose gradient is Hölder continuous of order $\nu \in (0, 1]$ with constant $L > 0$ on a convex set S, is ν-paraconvex function with constant $\frac{2^{1-\nu}L}{1+\nu}$ on S.

(c) In robust principal component analysis (PCA), the objective is to identify sparse corruptions in a low-rank matrix. Given an $m \times n$ matrix A, the goal is to decompose it as $A = UV^T$ where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ are low-rank factor matrices with r denoting the desired rank. This problem can be formulated as:

$$\min \|UV^T - A\|_1 \ \text{ s.t. } \ U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}.$$

The Jacobian of the function $(U, V) \mapsto UV^T - A$ is Lipschitz continuous with constant $L_1 = 1$. Therefore, the objective function, which fits into the composite form (3.7), is 1-paraconvex by Proposition 3.11; see also [21]. Moreover, since the function is semialgebraic, it satisfies the HEB locally; see [8, Theorem 6.5]. In addition, by incorporating a suitable regularization term of the form $\|U^TU - V^TV\|_F$, the resulting objective function $\frac{1}{m}\|UV^T - A\|_1 + \lambda\|U^TU - V^TV\|_F$ satisfies the HEB of order $\delta = 1$ and remains 1-paraconvex function; see [42, Propositions 5 and 6].

(d) As another example, consider the Robust Phase Retrieval problem; see [20, 22, 23]. While the classical formulation targets squared magnitude conditions, we extend this to the more flexible p-power regime for $p \in (1, 2]$ by seeking a solution x that satisfies $|\langle a_i, x\rangle|^p \approx b_i$ for $i = 1, \ldots, m$, where $\{a_i\}_{i=1}^m \subseteq \mathbb{R}^n$ and $\{b_i\}_{i=1}^m \subseteq \mathbb{R}$. This leads to the Robust generalized Phase Retrieval problem as follows

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{m}\sum_{i=1}^{m} \Big||\langle a_i, x\rangle|^p - b_i\Big|,$$

a formulation that serves as a robust alternative to the standard least-squares approach. By choosing $1 < p \le 2$, the model might become resilient to impulsive noise and heavy-tailed outliers often encountered in real-world sensor data. From an optimization perspective, this objective is of composite form (3.7), is $(p - 1)$-paraconvex and semialgebraic, satisfying a Hölderian error bound.

(e) The deployment of regularizers, ranging from ℓ_1- and ℓ_2- norms to nonconvex and weakly convex penalties, has attracted significant attention in optimization. In applications such as signal processing and machine learning, these penalties are essential for promoting sparsity, ensuring coercivity, and providing robustness against noise. In some cases, it has

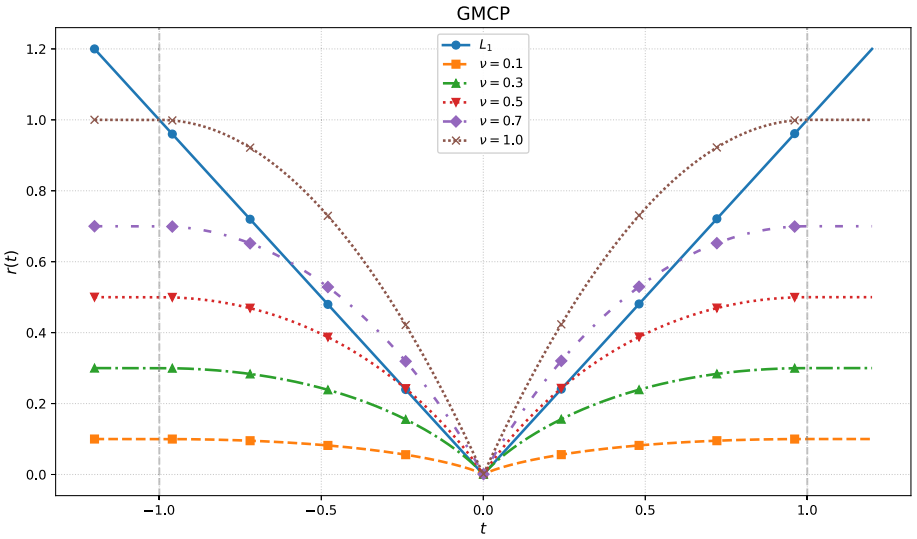


Fig. 1 The ℓ_1-norm regularization and the GMCP with $\nu \in \{0.1, 0.3, 0.5, 0.7, 1\}$ and $\omega = 1$

been known that nonconvex regularizers like the Minimax Concave Penalty (MCP) and SCAD have proven effective in achieving nearly unbiased sparse recovery [15, 44, 64, 73, 75]. In this work, we introduce a specialized penalty, the Generalized MCP (GMCP), defined as

$$r(t) = \begin{cases} (1+\nu)\omega|t| - \omega^{1+\nu}|t|^{1+\nu}, & \text{if } |t| \le \frac{1}{\omega}, \\ \nu, & \text{if } |t| > \frac{1}{\omega}, \end{cases}$$

where $\nu \in (0, 1]$ and the scaling parameter $\omega > 0$; see Figure 1. From a theoretical standpoint, this function is ν-paraconvex with a constant $\rho = \omega^{1+\nu}$ by Theorem 3.12 (b). Consequently, due to Proposition 3.9 (a) for any convex loss function $h : \mathbb{R}^n \to \mathbb{R}$ and any regularization parameter $\lambda > 0$, the composite objective $x \mapsto h(x) + \lambda \sum_{i=1}^n r(x_i)$ is globally ν-paraconvex. □

A further extension of convexity in the nonconvex setting is ν-weak convexity [18]. Given a nonempty open set $\mathcal{U} \subseteq \mathbb{R}^n$ and $0 < \nu \le 1$, a locally Lipschitz function $\varphi : \mathcal{U} \to \mathbb{R}$ is said to be ν-weakly convex, if for any $x_0 \in \mathcal{U}$, there exist $\delta > 0$ and $C > 0$ such that for all $x, y \in \mathbb{B}(x_0; \delta)$ and $\lambda \in [0, 1]$,

$$\varphi(\lambda x + (1-\lambda)y) \le \lambda\varphi(x) + (1-\lambda)\varphi(y) + C\lambda(1-\lambda)\|x - y\|^{1+\nu}.$$

This concept is a local version of ν-paraconvexity. Hence, ν-paraconvexity implies ν-weak convexity. Furthermore, Theorem 3.7 demonstrates that the converse holds on compact convex sets.

Lemma 3.6 *Let $\mathcal{I}$ be an interval in $\mathbb{R}$ and $a, b \in I$ such that $a < b$. Let $\varphi : \mathcal{I} \to \mathbb{R}$ be a ν-paraconvex function on $\mathcal{I} \cap [a, +\infty)$ and $\mathcal{I} \cap (-\infty, b]$ for some $\nu \in (0, 1]$ with constant $\rho > 0$. Then, φ is a ν-paraconvex function on $\mathcal{I}$ with constant 6ρ.*

Proof If $a = \inf \mathcal{I}$ or $b = \sup \mathcal{I}$, then there is nothing to prove. Therefore, without loss of generality, assume that $a, b \in \operatorname{int} \mathcal{I}$. Let $x, y \in \mathcal{I}$ with $x < y$, and let $0 < \lambda < 1$. Set

$z = \lambda x + (1-\lambda)y$. We aim to show that

$$\varphi(z) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y) + 6\rho \min\{\lambda, 1-\lambda\}(y-x)^{1+\nu}. \tag{3.8}$$

If $a < x$ or $y < b$, the inequality (3.8) is trivially valid due to the ν-paraconvexity of h on $\mathcal{I} \cap [a, +\infty)$ and $\mathcal{I} \cap (-\infty, b]$. Now, assume that $x < a < b < y$. Three possible cases arise: (i) $z \in (x, a)$; (ii) $z \in [a, b]$; (iii) $z \in (b, y)$.

In Case (i), $z \in (x, a)$, there exists $r, s, t \in (0, 1)$ such that $z = ra + (1-r)x$, $a = sb + (1-s)x$, and $b = ty + (1-t)a$. By substituting these relations, we get $a = sty + s(1-t)a + (1-s)x$ and $b = ty + (1-t)sb + (1-t)(1-s)x$, leading to

$$\begin{aligned} a &= \tfrac{st}{1-s(1-t)}y + \tfrac{1-s}{1-s(1-t)}x, \quad b = \tfrac{t}{1-s(1-t)}y + \tfrac{(1-t)(1-s)}{1-s(1-t)}x, \\ z &= \left((1-r) + \tfrac{r(1-s)}{1-s(1-t)}\right)x + \tfrac{rst}{1-s(1-t)}y. \end{aligned} \tag{3.9}$$

Thus, $\lambda = (1-r) + \frac{r(1-s)}{1-s(1-t)}$. Furthermore, the ν-paraconvexity of φ on $\mathcal{I} \cap [a, +\infty)$ and $\mathcal{I} \cap (-\infty, b]$ with constant ρ together with $\min\{\lambda, 1-\lambda\} \leq 2\lambda(1-\lambda)$ verifies that

$$\varphi(b) \leq t\varphi(y) + (1-t)\varphi(a) + 2\rho t(1-t)(y-a)^{1+\nu}, \tag{3.10}$$

$$\varphi(a) \leq s\varphi(b) + (1-s)\varphi(x) + 2\rho s(1-s)(b-x)^{1+\nu}, \tag{3.11}$$

$$\varphi(z) \leq r\varphi(a) + (1-r)\varphi(x) + 2\rho r(1-r)(a-x)^{1+\nu}. \tag{3.12}$$

Combining (3.10) and (3.11) ensures

$$\varphi(a) \leq s\left(t\varphi(y) + (1-t)\varphi(a) + 2\rho t(1-t)(y-a)^{1+\nu}\right) + (1-s)\varphi(x) + 2\rho s(1-s)(b-x)^{1+\nu},$$

and consequently,

$$\varphi(a) \leq \tfrac{st}{1-s(1-t)}\varphi(y) + \tfrac{1-s}{1-s(1-t)}\varphi(x) + 2\rho\left(\tfrac{st(1-t)}{1-s(1-t)}(y-a)^{1+\nu} + \tfrac{s(1-s)}{1-s(1-t)}(b-x)^{1+\nu}\right).$$

Now, the inequality (3.8) is confirmed by (3.12), i.e.,

$$\begin{aligned} \varphi(z) &\leq \left((1-r) + \tfrac{r(1-s)}{1-s(1-t)}\right)\varphi(x) + \tfrac{rst}{1-s(1-t)}\varphi(y) \\ &\quad + 2\rho\left(\tfrac{rst(1-t)}{1-s(1-t)}(y-a)^{1+\nu} + \tfrac{rs(1-s)}{1-s(1-t)}(b-x)^{1+\nu} + r(1-r)(a-x)^{1+\nu}\right) \\ &\leq \lambda\varphi(x) + (1-\lambda)\varphi(y) + 6\rho\lambda(1-\lambda)(y-x)^{1+\nu} \\ &\leq \lambda\varphi(x) + (1-\lambda)\varphi(y) + 6\rho\min\{\lambda, 1-\lambda\}(y-x)^{1+\nu}. \end{aligned}$$

To show Case (ii), we assume that $z \neq a$. There exist $r, s \in (0, 1)$ such that $z = ra + (1-r)y$ and $a = sx + (1-s)z$ implying $z = \frac{rs}{1-r(1-s)}x + \frac{1-r}{1-r(1-s)}y$. So, $\lambda = \frac{rs}{1-r(1-s)}$. Moreover, relying again on ν-paraconvexity of φ we deduce

$$\begin{aligned} \varphi(z) &\leq r\varphi(a) + (1-r)\varphi(y) + 2\rho r(1-r)(y-a)^{1+\nu} \\ &\leq r\left(s\varphi(x) + (1-s)\varphi(z) + 2\rho s(1-s)(z-x)^{1+\nu}\right) \\ &\quad + (1-r)\varphi(y) + 2\rho r(1-r)(y-a)^{1+\nu}, \end{aligned}$$

which comes out as (3.8). In the case where $z = a$, we have $z \neq b$, and following a similar argument yields the desired result. The proof of Case (iii) is similar to Cases (ii)-(iii) and is therefore omitted. □

Our next result indicates that ν-weak convexity leads to ν-paraconvexity on compact convex sets.

Theorem 3.7 (ν-weak convexity and ν-paraconvexity). *Let $\mathcal{U} \subseteq \mathbb{R}^n$ be an open set and $S \subseteq \mathcal{U}$ be a nonempty convex and compact set. Let $h : \mathcal{U} \to \mathbb{R}$ be a ν-weakly convex function for some $\nu \in (0, 1]$. Then, h is a ν-paraconvex function on S.*

Proof Let us begin by showing that the inequality of ν-weak convexity holds uniformly on S, with a fixed constant $C > 0$ independent of $x \in S$. The ν-weak convexity of h implies that for each point $x \in S$, there exist constants $C_x > 0$ and $\delta_x > 0$ such that for all $x_0, y_0 \in \mathbb{B}(x; \delta_x)$ and $\lambda \in [0, 1]$,

$$h(\lambda x_0 + (1-\lambda)y_0) \leq \lambda h(x_0) + (1-\lambda)h(y_0) + C_x\lambda(1-\lambda)\|x_0 - y_0\|^{1+\nu}. \tag{3.13}$$

The collection of open balls, $\{\mathbb{B}(x; \delta_x)\}_{x\in S}$ covers S, and by the compactness principle, a finite subcollection of them covers S. Thus, there exist points $x_i \in S$ and radii $\delta_i > 0$, $i = 1, 2, \ldots, m$, such that $S \subseteq \bigcup_{i=1}^m \mathbb{B}(x_i; \delta_i)$. Define $\overline{C} := \max\{C_i : \ i = 1, 2, \ldots, m\}$. Then, for each $x \in S$, there exists some $\delta_x > 0$ such that $\mathbb{B}(x; \delta_x) \subseteq \mathbb{B}(x_i; \delta_i)$, namely for an index i such that $x \in \mathbb{B}(x_i; \delta_i)$, and (3.13) holds on $\mathbb{B}(x; \delta_x)$ with the fixed constant $\overline{C}$.

Let $x, y \in S$ and $\mu \in [0, 1]$ be arbitrary and constant hereafter. Let us consider the function $\varphi : [0, 1] \to \mathbb{R}$ defined as $\varphi(t) := h(z_t)$, where $z_t = y + t(x - y)$ and $t \in [0, 1]$. We aim at showing that φ is ν-paraconvex on $[0, 1]$ with constant $9\overline{C}\|y - x\|^{1+\nu}$. Once this is established, we have

$$\begin{aligned} h(z_\mu) = \varphi(\mu) &\leq \mu\varphi(1) + (1-\mu)\varphi(0) + 9\overline{C}\|y - x\|^{1+\nu}\mu(1-\mu)|1-0|^{1+\nu} \\ &= \mu h(x) + (1-\mu)h(y) + 9\overline{C}\mu(1-\mu)\|y - x\|^{1+\nu}, \end{aligned}$$

as our desired result. Let $\hat{t} \in (0, 1)$. Inasmuch as S is convex, it follows that $z_{\hat{t}} \in S$. So, there exists $\delta > 0$ such that for all $x_0, y_0 \in \mathbb{B}(z_{\hat{t}}; \delta)$ and $\lambda \in [0, 1]$, the inequality (3.13) is satisfied with constant $\overline{C}$. Let us consider $\epsilon := \frac{1}{2}\min\{1 - \hat{t}, \hat{t}, \frac{\delta}{\|y-x\|}\}$. Suppose $t_1, t_2 \in [\hat{t} - \epsilon, \hat{t} + \epsilon]$ and $\lambda \in [0, 1]$. Then, $z_{t_1}, z_{t_2} \in \mathbb{B}(z_{\hat{t}}; \delta)$ and therefore

$$h(\lambda z_{t_1} + (1-\lambda)z_{t_2}) \leq \lambda h(z_{t_1}) + (1-\lambda)h(z_{t_2}) + \overline{C}\lambda(1-\lambda)\|z_{t_1} - z_{t_2}\|^{1+\nu},$$

implying

$$\varphi(\lambda t_1 + (1-\lambda)t_2) \leq \lambda\varphi(t_1) + (1-\lambda)\varphi(t_2) + \overline{C}\|y - x\|^{1+\nu}\lambda(1-\lambda)|t_1 - t_2|^{1+\nu}.$$

Hence, φ is ν-paraconvex on $[\hat{t} - \epsilon, \hat{t} + \epsilon]$ with constant $\overline{C}\|y - x\|^{1+\nu}$. Now, let us define

$$A := \left\{t \in [0, 1] : \ \varphi \text{ is } \nu - \text{paraconvex on } [\hat{t} - \epsilon, t] \text{ with constant } 3\overline{C}\|y - x\|^{1+\nu}\right\},$$

and consider $a := \sup A$. Clearly, $\hat{t}+\epsilon \in A$. So, the supremum is well-defined and $0 < a \leq 1$. It is easy to see that $a \in A$ using continuity of h (see Theorem 3.8). We claim that $a = 1$. Otherwise, if $a \in (0, 1)$, the argument applied for $\hat{t}$ can similarly be repeated with a to conclude that φ is ν-paraconvex on $[a - \bar{\epsilon}, a + \bar{\epsilon}] \subseteq (0, 1)$ for some $\bar{\epsilon} > 0$ with constant $\overline{C}\|y - x\|^{1+\nu}$. In this case, the function φ is ν-paraconvex on $[\hat{t} - \epsilon, a]$ and $[a - \bar{\epsilon}, a + \bar{\epsilon}]$ with constant $\overline{C}\|y - x\|^{1+\nu}$, and so on $[\hat{t} - \epsilon, a + \bar{\epsilon}]$ with constant $3\overline{C}\|y - x\|^{1+\nu}$, by Lemma 3.6. This contradicts $a = \sup A$. Thus, $a = 1$. Hence, φ is ν-paraconvex on $[\hat{t} - \epsilon, 1]$ with constant $3\overline{C}\|y - x\|^{1+\nu}$. Similarly, it can be shown that φ is ν-paraconvex on $[0, \hat{t} + \epsilon]$ with constant $3\overline{C}\|y - x\|^{1+\nu}$. Therefore, φ is ν-paraconvex on $[0, 1]$ with constant $9\overline{C}\|y - x\|^{1+\nu}$, due to Lemma 3.6, and the proof is completed. □

In what follows, we investigate several properties and characterizations of the class of ν-paraconvex functions. First, similar to the convex case, we observe that ν-paraconvex functions exhibit local Lipschitz continuity. Consequently, their Clarke's subdifferential is well defined and nonempty at interior points of their domain; see [17, Proposition 2.1.2]. Although this property was proven in [33, 34] under the assumptions of local boundedness or continuity, we establish here that ν-paraconvex functions are locally bounded, thus directly satisfying the local Lipschitz continuity property.

