A globally convergent fast iterative shrinkage-thresholding algorithm with a new momentum factor for single and multi-objective convex optimization

Anonymous authors
Paper under double-blind review

Abstract

Convex-composite optimization, which minimizes an objective function represented by the sum of a differentiable function and a convex one, is widely used in machine learning and signal/image processing. Fast Iterative Shrinkage Thresholding Algorithm (FISTA) is a typical method for solving this problem and has a global convergence rate of $O(1/k^2)$. Recently, this has been extended to multi-objective optimization, together with the proof of the $O(1/k^2)$ global convergence rate. However, its momentum factor is classical, and the convergence of its iterates has not been proven. In this work, introducing some additional hyperparameters (a, b), we propose another accelerated proximal gradient method with a general momentum factor, which is new even for the single-objective cases. We show that our proposed method also has a global convergence rate of $O(1/k^2)$ for any (a, b), and further that the generated sequence of iterates converges to a weak Pareto solution when a is positive, an essential property for the finite-time manifold identification. Moreover, we report numerical results with various (a, b), showing that some of these choices give better results than the classical momentum factors.

1 Introduction

We consider the following convex-composite single ($m = 1$) or multi-objective ($m \geq 2$) optimization problem:

$$\min_{x \in \mathbf{R}^n} \quad F(x), \tag{1}$$

where $F \colon \mathbf{R}^n \to (\mathbf{R} \cup \{\infty\})^m$ is a vector-valued function with $F \coloneqq (F_1, \ldots, F_m)^\top$. We assume that each component $F_i \colon \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ is given by

$$F_i(x) \coloneqq f_i(x) + g_i(x) \quad \text{for all } i = 1, \ldots, m$$

with convex and continuously differentiable functions $f_i \colon \mathbf{R}^n \to \mathbf{R}, i = 1, \ldots, m$ and closed, proper and convex functions $g_i \colon \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}, i = 1, \ldots, m$, and each ∇f_i is Lipschitz continuous. As suggested in Tanabe et al. (2019), this problem involves many important classes. For example, it can express a convex-constrained problem if each g_i is the indicator function of a convex set S, i.e.,

$$\chi_S(x) \coloneqq \begin{cases} 0, & \text{if } x \in S, \\ \infty, & \text{otherwise.} \end{cases} \tag{2}$$

Multi-objective optimization has many applications in engineering Eschenauer et al. (1990), statistics Carrizosa & Frenk (1998), and machine learning (particularly multi-task learning Sener (2018); Lin et al. (2019), neural architecture search Kim et al. (2017); Dong et al. (2018); Elsken et al. (2019), and the accuracy-fairness trade-offs Liu & Vicente (2022)). In the multi-objective case, no single point minimizes all objective functions simultaneously in general. Therefore, we use the concept of *Pareto optimality*. We call a point

weakly Pareto optimal if there is no other point where the objective function values are strictly smaller. This generalizes the usual optimality for single-objective problems. In other words, single-objective problems are considered to be included in multi-objective ones. Hence, in the following, unless otherwise noted, we refer to (1) as multi-objective, including the case where $m = 1$.

One of the main strategies for multi-objective problems is the *scalarization approach* Gass & Saaty (1955); Geoffrion (1968); Zadeh (1963), which reduces the original multi-objective problem into a parameterized (or weighted) scalar-valued problem. However, it requires an *a priori* parameters (or weights) selection, which might be challenging. In fact, an example of convex bicriteria problem is provided in (Fliege et al., 2009, Section 7), where almost all choices of parameters fail, leading to unbounded scalarized problems. The meta-heuristics Gandibleux et al. (2004) is also popular, but it has no theoretical convergence properties under reasonable assumptions.

Many descent methods have been developed in recent years Fukuda & Graña Drummond (2014), overcoming those drawbacks. They decrease one or more objectives at each iteration or within some iterations, and their global convergence property can be analyzed under reasonable assumptions. For example, the steepest descent method Fliege & Svaiter (2000); Fliege et al. (2019); Désidéri (2012) converges globally to Pareto solutions for differentiable multi-objective problems. From a practical point of view, its applicability has also been reported in multi-task learning Sener (2018); Lin et al. (2019). Afterwards, the projected gradient Fukuda & Graña Drummond (2013), Newton's Fliege et al. (2009); Gonçalves et al. (2022), trust-region Carrizo et al. (2016), and conjugate gradient methods Lucambio Pérez & Prudente (2018) were also considered. Moreover, the proximal point Bonnel et al. (2005) and the inertial forward-backward methods Boţ & Grad (2018) can solve infinite-dimensional vector optimization problems.

For (1), the proximal gradient method Tanabe et al. (2019; 2023a) is effective. Using it, the merit function Tanabe et al. (2023c), which returns zero at the Pareto solutions and strictly positive values otherwise, converges to zero with rate $O(1/k)$ under reasonable assumptions. It is also shown that the generated sequence of iterates converges to a weak Pareto solution Bello Cruz et al. (2022). On the other hand, the accelerated proximal gradient method Tanabe et al. (2023b), which generalizes the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) Beck & Teboulle (2009) for convex-composite single-objective problems to multi-objective optimization, has also been considered, along with a proof of the merit function's $O(1/k^2)$ convergence rate. However, the momentum factor used there is classical ($t_1 = 1, t_{k+1} = \sqrt{t_k^2 + 1/4} + 1/2$), and the iterates' convergence is not proven.

This paper generalizes the associated factor by $t_1 = 1, t_{k+1} = \sqrt{t_k^2 - at_k + b} + 1/2$ with hyperparameters $a \in [0, 1), b \in [a^2/4, 1/4]$. This is new even in the single-objective context, and it generalizes well-known factors. For example, when $a = 0$ and $b = 1/4$, it reduces to $t_1 = 1, t_{k+1} = \sqrt{t_k^2 + 1/4} + 1/2$, proposed in Nesterov (1983); Beck & Teboulle (2009), and when $b = a^2/4$, it gives $t_k = (1 - a)k/2 + (1 + a)/2$, suggested in Chambolle & Dossal (2015); Attouch & Peypouquet (2016); Attouch et al. (2018); Su et al