Theorem 3.8 (Locally bounded and Lipschitz properties). *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper ν-paraconevex function on a convex set $S \subseteq \mathrm{dom}(h)$ with $\nu \in (0, 1]$ and constant $\rho > 0$. Then the following assertions hold:*

(a) The function h is locally bounded on int S*;*
(b) The function h is locally Lipschitz on int S*.*

Proof (a) Let $\bar{x} \in \mathrm{int}\, S$. There exists $\epsilon > 0$ such that $P := \{x \in \mathbb{R}^n : \|x - \bar{x}\|_\infty \leq \epsilon\} \subseteq S$. Clearly, polyhedral P has 2^n vertexes. Assume that E is the set of all vertices of the polyhedral P and $M := \max_{x\in E} h(x)$. Let $x \in P$. Then, thanks to the Minkowski-Weyl representation [7, Proposition 3.2.2] and extension of Caratheodory's theorem [7, Theorem 1.22], there exist scalars $\lambda_j \in [0, 1]$ and vertices $x^j \in E$, $j = 1, 2 \ldots, n + 1$, such that $x = \sum_{j=1}^{n+1} \lambda_j x^j$ and $\sum_{j=1}^{n+1} \lambda_j = 1$. Using approximate Jensen's inequality [68, Lemma 2.1], we deduce

$$\begin{aligned} h(x) &\leq \sum_{j=1}^{n+1} \lambda_j h(x^j) + \rho \sum_{j=1}^{n+1} \lambda_j (1 - \lambda_j) \max_{1\leq k\leq n+1} \|x^k - x^j\|^{1+\nu} \\ &\leq M + \frac{n+1}{4} \sqrt{n}^{1+\nu} 2^{1+\nu} \epsilon^{1+\nu} \rho. \end{aligned}$$

Now, assume that $x \in cl\mathbb{B}(\bar{x}; \epsilon) \subseteq P$. Set, $\bar{\lambda} := \frac{\epsilon}{\epsilon+\|x-\bar{x}\|}$. Then,

$$h(\bar{x}) \leq \bar{\lambda} h(x) + (1 - \bar{\lambda}) h\left(\tfrac{1}{1-\bar{\lambda}}\bar{x} - \tfrac{\bar{\lambda}}{1-\bar{\lambda}}x\right) + \rho \min\{\bar{\lambda}, 1 - \bar{\lambda}\} \left\|x - \tfrac{1}{1-\bar{\lambda}}\bar{x} + \tfrac{\bar{\lambda}}{1-\bar{\lambda}}x\right\|^{1+\nu},$$

i.e.,

$$\begin{aligned} h(x) &\geq \tfrac{1}{\bar{\lambda}} h(\bar{x}) - \tfrac{1-\bar{\lambda}}{\bar{\lambda}} h\left(\tfrac{1}{1-\bar{\lambda}}\bar{x} - \tfrac{\bar{\lambda}}{1-\bar{\lambda}}x\right) - \tfrac{\rho \min\{\bar{\lambda}, (1-\bar{\lambda})\}}{\bar{\lambda}} (\epsilon + \|x - \bar{x}\|)^{1+\nu} \\ &\geq h(\bar{x}) - M - \left(\tfrac{n+1}{4}\sqrt{n}^{1+\nu} - 1\right) 2^{1+\nu} \epsilon^{1+\nu} \rho. \end{aligned}$$

Setting

$$\overline{M} := \max\left\{M + \tfrac{n+1}{4}\sqrt{n}^{1+\nu} 2^{1+\nu} \epsilon^{1+\nu} \rho, \left|h(\bar{x}) - M - \left(\tfrac{n+1}{4}\sqrt{n}^{1+\nu} - 1\right) 2^{1+\nu} \epsilon^{1+\nu} \rho\right|\right\},$$

we come to $|h(x)| \leq \overline{M}$ for all $x \in cl\, \mathbb{B}(\bar{x}; \epsilon)$, i.e., h is locally bounded on int S. Assertion (b) follows from Assertion (a) in combination with [33, Lemma 2.5] and [34, Proposition 2.2], adjusting our desired result. □

In what follows, we investigate some calculus around the class of ν-paraconvex functions. Since the proofs of these results are straightforward consequences of Definition 3.1, we omit the details of the proofs.

Proposition 3.9 (Paraconvexity calculus). *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and ν-paraconvex function on a convex set $S \subseteq \mathrm{dom}(h)$ for some $\nu \in (0, 1]$ with constant $\rho > 0$, let $h_i : \mathbb{R}^n \to$*

$\overline{\mathbb{R}}$, $i \in \mathcal{I}$ *where* $\mathcal{I}$ *is an index set, be proper and* ν_i*-paraconvex functions on a convex set* $S_{\mathcal{I}} \subseteq \mathrm{dom}(h_i)$ *for some* $\nu_i \in (0, 1]$ *with constant* $\rho_i > 0$*, and let* $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be a convex function on* $S \subseteq \mathrm{dom}(g)$*. Then, the following statements hold:*

(a) The function $g + \varrho h$ *is* ν*-paraconvex on* S *with constant* $\varrho\rho$ *for any* $\varrho > 0$*;*
(b) The function h *is* ν*-paraconvex on* S *with any constant* $\varrho \in (\rho, \infty)$*;*
(c) The function ϱh *is* ν*-paraconvex on* S *with constant* $\varrho\rho$ *for any* $\varrho > 0$*;*
(d) If S *is bounded with diameter* $K > 0$*, then* h *is* ϑ*-paraconvex on* S *with constant* $\rho K^{\nu-\vartheta}$ *for any* $0 < \vartheta < \nu$*;*
(e) If $\sum_i \rho_i < \infty$ *and* $\vartheta = \nu_i$*,* $i \in \mathcal{I}$*, then* $\sum_i h_i$ *is* ϑ*-paraconvex on* $S_{\mathcal{I}}$ *with constant* $\sum_i \rho_i$*;*
(f) If $S_{\mathcal{I}}$ *is bounded with diameter* $K > 0$*,* $0 < \vartheta = \inf_i \nu_i$*, and* $\sum_i \rho_i K^{\nu_i - \vartheta} < \infty$*, then* $\sum_i h_i$ *is* ϑ*-paraconvex on* $S_{\mathcal{I}}$ *with constant* $\sum_i \rho_i K^{\nu_i - \vartheta}$*;*
(g) If $\sup_i \rho_i < \infty$*, and* $\vartheta = \nu_i$*,* $i \in \mathcal{I}$*, then* $\sup_i \{h_i(x)\}$ *is* ϑ*-paraconvex on* $S_{\mathcal{I}}$ *with constant* $\rho := \sup_i \rho_i$*.*

We next present characterizations of the ν-paraconvexity that are the straightforward consequences of the definition. Some of these results on the whole space have been previously investigated in the literature; see [34]. We include them here for completeness.

Recall that, by [17, Proposition 2.1.2], if a function is locally Lipschitz continuous at an interior point of its domain, then its Clarke's subdifferential is well defined and nonempty at that point.

Theorem 3.10 (First order characterization). *Let* $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be a proper and locally Lipschitz function on a nonempty convex set* $S \subseteq \mathrm{int}\,\mathrm{dom}(h)$*. Let* $\nu \in (0, 1]$ *and* $\rho \geq 0$*. Then, the following statements hold:*

(a) If S *is open and* h *is a* ν*-paraconvex function on* S *with constant* ρ*, then for any* $x, y \in S$ *and any* $\zeta \in \partial h(x)$*,*

$$h(y) \geq h(x) + \langle \zeta, y - x \rangle - \rho \|y - x\|^{1+\nu}. \tag{3.14}$$

Conversely, if inequality (3.14) holds for any $x, y \in S$ *and any* $\zeta \in \partial h(x)$*, then* h *is* ν*-paraconvex function on* S *with constant* $2^{1-\nu}\rho$*;*

(b) If S *is open and* h *is a* ν*-paraconvex function on* S *with constant* ρ*, then* ∂h *is* ν*-paramonotone on* S *with constant* 2ρ *, i.e., for any* $x, y \in S$*, any* $\zeta \in \partial h(x)$*, and any* $\eta \in \partial h(y)$*,*

$$\langle \eta - \zeta, y - x \rangle \geq -2\rho \|x - y\|^{1+\nu}. \tag{3.15}$$

Conversely, if ∂h *is* ν*-paramonotone on* S *with constant* ρ*, then* h *is* ν*-paraconvex on* S *with constant* $\frac{2^{1-\nu}\rho}{1+\nu}$*.*

Proof (a) For the first implication, let $x, y \in S$ and $\zeta \in \partial h(x)$. Define $d := y - x$. Then, for any $z \in S$ sufficiently close to x and any $0 < t < 1$, we obtain

$$h(z + d) - h(z) \geq \tfrac{h(z+td)-h(z)}{t} - \rho \tfrac{\min\{t, 1-t\}}{t} \|d\|^{1+\nu}.$$

Taking the limit superior on both sides as $t \to 0$ and $z \to x$, and using continuity of h yields inequality (3.14). To verify the converse implication, let $x, y \in S$, $0 < \lambda < 1$, and $\zeta \in \partial h(\lambda x + (1 - \lambda)y)$. By (3.14), we obtain

$$h(x) \geq h(\lambda x + (1 - \lambda)y) - (1 - \lambda)\langle \zeta, y - x \rangle - \rho(1 - \lambda)^{1+\nu} \|y - x\|^{1+\nu},$$

$$h(y) \geq h(\lambda x + (1 - \lambda)y) + \lambda \langle \zeta, y - x \rangle - \rho\lambda^{1+\nu} \|y - x\|^{1+\nu}.$$

Combining these inequalities gives

$$\begin{aligned}\lambda h(x)+(1-\lambda)h(y) &\geq h(\lambda x+(1-\lambda)y)-\rho(\lambda(1-\lambda)^{1+\nu}+(1-\lambda)\lambda^{1+\nu})\|y-x\|^{1+\nu}\\ &\geq h(\lambda x+(1-\lambda)y)-2^{1-\nu}\rho\min\{\lambda,(1-\lambda)\}\|y-x\|^{1+\nu},\end{aligned}$$

where the second inequality follows from the concavity of function $t\mapsto(1-t)^{\nu}+t^{\nu}$ on $[0,1]$ which attains its maximum at $t=1/2$ implying $(1-\lambda)^{\nu}+\lambda^{\nu}\leq 2^{1-\nu}$.

(b) The necessity implication follows immediately from (3.14). For the converse implication, let $x,y\in S$ and $\zeta\in\partial h(x)$. Then, there exists a measurable selection $\eta_t\in\partial h(ty+(1-t)x)$ for almost every $t\in[0,1]$ such that

$$h(y)-h(x)-\langle\zeta,y-x\rangle=\int_0^1\langle\eta_t-\zeta,y-x\rangle dt\geq\int_0^1-\rho t^{\nu}\|y-x\|^{1+\nu}dt=-\tfrac{\rho}{1+\nu}\|y-x\|^{1+\nu},$$

completing the proof in view of Assertion (a). □

Applying the above characterization, we show the ν-paraconvexity of a composite function.

Proposition 3.11 (Composite paraconvex function). *Let $S\subseteq\mathbb{R}^n$ be a convex set and let $\varphi:\mathbb{R}^n\to\mathbb{R}$ be a convex and Lipschitz function with constant $L_0>0$ on S. Assume that the Jacobian of $g:\mathbb{R}^m\to\mathbb{R}^n$ is Hölder continuous of order $\nu\in(0,1]$ with constant $L_1>0$ on S. Then, the composite function $h:=\varphi\circ g$ is ν-paraconvex with constant $\rho=\frac{2^{1-\nu}L_0L_1}{1+\nu}$ on S.*

Proof The Hölder continuity of Jacobian of g ensures

$$\|g(y)-g(x)-J(x)(y-x)\|\leq\tfrac{L_1}{1+\nu}\|y-x\|^{1+\nu},\quad\forall x,y\in S,$$

as stated in [76, Lemma 1]. Let $x,y\in S$ and $\zeta\in\partial h(x)$. By the chain rule theorem, [17, Theorem 2.3.10], there exists $\eta\in\partial\varphi(g(x))$ such that $\zeta=J(x)^T\eta$. Now, using these properties and the convexity of φ, we get

$$\begin{aligned}&h(y)-h(x)-\langle\zeta,y-x\rangle\\ &\quad=\varphi(g(y))-\varphi(g(x))-\langle J(x)^T\eta,y-x\rangle\geq\langle\eta,g(y)-g(x)\rangle-\langle\eta,J(x)(y-x)\rangle\\ &\quad\geq-\|\eta\|\,\|g(y)-g(x)-J(x)(y-x)\|\geq-\tfrac{L_0L_1}{1+\nu}\|y-x\|^{1+\nu},\end{aligned}$$

i.e., our desired result follows from Theorem 3.10 (a). □

We conclude this subsection by indicating that the convexity of the function $x\mapsto h(x)+C\|x\|^{1+\nu}$ is sufficient for the ν-paraconvexity of h. Building on this observation, we also derive a sufficient second-order characterization.

Theorem 3.12 (Second order characterization). *Let $h:\mathbb{R}^n\to\overline{\mathbb{R}}$ be a proper and continuous function and $S\subseteq\operatorname{dom}(h)$ be a convex set. Let $\nu\in(0,1]$ and $C,\overline{C}>0$. Then, the following assertions hold:*

(a) Let $g:\mathbb{R}^n\to\overline{\mathbb{R}}$ be a continuously differentiable function whose gradient is Hölder continuous of order ν with constant $\overline{C}$ on $S\subseteq\operatorname{int}\operatorname{dom}(g)$. If the function $h+Cg$ is convex on S, then h is ν-paraconvex on S with constant $\frac{2^{1-\nu}\overline{C}C}{1+\nu}$;

(b) If the function $x\mapsto h(x)+C\|x\|^{1+\nu}$ is convex on S, then h is ν-paraconvex on S with constant $4^{1-\nu}C$;

(c) If h is a twice continuously differentiable function on $S \subseteq \operatorname{int}\operatorname{dom}(h)$ and the matrix

$$\nabla^2 h(x) + C\frac{1+\nu}{\|x\|^{3-\nu}}\left(\|x\|^2 I - (1-\nu)xx^T\right),$$

is positive semidefinite for each $x \in S$ with $x \neq 0$, then h is a ν-paraconvex function on S.

Proof (a) By invoking Proposition 3.11, the function g is ν-paraconvex on S with constant $\frac{2^{1-\nu}\overline{C}}{1+\nu}$. Thus, the function h is ν-paraconvex on S with constant $\frac{2^{1-\nu}\overline{C}C}{1+\nu}$ by Proposition 3.9 (a).

(b) According to [58, Theorem 6.3], the gradient of the function $x \mapsto -\|x\|^{1+\nu}$ is a Hölder continuous of order ν with constant $2^{1-\nu}(1+\nu)$. Therefore, the claim holds by Assertion (a).

(c) This assertion follows directly from Assertion (a). □

Corollary 3.13 *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex function on a convex set $S \subseteq \operatorname{dom}(h)$. Then, for each $\nu \in (0, 1]$ and $C > 0$, the function $x \mapsto h(x) - C\|x\|^{1+\nu}$ is ν-paraconvex on S.*

Remark 3.14 In the case that $\nu = 1$, a function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is 1-paraconvex if and only if there exists a nonnegative constant C such that $h(x) + C\|x\|^2$ is convex; see [52, 59, 60]. On the other hand, for the case that $0 < \nu < 1$, Rolewicz [59, Example 1] presented an example of a ν-paraconvex function for which a similar result does not hold. However, we establish that the convexity of $h(x)+C\|x\|^{1+\nu}$ for some $0 < \nu < 1$ and $C > 0$ implies the ν-paraconvexity of h (see Theorem 3.12). Moreover, in Example 3.15, we provide an example that highlights this result, emphasizing the existence of ν-paraconvex functions with $0 < \nu < 1$ that are not 1-paraconvex (weakly convex). Additionally, the function $h(x) = -|x|$ is not ν-paraconvex for any $\nu \in (0, 1]$.

The following example provides a ν-paraconvex function which is not 1-paraconvex (weakly convex).

Example 3.15 Let $\nu \in (0, 1)$. Let us consider the function $h : \mathbb{R} \to \mathbb{R}$ given by

$$h(x) = \begin{cases} 1 - |x|^{1+\nu}, & -1 \leq x \leq 1, \\ x^2 - 1, & o.w, \end{cases}$$

see Figure 2. The function $x \mapsto h(x) + |x|^{1+\nu}$ is convex implying that h is a ν-paraconvex function. Furthermore, it can be seen that for any $\rho > 0$, the function $x \mapsto h(x)+\rho|x|^2$ is not convex and possesses a local maximum at $x = 0$ since $h''(x) + 2\rho \to -\infty$ as $x \to 0$. This indicates that the original function h is not 1- paraconvex (i.e., weakly convex). Nonetheless, the function $x \mapsto h(x) + \rho|x|^2$ also qualifies as a ν-paraconvex function.

□

3.2 Paraconvex optimization: Finding global minima

Here, we consider the optimization problem of the form (1.1). Throughout the paper, we consider the following assumption.

Assumption I We assume that

1. The set $X \subseteq \operatorname{dom}(f)$ is nonempty closed and convex;
2. The set of minimizers $\mathcal{X}^* := \arg\min_{x \in X} f(x)$ is nonempty and f^* is the optimal value of (1.1);

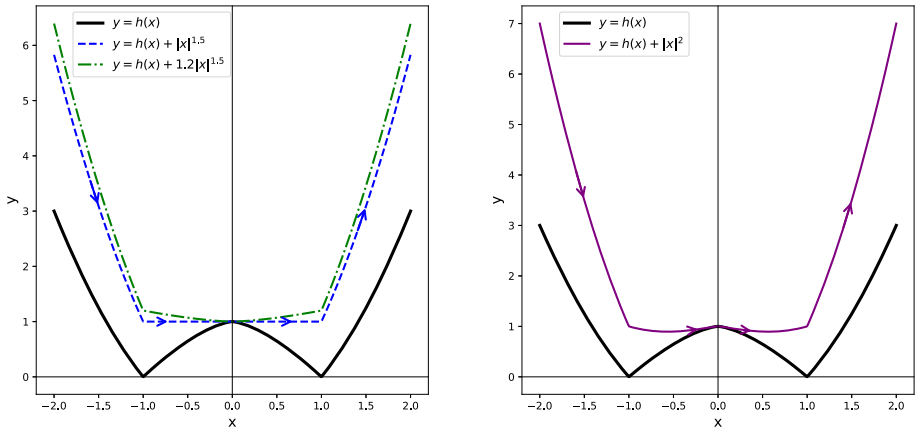


Fig. 2 Comparisons of the function $h(x)$ with $\nu = 0.5$ and its modified versions demonstrating ν-paraconvexity. The left plot shows $h(x)$ (in black), $h(x) + |x|^{1.5}$ (in green), highlighting the convexity of $h(x) + |x|^{1.5}$ which implies h is 0.5-paraconvex. The right plot shows $h(x)$ alongside $h(x) + |x|^2$ (in purple), illustrating that h is not 1-paraconvex

and the following properties are satisfied for parameters $\rho > 0$, $0 < \nu \leq 1$, $\frac{1}{1+\nu} < \delta \leq 1$ and $\mu > 0$:

3. The function f is ν-paraconvex with constant ρ on a open convex set $X' \subseteq \text{dom}(f)$ such that $X \subseteq X'$;
4. The function f admits the HEB of order δ with constant μ.

As discussed in Section 3.1, the class of paraconvex functions includes nonconvex and nonsmooth objectives that arise in structured optimization models. In the following example, we illustrate that such functions may admit saddle points, while their global minimizers can still be located within a relatively large basin of attraction.

Example 3.16 Let us consider the function $h : \mathbb{R}^2 \to \mathbb{R}$ given by $h(x, y) = x^2 + \left(|y|^{1+\nu} - 1\right)^2$ with $\nu \in (0, 1]$. A direct calculation yields

$$\nabla h(x, y) = \begin{bmatrix} 2x \\ 2(1+\nu)\text{sgn}(y)(|y|^{1+2\nu} - |y|^{\nu}) \end{bmatrix},$$

$$\nabla^2 h(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 2(1+\nu)\Big((1+2\nu)|y|^{2\nu} - \nu|y|^{\nu-1}\Big) \end{bmatrix}.$$

It is evident that h is nonconvex; see Figure 3. Moreover, the points $(0, \pm 1)$ and $(0, 0)$ are stationary. The points $(0, \pm 1)$ are global minimizers, whereas $(0, 0)$ is a saddle point. In particular, when $\nu = 1$, the Hessian at $(0, 0)$ has eigenvalues $\lambda_1 = 2$ and $\lambda_2 = -4$, showing that $(0, 0)$ is a strict nondegenerate saddle point. In contrast, for $\nu \in (0, 1)$, the second eigenvalue tends to $-\infty$ as $y \to 0$. Furthermore, the function h is ν-paraconvex with any constant $\rho > 2$. It is also worth noting that, for each global minimizer, the open ball of radius $r < 1$ centered at the minimizer contains no saddle points. □

For many problems of the form (1.1), there is a wide basin around the global minima, as suggested in Figure 3. Hence, if a local optimization algorithm initializes from a point in this

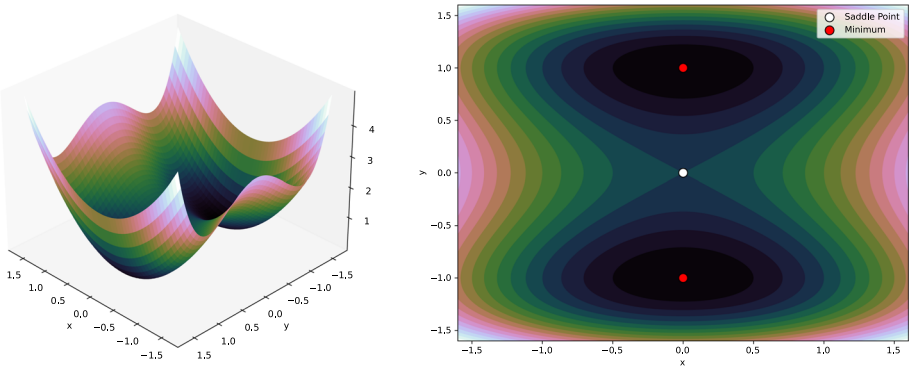


Fig. 3 Stationary points of the function $h(x, y) = x^2 + \left(y^2 - 1\right)^2$: global minima at $(0, \pm 1)$ (red points) and a saddle point at $(0, 0)$ (white point)

basin, it should converge fast to a global minimizer. This motivates the quest for developing two-stage algorithms in which (i) the outer algorithm (e.g., a spectral approach) is used to find a reasonable approximation of a global minimizer in the basin of the cost function, and (ii) the inner algorithm is used to achieve fast convergence to a global minimizer. In more detail, some sufficient conditions can be considered under which variants of spectral methods [16, Section VIII] can bring us to the vicinity of a global minimizer, and then we can apply a proper algorithm (e.g., PSMs; See Section 4) to converge to this global minimizer.

In the next result, we introduce a neighborhood of the optimal solution set of (1.1) that is devoid of extraneous stationary points as a direct result of ν-paraconvexity and Hölderian error boundness. Note that this region ensures a suitable region around the global minia such that if a proper algorithm is initialized from there, then it can generate a sequence converging to the global minimizer. In the following result, let us recall that Assumption I holds.

Lemma 3.17 *Let $x \in X$ be a stationary point of (1.1). Then, either $x \in \mathcal{X}^*$ or*

$$\left(\frac{\mu}{\rho}\right)^{\frac{\delta}{\delta(1+\nu)-1}} \leq \operatorname{dist}(x; \mathcal{X}^*). \tag{3.16}$$

Proof Without loss of generality, let $x \in X \setminus \mathcal{X}^*$ be a stationary point of (1.1). Hence, $0 \in \partial(f + \delta_X)(x)$ and Theorem 3.10 yield that

$$f(y) - f(x) \geq -\rho \|y - x\|^{(1+\nu)}, \quad \forall y \in X.$$

Setting $\bar{x} \in \operatorname{proj}_{\mathcal{X}^*}(x)$ and applying the HEB, we come to

$$\mu\,\mathrm{dist}^{1/\delta}(x;\mathcal{X}^*) \le f(x) - f(\bar{x}) \le \rho\|x-\bar{x}\|^{1+\nu} = \rho\,\mathrm{dist}^{1+\nu}(x;\mathcal{X}^*),$$

leading to (3.16). □

Now, for any $\gamma \in (0,1]$, let us define the tube

$$\mathcal{T}_\gamma := \left\{ x \in X : \mathrm{dist}(x;\mathcal{X}^*) < \left(\gamma\frac{\mu}{\rho}\right)^{\frac{\delta}{\delta(1+\nu)-1}} \right\}, \tag{3.17}$$

which contains no extraneous stationary points of the problem, due to Lemma 3.17. Additionally, we set

$$L := \sup\left\{ \|\zeta\| : \zeta \in \partial f(x),\ x \in \mathcal{T}_1 \right\}.$$

The following lemma provides a key relationship between μ and L. Specifically, for each $x \in \mathcal{T}_1$, the ratio $\tau_x := \tau\,\mathrm{dist}^{\frac{1}{\delta}-1}(x;\mathcal{X}^*)$ lies within the interval $[0,1]$, where $\tau = \frac{\mu}{L}$.

Lemma 3.18 *Let $\gamma \in (0,1]$ and $x \in \mathcal{T}_\gamma$. Then, $\tau_x = \tau\,\mathrm{dist}^{\frac{1}{\delta}-1}(x;\mathcal{X}^*) \in [0,1]$. In addition, if $\delta = 1$, then $\tau_x = \tau \in (0,1]$.*

Proof Let $x \in \mathcal{T}_\gamma$. If $x \in \mathcal{X}^*$, then trivially $\tau_x = 0$ and there is nothing to prove, i.e., without loss of generality, we assume that $x \in \mathcal{T}_\gamma \setminus \mathcal{X}^*$. For $\bar{x} \in \mathrm{proj}_{\mathcal{X}^*}(x)$, invoking the mean value theorem [46, Theorem 3.51], there exist $z \in [x,\bar{x})$ and $\zeta \in \partial f(z)$ satisfying $f(x) - f(\bar{x}) \le \langle \zeta, x-\bar{x}\rangle$. Using the HEB, we come to

$$\mu\,\mathrm{dist}^{1/\delta}(x;\mathcal{X}^*) \le f(x) - f(\bar{x}) \le \|\zeta\|\,\|x-\bar{x}\| = \|\zeta\|\ \mathrm{dist}(x;\mathcal{X}^*).$$

Furthermore, it is easy to see that $z \in \mathcal{T}_\gamma \subseteq \mathcal{T}_1$, and so $\|\zeta\| \le L$. Thus, $\mu\,\mathrm{dist}^{1/\delta}(x;\mathcal{X}^*) \le L\,\mathrm{dist}(x;\mathcal{X}^*)$, and consequently $\tau_x \le 1$. In the special case where $\delta = 1$, it follows that $\tau_x = \tau = \frac{\mu}{L} \ne 0$. □

4 Projected subgradient algorithm for paraconvex optimization

In this section, we present the projected subgradient methods (PSMs) for the nonsmooth and constrained paraconvex optimization problems of the form (1.1) and establish their convergence analysis for several choices of step-sizes. Let us begin with the following generic PSM, where the starting point x_0 satisfies

$$x_0 \in \mathcal{T}_\gamma,$$

as $\mathcal{T}_\gamma$ is defined in (3.17).

Algorithm 1: Generic Projected Subgradient Method (PSM)

Input: $x_0 \in \mathcal{T}_\gamma$; $\gamma \in (0,1]$.

begin

  while *the stopping criteria do not hold* **do**

    Choose $\zeta_k \in \partial f(x_k)$;

    Set $x_{k+1} = \mathrm{proj}_X\left(x_k - \alpha_k \frac{\zeta_k}{\|\zeta_k\|}\right)$ and $k = k+1$;

  Set $x_{k+1} = x_k$;

Output: x_{k+1}.

In the above algorithm, the step-size $\alpha_k > 0$ plays a key role in the algorithm's progression. In fact, by employing different step-sizes, such as fixed, diminishing, and Scaled Polyak's step-sizes, we can define various PSMs, where they can then be compared in terms of the convergence rates and overall numerical performance. Recall that, by Assumption I, the function f is ν-paraconvex on the open convex set $X' \supseteq X$. This property implies that f is locally Lipschitz continuous at each iterate $x_k \in X$; see Theorem 3.8 (b). Consequently, the Clarke's subdifferential $\partial f(x_k)$ is nonempty for all k.

We begin with the subsequent lemma establishing a fundamental recurrence. This result is instrumental in deriving the convergence rate of the PSMs.

Lemma 4.1 (**Basic inequalities I**). *Let x_k and x_{k+1} be two consecutive iterations of Algorithm 1 such that $x_k \in \mathcal{T}_\gamma$ with $\gamma = 1/2$. Then, one has*

$$\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) \leq \operatorname{dist}^2(x_k; \mathcal{X}^*) - \tfrac{\alpha_k}{L}(f(x_k) - f^*) + \alpha_k^2, \tag{4.1}$$

$$\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) \leq \operatorname{dist}^2(x_k; \mathcal{X}^*) - \alpha_k \tau \operatorname{dist}^{\frac{1}{\delta}}(x_k; \mathcal{X}^*) + \alpha_k^2. \tag{4.2}$$

If, in addition, $\alpha_k \leq \min\left\{1, \min\{1, \tau\}(\frac{\mu}{2\rho})^{\frac{\delta}{\delta(1+\nu)-1}}\right\}$, *then* $x_{k+1} \in \mathcal{T}_\gamma$.

Proof For Algorithm 1, if $\zeta_k = 0$, the claims are evident then from $x_{k+1} = x_k \in \mathcal{X}^*$. Without loss of generality, we assume $\zeta_k \neq 0$. To show (4.1), let $x^* \in \operatorname{proj}_{\mathcal{X}^*}(x_k)$. Then, $\operatorname{dist}(x_k; \mathcal{X}^*) = \|x_k - x^*\|$ and $f(x^*) = f^*$. Applying on the nonexpansiveness of proj_X, it holds that

$$\begin{aligned} \operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) &\leq \|x_{k+1} - x^*\|^2 \leq \|(x_k - x^*) - \alpha_k \tfrac{\zeta_k}{\|\zeta_k\|}\|^2 \\ &= \|x_k - x^*\|^2 + \tfrac{2\alpha_k}{\|\zeta_k\|}\langle \zeta_k, x^* - x_k\rangle + \alpha_k^2 \\ &\leq \operatorname{dist}^2(x_k; \mathcal{X}^*) + \tfrac{2\alpha_k}{\|\zeta_k\|}\left(f^* - f(x_k) + \rho \operatorname{dist}^{1+\nu}(x_k; \mathcal{X}^*)\right) + \alpha_k^2, \end{aligned} \tag{4.3}$$

where the last inequality comes from ν-paraconvexity of f, Theorem 3.10 (a). Now, inasmuch as $x_k \in \mathcal{T}_\gamma \subseteq \mathcal{T}_1$, we have $\frac{\alpha_k}{L} \leq \frac{\alpha_k}{\|\zeta_k\|}$ and $\operatorname{dist}(x_k; \mathcal{X}^*) < \left(\frac{\mu}{2\rho}\right)^{\frac{\delta}{\delta(1+\nu)-1}}$, i.e.,

$$\rho \operatorname{dist}^{1+\nu}(x_k; \mathcal{X}^*) < \tfrac{1}{2}\mu \operatorname{dist}^{\frac{1}{\delta}}(x_k; \mathcal{X}^*) \leq \tfrac{1}{2}(f(x_k) - f^*),$$

using the HEB. Substituting these bounds into (4.3) results in (4.1), i.e.,

$$\begin{aligned} \operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) &\leq \operatorname{dist}^2(x_k; \mathcal{X}^*) + \tfrac{2\alpha_k}{\|\zeta_k\|}\left(f^* - f(x_k) + \tfrac{1}{2}(f(x_k) - f^*)\right) + \alpha_k^2 \\ &\leq \operatorname{dist}^2(x_k; \mathcal{X}^*) - \tfrac{\alpha_k}{L}\left(f(x_k) - f^*\right) + \alpha_k^2. \end{aligned}$$

Moreover, the argument for (4.2) is derived from (4.1), relying on the HEB, i.e.,

$$\begin{aligned} \operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) &\leq \operatorname{dist}^2(x_k; \mathcal{X}^*) - \tfrac{\alpha_k}{L}\left(f(x_k) - f^*\right) + \alpha_k^2 \\ &\leq \operatorname{dist}^2(x_k; \mathcal{X}^*) - \alpha_k \tau \operatorname{dist}^{\frac{1}{\delta}}(x_k; \mathcal{X}^*) + \alpha_k^2. \end{aligned}$$

For deriving the last part, $x_{k+1} \in \mathcal{T}_\gamma$, we consider two cases: (i) $\alpha_k \leq \tau \operatorname{dist}^{\frac{1}{\delta}}(x_k; \mathcal{X}^*)$; (ii) $\alpha_k > \tau \operatorname{dist}^{\frac{1}{\delta}}(x_k; \mathcal{X}^*)$.

In Case (i), $\alpha_k \leq \tau \operatorname{dist}^{\frac{1}{\delta}}(x_k; \mathcal{X}^*)$, applying the above inequality to (4.1) yields

$$\text{dist}^2(x_{k+1};\mathcal{X}^*) \le \text{dist}^2(x_k;\mathcal{X}^*) - \alpha_k\tau\,\text{dist}^{\frac{1}{\delta}}(x_k;\mathcal{X}^*) + \alpha_k^2 \le \text{dist}^2(x_k;\mathcal{X}^*) < \left(\tfrac{\mu}{2\rho}\right)^{\frac{2\delta}{1-\delta(1+\nu)}}.$$

i.e., $x_{k+1} \in \mathcal{T}_\gamma$.

For Case (ii), $\alpha_k > \tau\,\text{dist}^{\frac{1}{\delta}}(x_k;\mathcal{X}^*)$, leading to

$$\text{dist}(x_k;\mathcal{X}^*) < \left(\tfrac{\alpha_k}{\tau}\right)^{\delta}. \tag{4.4}$$

Take the convex function $\varphi(t) := t - \alpha_k\tau t^{\frac{1}{2\delta}} + \alpha_k^2$ on $[0, (\frac{\alpha_k}{\tau})^{2\delta}]$ where the maximum is achieved at 0 or $(\frac{\alpha_k}{\tau})^{2\delta}$. On the basis of (4.1) and (4.4), $\text{dist}^2(x_{k+1};\mathcal{X}^*) \le \varphi(\text{dist}^2(x_k;\mathcal{X}^*))$ and $\text{dist}^2(x_k;\mathcal{X}^*) \in [0, (\frac{\alpha_k}{\tau})^{2\delta}]$, i.e.,

$$\begin{aligned}\text{dist}^2(x_{k+1};\mathcal{X}^*) &\le \max\left\{\varphi(t) :\ t \in [0, (\tfrac{\alpha_k}{\tau})^{2\delta}]\right\} = \max\left\{\alpha_k^2,\ (\tfrac{\alpha_k}{\tau})^{2\delta} - \alpha_k\tau\tfrac{\alpha_k}{\tau} + \alpha_k^2\right\}\\ &\le \max\left\{\alpha_k^{2\delta},\ (\tfrac{\alpha_k}{\tau})^{2\delta}\right\}\\ &\le \max\left\{1,\ \tfrac{1}{\tau^{2\delta}}\right\}\min\{1,\tau^{2\delta}\}(\tfrac{\mu}{2\rho})^{\frac{2\delta}{\delta(1+\nu)-1}} = (\tfrac{\mu}{2\rho})^{\frac{2\delta}{\delta(1+\nu)-1}},\end{aligned}$$

where the second and third inequalities are derived from the bound on α_k, i.e., $x_{k+1} \in \mathcal{T}_\gamma$. □

Next, we drive an upper bound for the difference between the lowest value of the function over the iterations and the optimal value. Let us define $f_k^* := \min\{f(x_i) : i = 1, 2, \ldots, k\}$ for all $k \ge 0$.

Lemma 4.2 (Basic inequalities II). *Let $\{x_k\}_{k\ge0} \subseteq \mathcal{T}_\gamma$ with $\gamma = 1/2$ is generated by Algorithm 1. Then, one has*

$$f(x_k) - f^* \le \frac{L\,\text{dist}^2(x_k;\mathcal{X}^*) + L\alpha_k^2}{\alpha_k},$$
$$f_k^* - f^* \le \frac{L\,\text{dist}^2(x_1;\mathcal{X}^*) + L\sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i}.$$

Proof These inequalities follow directly from Lemma 4.1. □

4.1 PSM with constant step-size

Here, we analyze a PSM with a constant step-size, $\alpha_k = \alpha > 0$ for all k. We show that, with appropriate initialization, the sequence $D_k := \text{dist}^2(x_k;\mathcal{X}^*)$ converges linearly below a certain fixed threshold.

Theorem 4.3 (Convergence of constant step-size method). *Let $\tau = \frac{\mu}{L}$. Fix a constant step-size α satisfying*

$$0 < \alpha < \min\left\{\tfrac{2\delta}{\tau}\left(\tfrac{\mu}{2\rho}\right)^{\frac{2\delta-1}{\delta(1+\nu)-1}},\ \frac{\tau}{(\tau^{2\delta}+1)^{\frac{1}{2\delta}}}\left(\tfrac{\mu}{2\rho}\right)^{\frac{1}{\delta(1+\nu)-1}},\ 1\right\}.$$

Let $\{x_k\}_{k\ge0}$ be generated by Algorithm 1 with $\gamma = 1/2$ and the constant step-size $\alpha_k = \alpha$. Define the constants

$$D_* := \left(\tfrac{\alpha}{\tau}\right)^{2\delta},\qquad q := 1 - \tfrac{\alpha\tau}{2\delta}\left(\tfrac{2\rho}{\mu}\right)^{\frac{2\delta-1}{\delta(1+\nu)-1}},\qquad \mathcal{D}^2 := \max\left\{D_0, \alpha^2 + D_*\right\}.$$

Then, $0 < q < 1$, *and for any* $k \in \mathbb{N}$, *we have*

$$\sqrt{D_k} \leq \mathcal{D} \leq \left(\frac{\mu}{2\rho}\right)^{\frac{\delta}{\delta(1+\nu)-1}}, \quad \textit{and} \quad D_k - D_* \leq \max\left\{q^k\left(D_0 - D_*\right), \alpha^2\right\}. \tag{4.5}$$

Proof We first verify claims about $\mathcal{D}$ and q, and then use the induction to establish the iteration-dependent inequalities. First, we will verify $\mathcal{D} \leq \left(\frac{\mu}{2\rho}\right)^{\frac{\delta}{\delta(1+\nu)-1}}$. Inasmuch as $x_0 \in \mathcal{T}_\gamma$, $D_0 \leq \left(\frac{\mu}{2\rho}\right)^{\frac{2\delta}{\delta(1+\nu)-1}}$ due to definition of $\mathcal{T}_\gamma$. Moreover, using the upper bound of α, we get

$$\alpha^2 + D_* \leq \alpha^{2\delta} + \frac{\alpha^{2\delta}}{\tau^{2\delta}} = \alpha^{2\delta}\frac{\tau^{2\delta}+1}{\tau^{2\delta}} \leq \left(\frac{\mu}{2\rho}\right)^{\frac{2\delta}{\delta(1+\nu)-1}},$$

i.e., $\mathcal{D} \leq \left(\frac{\mu}{2\rho}\right)^{\frac{2\delta}{\delta(1+\nu)-1}}$. Regarding to $q \in (0, 1)$, it is clear that $q < 1$. Considering

$$0 < \alpha < \frac{2\delta}{\tau}\left(\frac{\mu}{2\rho}\right)^{\frac{2\delta-1}{\delta(1+\nu)-1}},$$

and applying the basic calculus, it holds that $q > 0$. We now prove the inequalities in (4.5) using induction. At $k = 0$, the inequalities hold trivially. Now by induction, assume that for some $k \in \mathbb{N}$, the following holds (inductive hypothesis):

$$\sqrt{D_i} \leq \mathcal{D} \ \text{ and } \ D_i - D_* \leq \max\left\{q^i\left(D_0 - D_*\right), \alpha^2\right\}, \quad \forall i = 0, 1, \ldots, k.$$

We will verify these claims for $(k+1)$-th step. From Lemma 4.1, and the concavity of function $t^{\frac{1}{2\delta}}$ on $[0, +\infty)$, it follows that

$$\begin{aligned} D_{k+1} &\leq D_k - \alpha\tau D_k^{\frac{1}{2\delta}} + \alpha^2 = D_k - \alpha\tau\left(D_k^{\frac{1}{2\delta}} - \frac{\alpha}{\tau}\right) \\ &= D_k - \alpha\tau\left(D_k^{\frac{1}{2\delta}} - D_*^{\frac{1}{2\delta}}\right) \leq D_k - \frac{\alpha\tau}{2\delta D_k^{1-\frac{1}{2\delta}}}(D_k - D_*), \end{aligned}$$

i.e.,

$$D_{k+1} - D_* \leq \left(1 - \frac{\alpha\tau}{2\delta D_k^{1-\frac{1}{2\delta}}}\right)(D_k - D_*).$$

We have two cases: (i) $D_k \geq D_*$; (ii) $D_k < D_*$.

In Case (i), since $D_k \leq \mathcal{D}^2 \leq \left(\frac{\mu}{2\rho}\right)^{\frac{2\delta}{\delta(1+\nu)-1}}$, we obtain

$$\begin{aligned} D_{k+1} - D_* &\leq \left(1 - \frac{\alpha\tau}{2\delta D_k^{1-\frac{1}{2\delta}}}\right)(D_k - D_*) \leq \left(1 - \frac{\alpha\tau}{2\delta\left(\frac{\mu}{2\rho}\right)^{\frac{2\delta}{\delta(1+\nu)-1}\frac{2\delta-1}{2\delta}}}\right)(D_k - D_*) \\ &= q\left(D_k - D_*\right). \end{aligned}$$

Moreover, $\sqrt{D_{k+1}} \leq \sqrt{D_k} \leq \mathcal{D}$, inasmuch as $0 < q < 1$. In Case (ii), $D_k < D_*$, relying once again on Lemma 4.1, we get

$$D_{k+1} - D_* \leq D_k - \alpha\tau D_k^{\frac{1}{2\delta}} + \alpha^2 - D_* \leq D_k + \alpha^2 - D_* < \alpha^2.$$

In addition, $\sqrt{D_{k+1}} \leq \sqrt{\alpha^2 + D_*} \leq \mathcal{D}$. Hence, in both cases and by the induction assumption, it can be deduced

$$D_{k+1} - D_* \leq \max\left\{q\left(D_k - D_*\right), \alpha^2\right\} \leq \max\left\{q^{k+1}\left(D_0 - D_*\right), \alpha^2\right\}.$$

Since $\sqrt{D_{k+1}} \leq \mathcal{D}$, by induction, the claims hold for all $k \geq 0$. □

Theorem 4.3 establishes that the sequence D_k decrease to a value below D_* at a linear rate. Furthermore, even if D_k becomes smaller than D_*, it remains close to this threshold.

The subsequent result is a straightforward consequence of Theorem 4.3 for $\delta = 1$ under Assumption I.

Corollary 4.4 *Let $\delta = 1$ and $\tau = \mu/L$. Let $\{x_k\}_{k\geq 0}$ be generated by Algorithm 1 with $\gamma = 1/2$ and the constant step-size $\alpha_k = \alpha$ satisfying $0 < \alpha < \min\left\{\frac{\tau}{\sqrt{2}}\left(\frac{\mu}{2\rho}\right)^{1/\nu}, 1\right\}$. Then, $0 < q := 1 - \frac{\alpha\tau}{2}\left(\frac{2\rho}{\mu}\right)^{1/\nu}$ and for any $k \in \mathbb{N}$, one has*

$$D_k \leq \max\left\{D_0, \alpha^2 + D_*\right\} \leq \left(\frac{\mu}{2\rho}\right)^{2/\nu}, \quad \text{and} \quad D_k - \frac{\alpha^2}{\tau^2} \leq \max\left\{q^k\left(D_0 - \frac{\alpha^2}{\tau^2}\right), \alpha^2\right\}.$$

Proof By straightforward calculations, the result follows directly from Theorem 4.3. □

The next theorem provides an upper bound, determined by the step-size, for the distance of the lowest estimated value over iterations from the optimal value.

Theorem 4.5 (Convergence rate of constant step-size method). *Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $\gamma = 1/2$ and the constant step-size $\alpha_k = \alpha$ satisfying $0 < \alpha < \min\left\{1, \min\{1, \frac{\mu}{L}\}(\frac{\mu}{2\rho})^{\frac{1}{\delta(1+\nu)-1}}\right\}$. Then,*

$$0 \leq f_k^* - f^* \leq 2L\alpha, \qquad \forall k \geq \frac{1}{\alpha^2}\left(\frac{\mu}{2\rho}\right)^{\frac{2\delta}{\delta(1+\nu)-1}}.$$

Proof This inequality follows directly form Lemma 4.2, together with Lemma 4.1. □

4.2 PSMs with diminishing step-sizes

In the previous section, we analyzed the constant step-size scheme and established convergence results within a fixed threshold. Here, we shift our focus to achieving convergence to the optimal solution by studying the PSM with a diminishing step-size. The next result validates the convergence of $\{f_k^*\}_{k\geq 1}$ and of a subsequence of $\{x_k\}_{k\geq 1}$ in which $\{x_k\}_{k\geq 1}$ is generated by the PSM with Nonsummable Diminishing (ND) step-size satisfying

$$\alpha_k \geq 0, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty.$$

Theorem 4.6 (Convergence of ND method). *Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $\gamma = 1/2$ and ND step-size α_k satisfying $0 < \alpha_k < \min\left\{1, \min\{1, \frac{\mu}{L}\}(\frac{\mu}{2\rho})^{\frac{1}{\delta(1+\nu)-1}}\right\}$. Then,*

$$\lim_{k\to\infty} f_k^* = f^*, \qquad \liminf_{k\to\infty} f(x_k) = f^*, \quad \text{and} \quad \liminf_{k\to\infty} \operatorname{dist}(x_k; \mathcal{X}^*) = 0.$$

Moreover, if $\{x_k\}_{k\geq 1}$ is a bounded sequence, it has a convergent subsequence to some optimal solution $x^* \in \mathcal{X}^*$.

Proof According to Lemmas 4.1 and 4.2, we have

$$0 \leq f_k^* - f^* \leq \frac{L\,\mathrm{dist}^2(x_1;\mathcal{X}^*) + L\sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i}.$$

Let us consider a fixed $j \in \mathbb{N}$. For $k > j$, it holds that

$$\begin{aligned} 0 \leq \frac{\sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i} &= \frac{\sum_{i=1}^{j-1} \alpha_i^2}{\sum_{i=1}^k \alpha_i} + \frac{\sum_{i=j}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i} \\ &\leq \frac{\sum_{i=1}^{j-1} \alpha_i^2}{\sum_{i=1}^k \alpha_i} + \max_{j\leq i\leq k}\{\alpha_i\}\frac{\sum_{i=j}^k \alpha_i}{\sum_{i=1}^k \alpha_i} \\ &\leq \frac{\sum_{i=1}^{j-1} \alpha_i^2}{\sum_{i=1}^k \alpha_i} + \max_{j\leq i}\{\alpha_i\}. \end{aligned}$$

Taking the limit as $k \to \infty$, we get $0 \leq \frac{\sum_{i=1}^\infty \alpha_i^2}{\sum_{i=1}^\infty \alpha_i} \leq \max_{j\leq i}\{\alpha_i\}$, for each $j \in \mathbb{N}$. Now, taking limit as $j \to \infty$, we conclude that $\frac{\sum_{i=1}^\infty \alpha_i^2}{\sum_{i=1}^\infty \alpha_i} = 0$. On the other hand, $\frac{L\,\mathrm{dist}^2(x_1;\mathcal{X}^*)}{\sum_{i=1}^\infty \alpha_i} = 0$. Thus, $\lim_{k\to\infty} f_k^* = f^*$.

Next, we demonstrate that $\liminf_{k\to\infty} f(x_k) = f^*$. To this end, we construct a subsequence $\{x_{k_j}\}_{j\in\mathbb{N}}$ of $\{x_k\}_{k\in\mathbb{N}}$ converging to some optimal solution. If $f(x_k) = f^*$ for some $k \in \mathbb{N}$, the result is valid. Otherwise, assume $f^* < f(x_k)$, for all $k \in \mathbb{N}$. For each $k \in \mathbb{N}$, there is nondecreasing index $i_k \in \{1, 2, \ldots, k\}$ such that $f_k^* = f(x_{i_k})$. Therefore, we have $\lim_{k\to\infty} f(x_{i_k}) = f^*$. Let us define $k_1 := i_1 = 1$. Since $f^* < f(x_1)$, there exists $k \in \mathbb{N}$, such that $f(x_{i_k}) < f(x_1)$. Now, let us set $k_2 := \min\{i_k \in \mathbb{N} : f(x_{i_k}) < f(x_1)\}$, i.e., $k_1 = 1 < k_2$ and $f(x_{k_2}) < f(x_{k_1})$. Continuing this process, we construct a subsequence x_{k_j} such that

$$k_1 < k_2 < \ldots < k_j, \quad \text{and} \quad f(x_{k_j}) < f(x_{k_{j-1}}) < \ldots < f(x_{k_1}),$$

where $k_t = \min\{i_k \in \mathbb{N} : f(x_{i_k}) < f(x_{k_{t-1}})\}$ for each $1 \leq t \leq j$. Now, noticing $\lim_{k\to\infty} f(x_{i_k}) = f^* < f(x_{k_j})$, there exists $k \in \mathbb{N}$ such that $f(x_{i_k}) < f(x_{k_j})$. As such, setting $k_{j+1} := \min\{i_k \in \mathbb{N} : f(x_{i_k}) < f(x_{k_j})\}$, we have $f(x_{k_{j+1}}) < f(x_{k_j})$ and $k_j < k_{j+1}$. Otherwise, if $k_{j+1} < k_j$, $f(x_{k_{j+1}}) < f(x_{k_j}) < f(x_{k_{j-1}})$, this leads to a contradiction. Thus, $\{x_{k_j}\}_{j\in\mathbb{N}}$ is a subsequence of the sequences $\{x_k\}_{k\in\mathbb{N}}$ and $\{x_{i_k}\}_{k\in\mathbb{N}}$. Therefore, $\lim_{j\to\infty} f(x_{k_j}) = \lim_{k\to\infty} f(x_{i_k}) = f^*$ and

$$f^* \leq \liminf_{k\to\infty} f(x_k) \leq \lim_{j\to\infty} f(x_{k_j}) = f^*.$$

This implies $\liminf_{k\to\infty} f(x_k) = f^*$. Furthermore, the HEB yields $\liminf_{k\to\infty} \mathrm{dist}(x_k;\mathcal{X}^*) = 0$.

We address the final claim by assuming that $\{x_k\}_{k\in\mathbb{N}}$ is a bounded sequence. Consequently, the sequence $\{x_{k_j}\}_{j\in\mathbb{N}}$ has a convergent subsequence. Without loss of generality, we assume $x_{k_j} \to x^*$ for some $x^* \in X$, i.e., $f^* = \lim_{j\to\infty} f(x_{k_j}) = f(x^*)$. As such, $x^* \in \mathcal{X}^*$, completing the proof. □

The previous theorem established that PSMs with nonsummable step-size exhibit subsequential convergence. The following theorem extends this result by demonstrating the convergence of $\{f(x_k) - f^*\}_{k\geq 1}$ and $\{\mathrm{dist}(x_k;\mathcal{X}^*)\}_{k\geq 1}$ where $\{x_k\}_{k\geq 1}$ is generated by the

PSM when the step-size is Square-Summable but Not Summable (SSNS) satisfying the following conditions:

$$\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Theorem 4.7 (Convergence of SSNS method). *Let $\{x_k\}_{k\geq 1}$ be generated by Algorithm 1 with $\gamma = 1/2$ and the SSNS step-size satisfying $0 \leq \alpha_k < \min\left\{1, \min\{1, \frac{\mu}{L}\}(\frac{\mu}{2\rho})^{\frac{1}{\delta(1+\nu)-1}}\right\}$. Then,*

$$\lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} f_k^* = f^*, \quad \textit{and} \quad \lim_{k\to\infty} \operatorname{dist}(x_k; \mathcal{X}^*) = 0.$$

Proof Invoking Theorem 4.6, it is clear that $\liminf_{k\to\infty} f(x_k) = \lim_{k\to\infty} f_k^* = f^*$, which ensures that

$$\lim_{j\to\infty} f(x_{k_j}) = f^*, \tag{4.6}$$

for some subsequence $\{x_{k_j}\}_{j\in\mathbb{N}}$ of $\{x_k\}_{k\in\mathbb{N}}$. We first show that $\operatorname{dist}(x_k; \mathcal{X}^*)$ tends to zero. By Lemma 4.1, for each $k \in \mathbb{N}$, we have $x_k \in \mathcal{T}_\gamma$, and

$$\begin{aligned}\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) &\leq \operatorname{dist}^2(x_k; \mathcal{X}^*) - \tfrac{\alpha_k}{L}\left(f(x_k) - f^*\right) + \alpha_k^2 \leq \operatorname{dist}^2(x_{k-1}; \mathcal{X}^*) + \alpha_{k-1}^2 + \alpha_k^2 \\ &\leq \ldots \leq \operatorname{dist}^2(x_1; \mathcal{X}^*) + \sum_{i=1}^{k} \alpha_i^2 \leq \left(\tfrac{\mu}{2\rho}\right)^{\frac{2\delta}{\delta(1+\nu)-1}} + \sum_{i=1}^{\infty} \alpha_i^2. \end{aligned} \tag{4.7}$$

As such, $\operatorname{dist}(x_k; \mathcal{X}^*)$ is a bounded sequence, inasmuch as $\sum_{i=1}^{\infty} \alpha_i^2 < \infty$, i.e., $\operatorname{dist}(x_{k_j}; \mathcal{X}^*)$ has a convergent subsequence. Without loss of generality, we assume that $\operatorname{dist}(x_{k_j}; \mathcal{X}^*)$ converges to some $d \geq 0$, and we hence need to show $d = 0$. Suppose on the contrary that $d > 0$, i.e., for sufficiently large j, $\operatorname{dist}(x_{k_j}; \mathcal{X}^*) > \frac{d}{2}$. Using this inequality and the HEB, it follows that

$$\mu\left(\tfrac{d}{2}\right)^{1/\delta} < \mu \operatorname{dist}^{1/\delta}(x_{k_j}; \mathcal{X}^*) \leq f(x_{k_j}) - f^*,$$

for sufficiently large j, which contradicts (4.6), i.e., $\operatorname{dist}(x_{k_j}; \mathcal{X}^*) \to 0$ as $j \to \infty$. Let $\epsilon > 0$ be arbitrary and fixed. There exists $j_0 \in \mathbb{N}$ such that $\operatorname{dist}(x_{k_{j_0}}; \mathcal{X}^*) < \frac{\epsilon^2}{2}$ and $\sum_{i=k_{j_0}}^{k} \alpha_i^2 < \frac{\epsilon^2}{2}$, for each $k \geq k_{j_0}$. Following (4.7), for each $k \geq k_{j_0}$, yields

$$\operatorname{dist}^2(x_k; \mathcal{X}^*) \leq \operatorname{dist}^2(x_{k_{j_0}}; \mathcal{X}^*) + \textstyle\sum_{i=k_{j_0}}^{k} \alpha_i^2 < \frac{\epsilon^2}{2} + \frac{\epsilon^2}{2} = \epsilon^2.$$

Consequently, $\operatorname{dist}(x_k; \mathcal{X}^*) \to 0$. On the other hand, Using the mean value theorem [46, Theorem 3.51] it can be concluded that

$$0 \leq f(x_k) - f^* \leq L \operatorname{dist}(x_k; \mathcal{X}^*), \qquad k \geq 1,$$

confirming $f(x_k) \to f^*$ and adjusting our desired results. □

We next analyze the convergence of the PSMs with specific step-size schemes, including diminishing step-sizes and geometrically decaying step-sizes. First, we focus on the PSM with diminishing step-size of the form, $\alpha_k = \lambda(k + k_0)^{-r}$ where $\lambda, k_0 > 0$ and $0 < r < 1$. Such step-sizes are widely used in both stochastic and deterministic versions of the subgradient method. The sublinear convergence of the PSM with this step-size is established in the following result.

Theorem 4.8 (Convergence rate of diminishing method). *Let $r \in (0,1)$ and $\lambda > 0$. Set*

$$A := 2^{r\delta}\sqrt{\left(\tfrac{2\lambda}{\tau}\right)^{2\delta} + \lambda^2}, \quad \textit{and}$$
$$k_0 := \max\left\{\left(\frac{4r\delta A^{\frac{2\delta-1}{\delta}}}{\lambda\tau}\right)^{\frac{1}{1-2r(1-\delta)}}, A^{\frac{1}{\delta r}}\left(\frac{2\rho}{\mu}\right)^{\frac{1}{r(\delta(1+\nu)-1)}}, 1\right\}.$$

Let $\{x_k\}_{k\geq 0}$ be generated by Algorithm 1 with $\gamma = 1/2$ and step-size $\alpha_k = \lambda(k+k_0)^{-r}$ and let the initial point $x_0 \in \mathcal{T}_\gamma$ satisfying $\operatorname{dist}(x_0;\mathcal{X}^*) \leq \min\left\{(\frac{2\lambda}{\tau})^{\delta}, Ak_0^{-\delta r}\right\}$. Then,

$$\operatorname{dist}(x_k;\mathcal{X}^*) \leq \frac{A}{(k+k_0)^{\delta r}}. \tag{4.8}$$

Proof Clearly, $\operatorname{dist}(x_0;\mathcal{X}^*) \leq Ak_0^{-\delta r}$, by virtue of the initialization condition. Let us proceed by induction. Assume that the inequality (4.8) holds for k, $\operatorname{dist}(x_k;\mathcal{X}^*) \leq A(k+k_0)^{-\delta r}$. Noticing $k_0 \geq A^{\frac{1}{\delta r}}\left(\frac{2\rho}{\mu}\right)^{\frac{1}{r(\delta(1+\nu)-1)}}$, it follows that $x_k \in \mathcal{T}_\gamma$. From Lemma 4.1, we obtain

$$\operatorname{dist}^2(x_{k+1};\mathcal{X}^*) \leq \operatorname{dist}^2(x_k;\mathcal{X}^*) - \lambda(k+k_0)^{-r}\tau\operatorname{dist}^{1/\delta}(x_k;\mathcal{X}^*) + \lambda^2(k+k_0)^{-2r}. \tag{4.9}$$

Let us consider the index set

$$I := \left\{i \in \mathbb{N} : \tfrac{1}{2}\tau\operatorname{dist}^{1/\delta}(x_i;\mathcal{X}^*) \leq \frac{\lambda}{(i+k_0)^r}\right\}.$$

There are two possible cases for the index $k+1$: (i) $k+1 \in I$; (ii) $k+1 \notin I$. In Case (i), $k+1 \in I$, it follows that: $\operatorname{dist}(x_{k+1};\mathcal{X}^*) \leq \frac{A}{(k+1+k_0)^{\delta r}}$, since $\left(\frac{2\lambda}{\tau}\right)^{\delta} < A$. Alternatively, in Case (ii), $k+1 \notin I$, let us further consider two subcases: k: either $k \in I$ or $k \notin I$. In the first case, $k \in I$, employing the inequality (4.9), we come to

$$\begin{aligned}\operatorname{dist}^2(x_{k+1};\mathcal{X}^*) &\leq \operatorname{dist}^2(x_k;\mathcal{X}^*) + \frac{\lambda^2}{(k+k_0)^{2r}} \leq \frac{\left(\frac{2\lambda}{\tau}\right)^{2\delta}}{(k+k_0)^{2\delta r}} + \frac{\lambda^2}{(k+k_0)^{2r}} \\ &\leq \left(\left(\tfrac{2\lambda}{\tau}\right)^{2\delta} + \lambda^2\right)\frac{1}{(k+k_0)^{2\delta r}} \leq \frac{A^2}{(k+1+k_0)^{2\delta r}}.\end{aligned}$$

We now verify the second case, $k \notin I$, i.e.,

$$\tfrac{1}{2}\tau\operatorname{dist}^{1/\delta}(x_k;\mathcal{X}^*) > \frac{\lambda}{(k+k_0)^r}.$$

Substituting this into (4.9) yields

$$\operatorname{dist}^2(x_{k+1};\mathcal{X}^*) < \operatorname{dist}^2(x_k;\mathcal{X}^*) - \frac{\frac{1}{2}\lambda\tau}{(k+k_0)^r}\operatorname{dist}^{1/\delta}(x_k;\mathcal{X}^*).$$

Let us consider the convex function $\varphi : \mathbb{R} \to \mathbb{R}$ given by

$$\varphi(t) := t - \frac{\frac{1}{2}\lambda\tau}{(k+k_0)^r}t^{\frac{1}{2\delta}},$$

defined over $[0, A^2(k+k_0)^{-2\delta r}]$ where the function achieves its maximum at 0 or $A^2(k+k_0)^{-2\delta r}$. Using the inductive assumption and (4.9), we deduce

$$\operatorname{dist}^2(x_{k+1};\mathcal{X}^*) < \varphi\big(\operatorname{dist}^2(x_k;\mathcal{X}^*)\big) \leq \max\left\{\varphi(0), \varphi\Big(A^2(k+k_0)^{-2\delta r}\Big)\right\}$$

$$= \frac{A^2}{(k+k_0)^{2\delta r}} - \frac{\frac{1}{2}\lambda\tau A^{1/\delta}}{(k+k_0)^{2r}}.$$

To ensure the result, it is enough to verify

$$\frac{A^2}{(k+k_0)^{2\delta r}} - \frac{\frac{1}{2}\lambda\tau A^{1/\delta}}{(k+k_0)^{2r}} \leq \frac{A^2}{(k+1+k_0)^{2\delta r}}. \tag{4.10}$$

Using the convexity of the function $\frac{A^2}{x^{2\delta r}}$ on positive real number set, we get

$$\frac{A^2}{(k+1+k_0)^{2\delta r}} \geq \frac{A^2}{(k+k_0)^{2\delta r}} - \frac{2\delta r A^2}{(k+k_0)^{2\delta r+1}}.$$

Moreover, $\frac{2\delta r A^2}{(k+k_0)^{2\delta r+1}} \leq \frac{\frac{1}{2}\lambda\tau A^{1/\delta}}{(k+k_0)^{2r}}$ if and only if $\left(\frac{4\delta r A^{2-\frac{1}{\delta}}}{\lambda\tau}\right)^{\frac{1}{1-2r(1-\delta)}} \leq k+k_0$. Thus, (4.10) holds as

$$\left(\frac{4r\delta A^{\frac{2\delta-1}{\delta}}}{\lambda\tau}\right)^{\frac{1}{1-2r(1-\delta)}} \leq k_0,$$

giving our desired results. □

Theorem 4.8 implies that the subgradient methods projected with decreasing step size $\alpha_k = \lambda(k+k_0)^{-r}$ when $0 < r < 1$ have a convergence rate of $O(K^{-r\delta})$. However, for the case where $\frac{1}{1+\nu} < \delta < 1$ we can achieve the convergence rate of $O\left(K^{\frac{-\delta}{2(1-\delta)}}\right)$. Our next result establishes the sublinear convergence of the PSM with diminishing step-size $\alpha_k = \lambda(k+k_0)^{-r}$ with $r > 1$.

Theorem 4.9 (Convergence rate of diminishing method). *Let* $\frac{1}{1+\nu} < \delta < 1$, $r := \frac{1}{2(1-\delta)}$, $A := (\frac{8\delta r}{\tau^2})^{\delta r}$, *and* $\lambda := \frac{1}{2}\tau A^{1/\delta}$ *with* $\tau = \frac{\mu}{L}$, *let* $\{x_k\}_{k\geq 0}$ *be generated by Algorithm 1 with* $\gamma = 1/2$ *and step-size* $\alpha_k = \lambda(k+k_0)^{-r}$, *and let the initial point* $x_0 \in \mathcal{T}_\gamma$ *satisfying* $\operatorname{dist}(x_0; \mathcal{X}^*) \leq A k_0^{-\delta r}$, *with* $k_0 = \max\left\{4\delta r, A^{\frac{1}{\delta r}}(\frac{2\rho}{\mu})^{\frac{2(1-\delta)}{\delta(1+\nu)-1}}\right\}$. *Then,*

$$\operatorname{dist}(x_k; \mathcal{X}^*) \leq \frac{A}{(k+k_0)^{\delta r}}. \tag{4.11}$$

Proof By induction, we show the inequality (4.11). The result for $k = 0$ holds. We now assume that the inequality (4.11) holds for iteration k. We will show that it remains true for $k+1$. First, we observe that

$$\operatorname{dist}(x_k; \mathcal{X}^*) \leq \frac{A}{(k+k_0)^{\delta r}} \leq \frac{A}{k_0^{\delta r}} \leq \frac{A}{A(\frac{2\rho}{\mu})^{\frac{2(1-\delta)\delta r}{\delta(1+\nu)-1}}} = (\tfrac{\mu}{2\rho})^{\frac{\delta}{\delta(1+\nu)-1}},$$

i.e., $x_k \in \mathcal{T}_\gamma$. Thanks to Lemma 4.1, it holds that

$$\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) \leq \operatorname{dist}^2(x_k; \mathcal{X}^*) - \lambda(k+k_0)^{-r}\tau \operatorname{dist}^{1/\delta}(x_k; \mathcal{X}^*) + \lambda^2(k+k_0)^{-2r}. \tag{4.12}$$

We consider the convex function $\varphi(t) := t - \lambda(k+k_0)^{-r}\tau t^{\frac{1}{2\delta}} + \lambda^2(k+k_0)^{-2r}$ on $[0, A^2(k+k_0)^{-2\delta r}]$, leading to

$$\begin{aligned}\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) &\leq \varphi(\operatorname{dist}^2(x_k; \mathcal{X}^*)) \leq \max\left\{\varphi(0), \varphi(A^2(k+k_0)^{-2\delta r})\right\} \\ &= \max\left\{\lambda^2(k+k_0)^{-2r},\ A^2(k+k_0)^{-2\delta r} - \lambda\tau A^{1/\delta}(k+k_0)^{-2r} + \lambda^2(k+k_0)^{-2r}\right\},\end{aligned}$$

where (4.12) was applied in the first inequality. To establish the desired result, it suffices to show

$$\frac{\lambda^2}{(k+k_0)^{2r}} \leq \frac{A^2}{(k+1+k_0)^{2\delta r}}, \tag{4.13}$$

$$\frac{A^2}{(k+k_0)^{2\delta r}} + \frac{\lambda^2 - \lambda\tau A^{1/\delta}}{(k+k_0)^{2r}} \leq \frac{A^2}{(k+1+k_0)^{2\delta r}}. \tag{4.14}$$

Let us begin with verifying (4.13). Inasmuch as the function $\frac{x^{2r}}{(1+x)^{2\delta r}}$ is increasing on $[0, \infty)$,

$$\frac{k_0^{2r}}{(1+k_0)^{2\delta r}} \leq \frac{(k+k_0)^{2r}}{(k+1+k_0)^{2\delta r}}. \tag{4.15}$$

Using the convexity of the function $\frac{1}{x^{2\delta r}}$ on positive real numbers, we get

$$\frac{1}{(1+k_0)^{2\delta r}} \geq \frac{1}{k_0^{2\delta r}} - \frac{2\delta r}{k_0^{2\delta r+1}} = \frac{1}{k_0^{2\delta r}} - \frac{2\delta r}{k_0^{2r}}, \tag{4.16}$$

with $2\delta r + 1 = 2r$. Combining (4.15) with (4.16), it follows that

$$\frac{(k+k_0)^{2r}}{(k+1+k_0)^{2\delta r}} \geq \frac{k_0^{2r}}{(1+k_0)^{2\delta r}} \geq \frac{k_0^{2r}}{k_0^{2\delta r}} - \frac{2\delta r k_0^{2r}}{k_0^{2r}} = k_0 - 2\delta r \geq 4\delta r - 2\delta r = 2\delta r = \frac{\lambda^2}{A^2},$$

ensuring (4.13). Next, we justify the inequality (4.14). From the convexity of $\frac{A^2}{x^{2\delta r}}$, we obtain

$$\frac{A^2}{(k+1+k_0)^{2\delta r}} - \frac{A^2}{(k+k_0)^{2\delta r}} \geq \frac{-2\delta r A^2}{(k+k_0)^{2\delta r+1}} = \frac{-2\delta r A^2}{(k+k_0)^{2r}}.$$

By combining this inequality with $\lambda^2 - \lambda\tau A^{1/\delta} = -2\delta r A^2$, which follows from the definition of A, λ, and r, the inequality (4.14) is verified. □

We next focus on a Geometrically Decaying (GD) step-size to ensure linear convergence to the optimal solution set. In this case, the step-size forms a decreasing geometric sequence that tends to zero, i.e., $\alpha_k = \lambda q^k$ where $\lambda > 0$ and $0 < q < 1$. Theorem 4.10 establishes the linear convergence of the PSM with GD step-size.

Theorem 4.10 (Convergence rate of GD method). *Let* $\delta = 1$, $\tau = \frac{\mu}{L} \neq 1$ *and*

$$\max\left\{\tfrac{5\tau^2-4}{2\tau^2}, 0\right\} < \beta < \tfrac{1}{2}, \quad 0 < \lambda < \tfrac{\tau}{2}\left(\tfrac{\mu}{2\rho}\right)^{1/\nu},$$
$$q := \sqrt{1 - \tfrac{(1-2\beta)\tau^2}{4}}, \quad A := \max\left\{\tfrac{2\lambda}{\tau}, \operatorname{dist}(x_0; \mathcal{X}^*)\right\}.$$

Let $\{x_k\}_{k\geq 0}$ *be generated by Algorithm 1 with* $\gamma = 1/2$ *and step-size* $\alpha_k = \lambda q^k$ *and let the initial point* $x_0 \in \mathcal{T}_\gamma$ *satisfying*

$$\operatorname{dist}(x_0; \mathcal{X}^*) \leq \frac{2\lambda}{\tau - \sqrt{\tau^2 - 4(1-q^2)}}. \tag{4.17}$$

Then,

$$\operatorname{dist}(x_k; \mathcal{X}^*) \leq Aq^k. \tag{4.18}$$

Proof By Lemma 3.18, it follows that $\tau \in (0, 1)$. Consequently, $\frac{5\tau^2-4}{2\tau^2} < \frac{1}{2}$, which ensures the existence of β within the specified range. Additionally, it is evident that $0 < q < 1$. Observing that $1 - q^2 < \frac{\tau^2}{4}$, we deduce $\tau^2 - 4(1 - q^2) > 0$, i.e., the inequality (4.17) is well-defined. We proceed by induction to verify the inequality (4.18). For $k = 0$, we note that $\operatorname{dist}(x_0; \mathcal{X}^*) \leq A = Aq^0$. Assume that (4.18) holds for $i = 0, 1, \ldots, k$, and we show that it holds for $k + 1$. Using the upper bound of λ and the assumption $x_0 \in \mathcal{T}_\gamma$, we get

$$\operatorname{dist}(x_k; \mathcal{X}^*) \leq Aq^k \leq A = \max\left\{\tfrac{2\lambda}{\tau}, \operatorname{dist}(x_0; \mathcal{X}^*)\right\} \leq \left(\tfrac{\mu}{2\rho}\right)^{1/\nu}.$$

Hence, $x_k \in \mathcal{T}_\gamma$, i.e., applying Lemma 4.1 to the case $\delta = 1$ yields

$$\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) \leq \operatorname{dist}^2(x_k; \mathcal{X}^*) - \lambda q^k \tau \operatorname{dist}(x_k; \mathcal{X}^*) + \lambda^2 q^{2k}.$$

Let us define the convex function $\varphi(t) := t^2 - \lambda q^k \tau t + \lambda^2 q^{2k}$ on $[0, Aq^k]$, which attains its maximum at 0 or Aq^k. By the inductive assumption, $\operatorname{dist}(x_k; \mathcal{X}^*) \in [0, Aq^k]$, i.e.,

$$\begin{aligned}\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) &\leq \varphi\big(\operatorname{dist}(x_k; \mathcal{X}^*)\big) \leq \max\left\{\varphi(0), \varphi(Aq^k)\right\} \\ &= q^{2k} \max\left\{\lambda^2,\ A^2 - \lambda\tau A + \lambda^2\right\}.\end{aligned}$$

To ensure $\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) \leq A^2 q^{2(k+1)}$, it is sufficient to verify

$$\lambda^2 \leq A^2 q^2, \quad A^2 - \lambda\tau A + \lambda^2 \leq A^2 q^2. \tag{4.19}$$

For the first inequality, since $A^2 \geq \frac{4\lambda^2}{\tau^2}$ and $\tau^2 \leq q^2$, it follows that $\lambda^2 \leq A^2 q^2$, where $\tau^2 \leq q^2$ arises from the lower bound of β. To show the second inequality in (4.19), we consider the quadratic equation

$$\left(1 - q^2\right) Z^2 - \lambda\tau Z + \lambda^2 = 0, \tag{4.20}$$

where its positive roots, $z_1 \leq z_2$ are given by

$$z_{1,2} := \frac{\lambda\tau \pm \lambda\sqrt{\tau^2 - 4(1-q^2)}}{2(1-q^2)} = \frac{2\lambda}{\tau \mp \sqrt{\tau^2 - 4(1-q^2)}}.$$

To satisfy the second inequality in (4.19), it is sufficient to verify that $z_1 \leq A \leq z_2$. Regarding the smaller root, it is clear that $z_1 \leq \frac{2\lambda}{\tau} \leq A$. For the larger root, we consider two cases: (i) $A = \operatorname{dist}(x_0; \mathcal{X}^*)$; (ii) $A = \frac{2\lambda}{\tau}$. In Case (i), $A = \operatorname{dist}(x_0; \mathcal{X}^*)$, the upper bound in (4.17) and $0 < q < 1$ ensures the result. In Case (ii), $A = \frac{2\lambda}{\tau}$, it is clear that $A \leq \frac{2\lambda}{\tau - \sqrt{\tau^2 - 4(1-q^2)}} = z_2$. Therefore, the inductive step is complete, adjusting our desired result. □

4.3 PSM with Scaled Polyak's step-size

In this section, we concentrate on the PSM with the Scaled Polyak's step-size, $\alpha_k = \frac{f(x_k) - f^*}{\sigma \|\zeta_k\|}$ where $\sigma > 0$. Unlike the PSMs discussed in the previous sections, this approach requires knowing the optimal value of the problem (1.1). This requirement is reasonable in certain scenarios, such as when employing the exact penalty method to solve nonlinear equations. We

conclude this section by analyzing the convergence of the sequence generated by Algorithm 1 with a scaled version of Polyak's step-size. We will show that the sequence generated by this method attains a sublinear convergence rate if $\frac{1}{1+\nu} < \delta < 1$, and enjoys the Q-linear convergence if $\delta = 1$.

Theorem 4.11 (Convergence rate of Scaled Polyak's method). *Fix the parameters $\sigma > \frac{1}{2}$ and $0 < \gamma < \frac{2\sigma-1}{2\sigma}$. Let $\{x_k\}_{k\geq 0}$ be generated by Algorithm 1 with Scaled Polyak's step-size. Then, the following assertions hold:*

(a) For each integer $k \geq 0$, $x_k \in \mathcal{T}_\gamma$ and

$$\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) \leq \left(1 - \Big(2\sigma(1-\gamma) - 1\Big)\sigma^{-2}\tau_k^2\right) \operatorname{dist}^2(x_k; \mathcal{X}^*), \tag{4.21}$$

where $\tau_k := \tau \operatorname{dist}^{\frac{1}{\delta}-1}(x_k; \mathcal{X}^*)$ and $\tau = \frac{\mu}{L}$;

(b) The sequence $\{\operatorname{dist}(x_k; \mathcal{X}^*)\}_{k\geq 0}$ either converges to zero in a finite number of steps or converges to zero monotonically and $\lim_{k\to\infty} f(x_k) = f^*$;

(c) If $\delta = 1$, the sequence $\{\operatorname{dist}(x_k; \mathcal{X}^*)\}_{k\geq 0}$ Q-linearly converges with rate $\sqrt{1 - \Big(2\sigma(1-\gamma) - 1\Big)\sigma^{-2}\tau^2}$;

(d) If $\frac{1}{1+\nu} < \delta < 1$, the sequence $\{\operatorname{dist}(x_k; \mathcal{X}^*)\}_{k\geq 0}$ sublinearly converges, i.e., there exists $\eta > 0$ such that

$$\operatorname{dist}^2(x_k; \mathcal{X}^*) \leq \eta k^{-\frac{\delta}{1-\delta}},$$

for all $k \in \mathbb{N}$ sufficiently large;

(e) For each $k \in \mathbb{N}$,

$$f_k^* - f^* \leq \frac{\sigma L \operatorname{dist}(x_0; \mathcal{X}^*)}{\sqrt{2\sigma(1-\gamma) - 1}\sqrt{k+1}}. \tag{4.22}$$

Proof (a) By assumption it holds that $x_0 \in \mathcal{T}_\gamma$. Using induction, assume that $x_k \in \mathcal{T}_\gamma$ for some $k \geq 0$. We prove that $x_{k+1} \in \mathcal{T}_\gamma$ for $k = k+1$. Two cases are recognized: (i) $\zeta_k = 0$; (ii) $\zeta_k \neq 0$. In Case (i), $\zeta_k = 0$, it clearly $x_{k+1} = x_k \in \mathcal{T}_\gamma$. Moreover, $0 = \zeta_k \in \partial f(x_k)$ and x_k is a stationary point for problem (1.1). Hence, $x_k \in \mathcal{X}^*$ by Lemma 3.17. Thus, $\operatorname{dist}(x_{k+1}; \mathcal{X}^*) = \operatorname{dist}(x_k; \mathcal{X}^*) = 0$, and (4.21) is satisfied. In Case (ii), $\zeta_k \neq 0$, let us further consider two subcases: either $x_k \in \mathcal{X}^*$ or $x_k \notin \mathcal{X}^*$. In the case $x_k \in \mathcal{X}^*$, it follows that $f(x_k) = f^*$, $\alpha_k = 0$, $\operatorname{dist}(x_k; \mathcal{X}^*) = 0$, and $\tau_k = 0$. Then, we obtain $x_{k+1} = \operatorname{proj}_X(x_k) = x_k \in \mathcal{T}_\gamma$, and also (4.21) holds.

Now, let us examine the latter case $x_k \notin \mathcal{X}^*$. We choose $\bar{x} \in \operatorname{proj}_{\mathcal{X}^*}(x_k)$ implying $\|x_k - \bar{x}\| = \operatorname{dist}(x_k; \mathcal{X}^*)$ and $f(\bar{x}) = f^*$. Drawing on the nonexpansiveness of the projection operator and Theorem 3.10, we have

$$\begin{aligned}
\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) \leq \|x_{k+1} - \bar{x}\|^2 &\leq \operatorname{dist}^2(x_k; \mathcal{X}^*) + \frac{2\big(f(x_k) - f(\bar{x})\big)}{\sigma\|\zeta_k\|^2}\langle \zeta_k, \bar{x} - x_k\rangle + \frac{\big(f(x_k) - f(\bar{x})\big)^2}{\sigma^2\|\zeta_k\|^2} \\
&\leq \operatorname{dist}^2(x_k; \mathcal{X}^*) + \frac{2\big(f(x_k) - f(\bar{x})\big)}{\sigma\|\zeta_k\|^2}\Big(f(\bar{x}) - f(x_k) + \rho\|x_k - \bar{x}\|^{1+\nu}\Big) + \frac{\big(f(x_k) - f(\bar{x})\big)^2}{\sigma^2\|\zeta_k\|^2} \\
&= \operatorname{dist}^2(x_k; \mathcal{X}^*) + \frac{f(x_k) - f^*}{\sigma^2\|\zeta_k\|^2}\Big(2\sigma\rho \operatorname{dist}^{1+\nu}(x_k; \mathcal{X}^*) - (2\sigma - 1)(f(x_k) - f^*)\Big).
\end{aligned} \tag{4.23}$$

Since $x_k \in \mathcal{T}_\gamma \subseteq \mathcal{T}_1$, we get $\operatorname{dist}(x_k; \mathcal{X}^*) < \left(\gamma\frac{\mu}{\rho}\right)^{\frac{\delta}{\delta(1+\nu)-1}}$ and $\|\zeta_k\| \leq L$. Therefore,

$$\rho \operatorname{dist}^{1+\nu}(x_k; \mathcal{X}^*) < \gamma\mu \operatorname{dist}^{\frac{1}{\delta}}(x_k; \mathcal{X}^*) \leq \gamma(f(x_k) - f^*), \tag{4.24}$$

where the second inequality follows from the HEB. Applying these inequalities in (4.23), we deduce

$$\begin{aligned}
&\operatorname{dist}^2(x_{k+1};\mathcal{X}^*)\\
&\leq \operatorname{dist}^2(x_k;\mathcal{X}^*) + \tfrac{f(x_k)-f^*}{\sigma^2\|\zeta_k\|^2}\Big(2\sigma\gamma\mu\operatorname{dist}^{\frac{1}{\delta}}(x_k;\mathcal{X}^*) - (2\sigma-1)\mu\operatorname{dist}^{\frac{1}{\delta}}(x_k;\mathcal{X}^*)\Big)\\
&\leq \operatorname{dist}^2(x_k;\mathcal{X}^*) + \tfrac{\mu^2\operatorname{dist}^{\frac{2}{\delta}}(x_k;\mathcal{X}^*)}{\sigma^2\|\zeta_k\|^2}\Big(2\sigma\gamma - 2\sigma + 1\Big)\\
&\leq \left(1 - \Big(2\sigma(1-\gamma)-1\Big)\sigma^{-2}\tau_k^2\right)\operatorname{dist}^2(x_k;\mathcal{X}^*).
\end{aligned}$$

Hence, the inequality (4.21) is valid. Additionally, inasmuch as $x_k \in \mathcal{T}_\gamma$ and $x_k \notin \mathcal{X}^*$, we conclude $\tau_k \in (0,1]$ by Lemma 3.18, i.e., $0 < 1 - \Big(2\sigma(1-\gamma)-1\Big)\sigma^{-2}\tau_k^2 < 1$. Hence, $\operatorname{dist}(x_{k+1};\mathcal{X}^*) < \operatorname{dist}(x_k;\mathcal{X}^*)$, which yields $x_{k+1} \in \mathcal{T}_\gamma$.

(b) In order to prove the convergence of $\operatorname{dist}(x_k;\mathcal{X}^*)$, based on Assertion (a) either there exists some $k_0 \in \mathbb{N}$ such that $\operatorname{dist}(x_k;\mathcal{X}^*) = 0$ for all $k \geq k_0$, or $\operatorname{dist}(x_k;\mathcal{X}^*)$ is a decreasing positive sequence. We show that $\operatorname{dist}(x_k;\mathcal{X}^*) \to 0$ in the latter case. Suppose for contradiction that $\operatorname{dist}(x_k;\mathcal{X}^*)$ converges to some $d > 0$. Taking (4.21) into account, we get $d^2 \leq \left(1-\Big(2\sigma(1-\gamma)-1\Big)\sigma^{-2}\tau^2 d^{\frac{2}{\delta}-2}\right)d^2$, leading to a contradiction. Hence, $\operatorname{dist}(x_k;\mathcal{X}^*)$ decreasingly converges to zero. Moreover, the closedness of $\mathcal{X}^*$ implies that $x_k \to x^*$, for some $x^* \in \mathcal{X}^*$, i.e., it results in $f(x_k) \to f(x^*) = f^*$, due to continuity of f.

(c) In this assertion, Lemma 3.18 ensures that $\tau_k = \tau = \frac{\mu}{L} \in (0,1]$. It follows from inequality (4.21) that

$$\operatorname{dist}^2(x_{k+1};\mathcal{X}^*) \leq \left(1 - \Big(2\sigma(1-\gamma)-1\Big)\sigma^{-2}\tau^2\right)\operatorname{dist}^2(x_k;\mathcal{X}^*),$$

where $0 < 1 - \Big(2\sigma(1-\gamma)-1\Big)\sigma^{-2}\tau^2 < 1$. Therefore, $\operatorname{dist}(x_k;\mathcal{X}^*)$ converges Q-linearly to zero with rate $\sqrt{1 - \Big(2\sigma(1-\gamma)-1\Big)\sigma^{-2}\tau^2}$. Consequently, x_k converges $\mathcal{X}^*$.

(d) The inequality (4.21) can be rewritten as

$$\operatorname{dist}^{1/\delta}(x_k;\mathcal{X}^*) \leq \frac{\sigma^2}{\Big(2\sigma(1-\gamma)-1\Big)\tau^2}\Big(\operatorname{dist}(x_k;\mathcal{X}^*) - \operatorname{dist}(x_{k+1};\mathcal{X}^*)\Big).$$

Thus, the sublinear convergence of $\operatorname{dist}(x_k;\mathcal{X}^*)$ is guaranteed from Lemma 2.1.

(e) Combining (4.23) and (4.24) with $\|\zeta_k\| \leq L$, it follows that

$$\begin{aligned}
\operatorname{dist}^2(x_{k+1};\mathcal{X}^*) &\leq \operatorname{dist}^2(x_k;\mathcal{X}^*) + \tfrac{f(x_k)-f^*}{\sigma^2\|\zeta_k\|^2}\Big(2\sigma\gamma(f(x_k)-f^*) - (2\sigma-1)(f(x_k)-f^*)\Big)\\
&\leq \operatorname{dist}^2(x_k;\mathcal{X}^*) - \tfrac{(f(x_k)-f^*)^2}{\sigma^2L^2}\Big(2\sigma(1-\gamma)-1\Big)\\
&\leq \ldots \leq \operatorname{dist}^2(x_0;\mathcal{X}^*) - \tfrac{2\sigma(1-\gamma)-1}{\sigma^2L^2}\sum_{i=0}^{k}(f(x_i)-f^*)^2\\
&\leq \operatorname{dist}^2(x_0;\mathcal{X}^*) - \tfrac{2\sigma(1-\gamma)-1}{\sigma^2L^2}(f_k^*-f^*)^2(k+1).
\end{aligned}$$

Since $\operatorname{dist}^2(x_{k+1};\mathcal{X}^*) \geq 0$, the inequality (4.22) is obtained, adjusting our claims. □

5 Application to robust matrix recovery

This section presents a detailed analysis of our numerical experiments to demonstrate the efficiency of various projected subgradient methods employing several step-size strategies, including the following:

- Diminishing: Algorithm 1 with the step-size $\alpha_k := \frac{\alpha_0}{\sqrt{k+1}}$ and $\alpha_0 > 0$;
- Decaying: Algorithm 1 with the step-size $\alpha_k := \alpha_0 \lambda^k$ and $\alpha_0 > 0$;
- Polyak: Algorithm 1 with the step-size $\alpha_k := \frac{f(x_k)-f^*}{\|\zeta_k\|}$;
- Scaled Polyak: Algorithm 1 with the step-size $\alpha_k := \frac{f(x_k)-f^*}{4\|\zeta_k\|}$.

5.1 Robust low-rank matrix recovery

The experiments are performed on various applications of *robust low-rank matrix recoveries* using different datasets. The applications include *robust matrix completion* on the MovieLens dataset, *image inpainting*, *robust nonnegative matrix factorization for face recognition*, and *matrix compression*. The basis of all these applications is the robust matrix recovery problem

$$\min_{U,V} \|X - UV\|_1, \tag{5.1}$$

where the matrix X is approximated by UV. Here, $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{r\times n}$ represent low-rank matrices [42]. This formulation results in a nonconvex and nonsmooth paraconvex optimization problem (as confirmed in Proposition 3.11), which presents significant computational challenges. In our implementation, the code was entirely written in Python (a 3.11.7 version) and tested on a MacBook Pro, macOS Ventura 13.6.6, equipped with 32 GB of RAM.

5.1.1 Robust matrix completion problem

In data mining and machine learning, many datasets are represented as matrices, often with missing values ranging from minimal to substantial. The problem of *robust matrix completion (RMC)* [70] is formulated as

$$\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{r\times n}} \frac{1}{2}\|M \odot (X - UV)\|_1, \tag{5.2}$$

where $\odot$ stands for the Hadamard product of two matrices. Missing values can be addressed by defining a binary weight matrix M given by

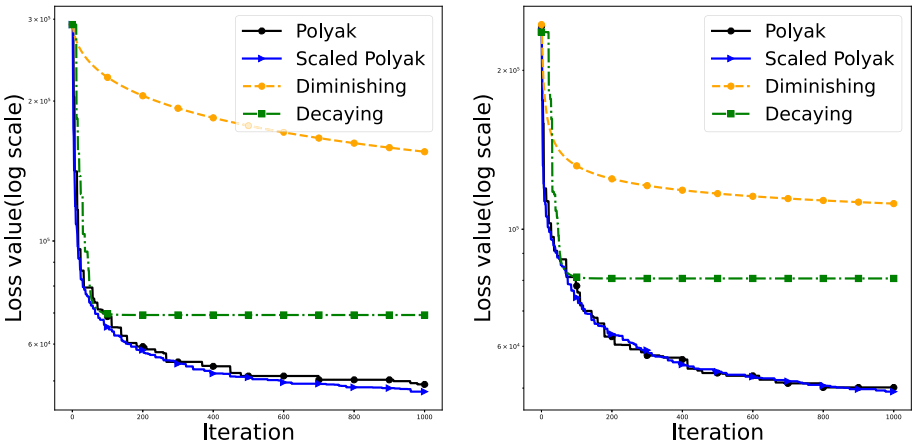
$$M_{ij} = \begin{cases} 1 & \text{if } X_{ij} \text{ is known,} \\ 0 & \text{if } X_{ij} \text{ is unknown.} \end{cases}$$

Recovering such matrices is challenging due to the sparsity of observed values, compounded by noise, corruption, and outliers.

We evaluated our method on the widely-used MovieLens (ML-100K) dataset, which contains approximately 100,000 ratings from 943 users for 1,682 movies (ratings range from 1 to 5). The dataset was split into an 80% training set and a 20% test set to evaluate reconstruction quality. The proposed algorithms are applied to the problem (5.2) with the rank $r = 20$ and a maximum of 1000 iterations. For the decaying strategy, the parameters are set as $\gamma_k = 0.95$ and $\alpha_0 = 10^{-3}$, while for the diminishing strategy, $\alpha_0 = 10^{-3}$. The reconstruction

Table 1 RMSE values for the MovieLens 100k dataset

Step-size	Polyak	Scaled Polyak	Diminishing	Decaying
svd initialization	1.460	1.254	1.655983	1.153
random initialization	1.350	1.288	1.085	1.066



(a) Loss values with SVD starting point. (b) Loss values with random starting point.

Fig. 4 Convergence of the loss function (log scale) over 1000 iterations for the MovieLens 100K dataset, comparing four step-size strategies: Polyak, scaled Polyak, diminishing, and decaying using two initial points

accuracy is measured by the *root mean squared error (RMSE)*, and the results are presented in Table 1. The convergence behavior is analyzed based on the loss values, where the results are presented in Figure 4. To demonstrate the robustness of our algorithm, we employed two different initialization methods: one based on SVD (shown in the left subfigure of Figure 4) and the other with a random starting point (shown in the right subfigure of Figure 4).

Figure 4 illustrates that the Polyak and scaled Polyak strategies achieved lower loss values compared to the decaying and diminishing strategies under both initialization methods. However, Table 1 indicates the decaying strategy achieved the best performance with an RMSE of 1.153, closely followed by Scaled Polyak with an RMSE of 1.254. Both methods indicate fast convergence and accurate predictions. In contrast, the diminishing algorithm results in the highest RMSE of 1.656.

5.1.2 Matrix compression using RNMF

Matrix compression reduces the storage of large matrices and computational demands by approximating them with compact representations that retain essential information while minimizing redundancy; cf. [78]. This process is crucial in fields such as data analysis, machine learning, and scientific computing, where efficient handling of high-dimensional data is vital. Techniques such as low-rank approximations, including nonnegative matrix factorization (NMF), are widely used. These methods decompose the original matrix into two lower-rank factors that, when combined, approximate the original data, enabling efficient

storage, faster computations, and, in some cases, enhanced interpretability of the underlying structures.

Here, we applied the projected subgradient method to the reformulated robust matrix completion (RMC) problem in (5.1), treating it as a robust nonnegative matrix factorization (RNMF) problem

$$\min_{U\geq 0\in\mathbb{R}^{m\times r}, V\geq 0\in\mathbb{R}^{r\times n}} \frac{1}{2}\|X - UV\|_1, \tag{5.3}$$

using the aforementioned step-size strategies on the 'Man' image. The goal is to compress the image by representing it with two low-rank matrices for different ranks ($r = 5, 10, 50, 100$).

The results in Figure 5 and 6 illustrate the impact of rank on reconstruction quality during matrix compression. Notably, the reconstruction quality is improved across all step-size strategies as the rank increased from 5 to 100. For lower ranks (5 and 10), the Decaying and Diminishing methods are struggling, leaving significant visual artifacts in the reconstructions. In contrast, the scaled Polyak and Polyak methods demonstrate better stability and higher-quality reconstructions, even at lower ranks. All methods improved at higher ranks (50 and 100), but the Scaled Polyak method consistently produced sharper and more accurate results. Figure 6 further highlights the convergence behavior of the step-size strategies across ranks. The Decaying method stagnates early and fails to reduce the loss significantly. The Diminishing method shows some progress, but at a slower rate. However, the Polyak and scaled Polyak methods effectively reduce the loss, with the scaled Polyak consistently achieving the lowest final loss values. These findings demonstrate that while rank significantly impacts reconstruction performance, the Polyak-based methods remain robust across different ranks, rendering them particularly effective for matrix compression tasks.

5.2 Robust image deblurring problem

We next illustrate the numerical behavior of the proposed projected subgradient methods on a robust image deblurring problem. Image deblurring can be formulated as the linear inverse problem

$$y = Ax^\star + e,$$

where $x^\star \in \mathbb{R}^n$ denotes the original image, $A \in \mathbb{R}^{n\times n}$ represents a blur operator, and $e \in \mathbb{R}^n$ corresponds to additive noise [63]. To enhance robustness against noise and outliers, we consider

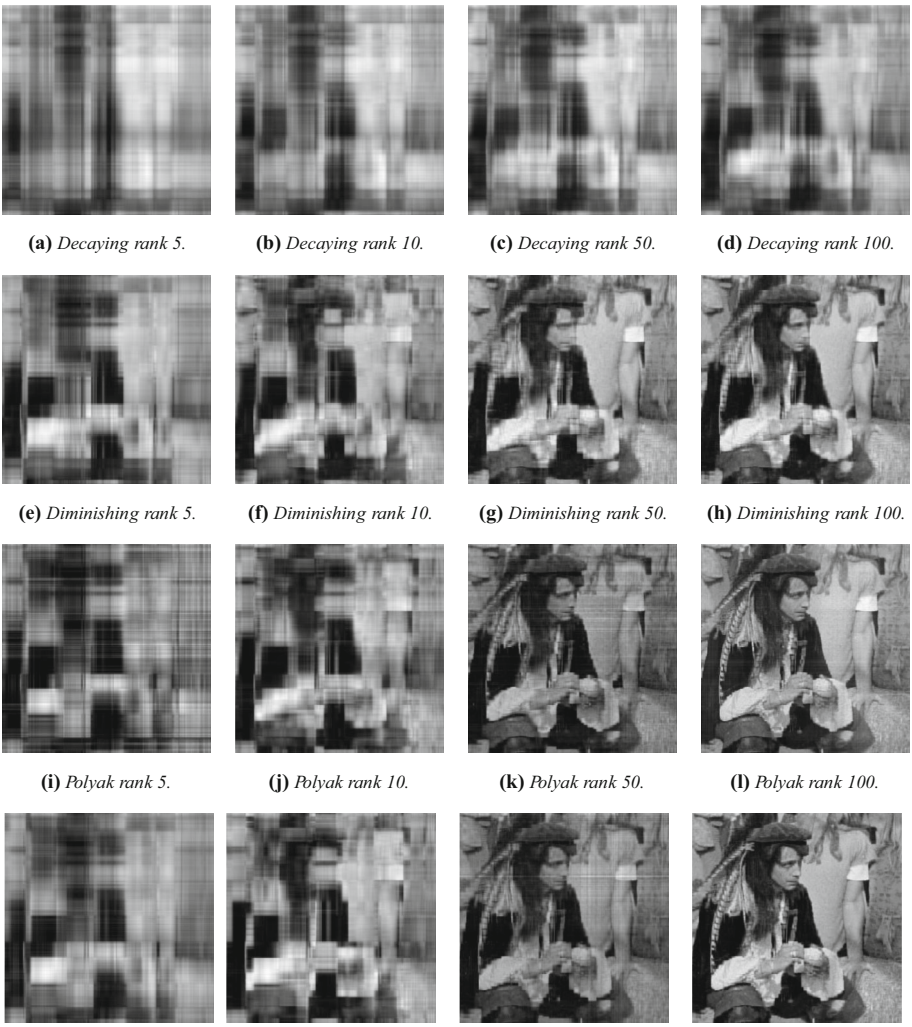
$$\min_{x\in\mathbb{R}^n} \|Ax - y\|_1 + \lambda \sum_{i=1}^{n} r(x_i), \tag{5.4}$$

where r is the GMCP function introduced in Example 3.5 (e), defined by

$$r(t) = \begin{cases} (1+\nu)|t| - |t|^{1+\nu}, & \text{if } |t| \leq 1, \\ \nu, & \text{if } |t| > 1, \end{cases}$$

with $\nu \in (0, 1]$. By Proposition 3.9 (a) and Theorem 3.12 (b), the objective function in (5.4) is ν-paraconvex and satisfies a local Hölderian error bound.

The forward operator A models spatially invariant blur and is implemented as a two-dimensional convolution with a normalized 5×5 averaging kernel using FFTs under periodic boundary conditions. The original image $x^\star$ is the standard *Cameraman* image normalized to [0, 1]. The observations are generated with a blurred signal-to-noise ratio (BSNR) of 40



(a) *Decaying rank 5.* **(b)** *Decaying rank 10.* **(c)** *Decaying rank 50.* **(d)** *Decaying rank 100.*

(e) *Diminishing rank 5.* **(f)** *Diminishing rank 10.* **(g)** *Diminishing rank 50.* **(h)** *Diminishing rank 100.*

(i) *Polyak rank 5.* **(j)** *Polyak rank 10.* **(k)** *Polyak rank 50.* **(l)** *Polyak rank 100.*

(m) *Scaled Polyak rank 5.* **(n)** *Scaled Polyak rank 10.* **(o)** *Scaled Polyak rank 50.* **(p)** *Scaled Polyak rank 100.*

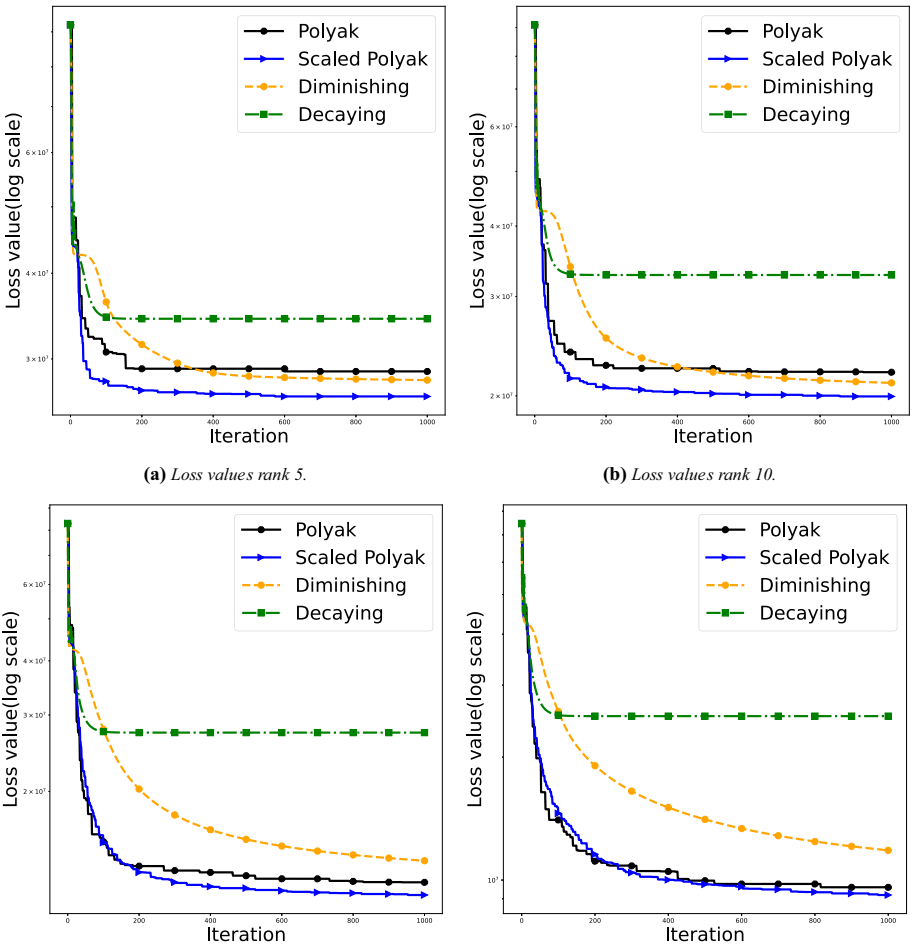
Fig. 5 Comparison of reconstructed images for different ranks ($r = 5, 10, 50, 100$) using four step-size strategies: Decaying, Diminishing, Polyak, and Scaled Polyak

dB. The parameters are set as

$$\lambda = 10^{-2}, \qquad \nu = \tfrac{1}{2},$$

and all methods are run for 150 iterations with initial point $x_0 = y$. We compare four step-size strategies: Polyak, scaled Polyak, diminishing, and geometrically decaying.

Figure 8a presents the running minimum of the objective values (log-scale). The Polyak and scaled Polyak strategies decrease the objective more rapidly during the initial iterations, while the diminishing rule exhibits slower convergence. The decaying strategy provides stable and competitive behavior. Figure 8 shows the evolution of the PSNR over the iterations. The Polyak-type strategies attain higher PSNR values, reaching approximately 29 dB after 150 iterations, whereas the diminishing rule yields lower reconstruction quality. The reconstructed



(a) *Loss values rank 5.*

(b) *Loss values rank 10.*

(c) *Loss values rank 50.*

(d) *Loss values rank 100.*

Fig. 6 Convergence plots showing the loss value (log scale) over 1000 iterations for different ranks ($r = 5, 10, 50, 100$) using the decaying, diminishing, Polyak, and scaled Polyak step-sizes

images obtained after 150 iterations are displayed in Figure 7. The Polyak and scaled Polyak strategies produce visually sharper reconstructions, while the diminishing rule results in noticeable over-smoothing.

6 Concluding remarks

In this paper, we studied the projected subgradient methods for paraconvex functions under the Hölderian error bound condition. In the first step, we investigate the properties and characterizations of paraconvex functions, including their local Lipschitz continuity. Notably, we established that ν-paraconvexity coincides with ν-weak convexity on convex and compact sets. In addition, under the Hölderian error bound condition, we identified regions around the optimal set that include no saddle points.

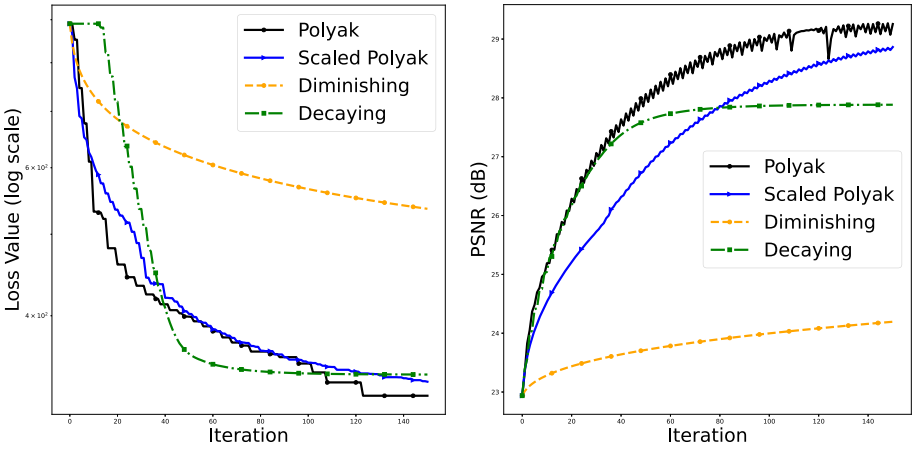


(a) *Original image x.* **(b)** *Blurred image y.* **(c)** *Decaying (PSNR* = 27.88 *dB).*



(d) *Diminishing (PSNR* = 24.19 *dB).* **(e)** *Polyak (PSNR* = 29.26 *dB).* **(f)** *Scaled Polyak (PSNR* = 28.86 *dB).*

Fig. 7 Recovered images after 150 iterations for the robust image deblurring problem



(a) *Loss values (log-scale) over 150 iterations.* **(b)** *PSNR values over 150 iterations.*

Fig. 8 Cameraman deblurring. (a) Objective value versus iteration (log scale). (b) PSNR (dB) versus iteration. Comparison of Polyak, scaled Polyak, diminishing, and decaying step-size strategies

Next, we described the generic projected subgradient method with several step-sizes (i.e., constant, nonsummable diminishing, square-summable but not summable, geometrically decaying, and Scaled Polyak's step sizes). It was demonstrated that the projected subgradient method with a constant step-size achieves linear convergence up to a fixed threshold. Furthermore, we demonstrated subsequential and global convergence for nonsummable and square-summable but not summable step-sizes, respectively, while establishing linear convergence for geometrically decaying and Scaled Polyak's step-sizes.

Finally, we assessed the practical effectiveness of the proposed methods in robust low-rank matrix recovery problems, such as robust matrix completion, matrix compression, and robust nonnegative matrix factorization, and robust image deblurring. Our preliminary numerical experiments highlighted the potential of our methods for nonsmooth and nonconvex problems that lack a suitable structure. Notably, the Scaled Polyak's step-size consistently outperformed the other considered algorithms, highlighting its practical value and potential for broader adoption in nonconvex optimization applications, e.g., solving high-order and non-Euclidean proximal-point subproblems; cf. [2, 5, 35, 36].

References

1. Ahookhosh, M.: Optimal subgradient methods: computational properties for large-scale linear inverse problems. Optim. Eng. **19**(4), 815–844 (2018)
2. Ahookhosh, M., Nesterov, Y.: High-order methods beyond the classical complexity bounds: inexact high-order proximal-point methods. Math. Program. **208**(1), 365–407 (2024)
3. Ahookhosh, M., Neumaier, A.: Optimal subgradient algorithms for large-scale convex optimization in simple domains. Numerical Algorithms **76**, 1071–1097 (2017)
4. Ahookhosh, M., Neumaier, A.: Solving structured nonsmooth convex optimization with complexity $\mathcal{O}(\varepsilon^{-1/2})$. TOP **26**, 110–145 (2018)
5. Ahookhosh, M., Themelis, A., Patrinos, P.: A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. SIAM J. Optim. **31**(1), 653–685 (2021)
6. Aragón Artacho, F.J., Fleming, R.M., Vuong, P.T.: Accelerating the dc algorithm for smooth functions. Math. Program. **169**, 95–118 (2018)
7. Bertsekas, D., Nedic, A., Ozdaglar, A.: Convex analysis and optimization, vol. 1. Athena Scientific (2003)
8. Bierstone, E., Milman, P.D.: Semianalytic and subanalytic sets. Publications Mathématiques de l'IHÉS **67**, 5–42 (1988)
9. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2007)
10. 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**(2), 471–507 (2017)
11. Burke, J., Deng, S.: Weak sharp minima revisited Part I: basic theory. Control. Cybern. **31**(3), 439–469 (2002)
12. Burke, J.V., Ferris, M.C.: Weak sharp minima in mathematical programming. SIAM J. Control. Optim. **31**(5), 1340–1359 (1993)
13. Cannarsa, P., Sinestrari, C.: Semiconcave functions, Hamilton-Jacobi equations, and optimal control, vol. 58. Springer Science & Business Media (2004)
14. Chartrand, R., Staneva, V.: Restricted isometry properties and nonconvex compressive sensing. Inverse Problems **24**(3), 035,020 (2008)
15. Chen, L., Gu, Y.: The convergence guarantees of a non-convex approach for sparse recovery. IEEE Trans. Signal Process. **62**(15), 3754–3767 (2014)
16. Chi, Y., Lu, Y.M., Chen, Y.: Nonconvex optimization meets low-rank matrix factorization: An overview. IEEE Trans. Signal Process. **67**(20), 5239–5269 (2019)
17. Clarke, F.H.: Optimization and nonsmooth analysis, SIAM (1990)
18. Daniilidis, A., Malick, J.: Filling the gap between lower-c^1 and lower-c^2 functions. J. Convex Anal. **12**(2), 315–329 (2005)
19. Davis, D., Drusvyatskiy, D., MacPhee, K.J., Paquette, C.: Subgradient methods for sharp weakly convex functions. J. Optim. Theory Appl. **179**, 962–982 (2018)

20. Davis, D., Drusvyatskiy, D., Paquette, C.: The nonsmooth landscape of phase retrieval. IMA J. Numer. Anal. **40**(4), 2652–2695 (2020)
21. Drusvyatskiy, D.: The proximal point method revisited, (2017). arXiv preprint arXiv:1712.06038
22. Duchi, J.C., Ruan, F.: Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval. Information and Inference: A Journal of the IMA **8**(3), 471–529 (2019)
23. Eldar, Y.C., Mendelson, S.: Phase retrieval: Stability and recovery guarantees. Appl. Comput. Harmon. Anal. **36**(3), 473–494 (2014)
24. Ermol'ev, Y.M.: Methods of solution of nonlinear extremal problems. Cybernetics **2**(4), 1–14 (1966)
25. Ferris, M.C.: Finite termination of the proximal point algorithm. Math. Program. **50**(1), 359–366 (1991)
26. Garrigos, G., Rosasco, L., Villa, S.: Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry. Math. Program. **198**(1), 937–996 (2023)
27. Gillis, N.: Nonnegative matrix factorization. SIAM, (2020)
28. Grimmer, B.: General hölder smooth convergence rates follow from specialized rates assuming growth bounds. J. Optim. Theory Appl. **197**(1), 51–70 (2023)
29. Hu, Y., Li, J., Yu, C.K.W.: Convergence rates of subgradient methods for quasi-convex optimization problems. Comput. Optim. Appl. **77**(1), 183–212 (2020)
30. Hu, Y., Yang, X., Sim, C.K.: Inexact subgradient methods for quasi-convex optimization problems. Eur. J. Oper. Res. **240**(2), 315–327 (2015)
31. Jain, P., Kar, P., et al.: Non-convex optimization for machine learning. Foundations and Trends® in Machine Learning **10**(3-4), 142–363 (2017)
32. Johnstone, P.R., Moulin, P.: Faster subgradient methods for functions with hölderian growth. Math. Program. **180**(1), 417–450 (2020)
33. Jourani, A.: Open mapping theorem and inversion theorem for γ-paraconvex multivalued mappings and applications. Stud. Math. **117**, 123–136 (1996)
34. Jourani, A.: Subdifferentiability and subdifferential monotonicity of-paraconvex functions. Control. Cybern. 25(4), (1996)
35. Kabgani, A., Ahookhosh, M.: Itsopt: An inexact two-level smoothing framework for nonconvex optimization via high-order moreau envelope, (2024). arXiv preprint arXiv:2410.19928
36. Kabgani, A., Ahookhosh, M.: Moreau envelope and proximal-point methods under the lens of high-order regularization. Set-Valued and Variational Analysis **33**(4), 47 (2025)
37. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16, pp. 795–811. Springer (2016)
38. Kiwiel, K.C.: Convergence and efficiency of subgradient methods for quasiconvex minimization. Math. Program. **90**, 1–25 (2001)
39. Kruger, A.Y.: Error bounds and hölder metric subregularity. Set-Valued and Variational Analysis **23**(4), 705–736 (2015)
40. Li, G.: Global error bounds for piecewise convex polynomials. Math. Program. **137**(1), 37–64 (2013)
41. Li, X., Chen, S., Deng, Z., Qu, Q., Zhu, Z., Man-Cho So, A.: Weakly convex optimization over stiefel manifold using riemannian subgradient-type methods. SIAM J. Optim. **31**(3), 1605–1634 (2021)
42. Li, X., Zhu, Z., Man-Cho So, A., Vidal, R.: Nonconvex robust low-rank matrix recovery. SIAM J. Optim. **30**(1), 660–686 (2020)
43. Liu, J., Ma, R., Zeng, X., Liu, W., Wang, M., Chen, H.: An efficient non-convex total variation approach for image deblurring and denoising. Applied Mathematics and Computation **397**, 125,977 (2021)
44. Liu, Q., Yang, C., Gu, Y., So, H.C.: Robust sparse recovery via weakly convex optimization in impulsive noise. Signal Process. **152**, 84–89 (2018)
45. Luo, Z.Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. **46**(1), 157–178 (1993)
46. Mordukhovich, B.: Variational Analysis and Generalized Differentiation I: Basic Theory. Grundlehren der mathematischen Wissenschaften. Springer, Berlin Heidelberg (2006)
47. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 221–259 (2009)
48. Nesterov, Y.: Subgradient methods for huge-scale optimization problems. Math. Program. **146**(1), 275–297 (2014)
49. Nesterov, Y.: Lectures on Convex Optimization, vol. 137, Springer (2018)
50. Nesterov, Y.: Primal subgradient methods with predefined step sizes. J. Optim. Theory Appl., 1–33 (2024)
51. Neumaier, A.: Osga: a fast subgradient algorithm with optimal complexity. Math. Program. **158**, 1–21 (2016)

52. Nurminskii, E.A.: The quasigradient method for the solving of the nonlinear programming problems. Cybernetics **9**(1), 145–150 (1973)
53. Pang, J.S.: Error bounds in mathematical programming. Math. Program. **79**(1), 299–332 (1997)
54. Polyak, B.T.: A general method for solving extremal problems. Soviet Mathematics Doklady **8**, 593–597 (1967)
55. Polyak, B.T.: Sharp minima. institute of control sciences lecture notes, moscow, ussr, 1979. In: IIASA workshop on generalized Lagrangians and their applications, IIASA, Laxenburg, Austria (1979)
56. Polyak, B.T.: Introduction to optimization. Optimization Software. Inc., Publications Division, New York (1987)
57. Quiroz, E.P., Ramirez, L.M., Oliveira, P.R.: An inexact proximal method for quasiconvex minimization. Eur. J. Oper. Res. **246**(3), 721–729 (2015)
58. Rodomanov, A., Nesterov, Y.: Smoothness parameter of power of euclidean norm. J. Optim. Theory Appl. **185**, 303–326 (2020)
59. Rolewicz, S.: On $\gamma-$paraconvex multifunctions. Math. Japonica **24**, 293–300 (1979)
60. Rolewicz, S.: On paraconvex multifunctions. Oper. Research Verf. (Methods of Oper Res) **31**, 540–546 (1979)
61. Rolewicz, S.: On $\alpha(\cdot)-$paraconvex and strongly $\alpha(\cdot)-$paraconvex functions. Control. Cybern. **29**(1), 367–377 (2000)
62. Ruder, S.: An overview of gradient descent optimization algorithms, (2016). arXiv preprint arXiv:1609.04747
63. Shen, J.J.: Deblurring images: Matrices, spectra, and filtering, (2007)
64. Shen, X., Chen, L., Gu, Y., So, H.C.: Square-root lasso with nonconvex regularization: an admm approach. IEEE Signal Process. Lett. **23**(7), 934–938 (2016)
65. Shor, N.Z.: Minimization Methods for Non-Differentiable Functions, vol. 3. Springer Series in Computational Mathematics, Springer, New York (1985)
66. Sun, R., Luo, Z.Q.: Guaranteed matrix completion via non-convex factorization. IEEE Trans. Inf. Theory **62**(11), 6535–6579 (2016)
67. Van Ngai, H., Penot, J.P.: Paraconvex functions and paraconvex sets. Stud. Math. **184**(1), 1–29 (2008)
68. Van Ngai, H., Tron, N.H., Van Vu, N., Théra, M.: Variational analysis of paraconvex multifunctions. J. Optim. Theory Appl. **193**(1), 180–218 (2022)
69. Vial, J.P.: Strong and weak convexity of sets and functions. Math. Oper. Res. **8**(2), 231–259 (1983)
70. Wang, N., Yao, T., Wang, J., Yeung, D.Y.: A probabilistic approach to robust matrix factorization. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII 12, pp. 126–139. Springer (2012)
71. Welling, M., Weber, M.: Positive tensor factorization. Pattern Recogn. Lett. **22**(12), 1255–1261 (2001)
72. Wen, F., Chu, L., Liu, P., Qiu, R.C.: A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. IEEE Access **6**, 69883–69906 (2018)
73. Wen, F., Pei, L., Yang, Y., Yu, W., Liu, P.: Efficient and robust recovery of sparse signal and image using generalized nonconvex regularization. IEEE Transactions on Computational Imaging **3**(4), 566–579 (2017)
74. Xu, Y., Lin, Q., Yang, T.: Accelerate stochastic subgradient method by leveraging local error bound, CoRR (2016). arXiv: abs/1607.01027
75. Yang, C., Shen, X., Ma, H., Chen, B., Gu, Y., So, H.C.: Weakly convex regularized robust sparse recovery methods with theoretical guarantees. IEEE Trans. Signal Process. **67**(19), 5046–5061 (2019)
76. Yashtini, M.: On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients. Optimization Letters **10**, 1361–1370 (2016)
77. Yu, Y., Peng, J., Yue, S.: A new nonconvex approach to low-rank matrix completion with application to image inpainting. Multidimension. Syst. Signal Process. **30**, 145–174 (2019)
78. Yuan, Z., Oja, E.: Projective nonnegative matrix factorization for image compression and feature extraction. In: Image Analysis: 14th Scandinavian Conference, SCIA 2005, Joensuu, Finland, June 19-22, 2005. Proceedings 14, pp. 333–342. Springer (2005)
79. Zhang, H.: New analysis of linear convergence of gradient-type methods via unifying error bound conditions. Math. Program. **180**(1), 371–416 (2020)
80. Zhou, Z., So, A.M.C.: A unified approach to error bounds for structured convex optimization problems. Math. Program. **165**, 689–728 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.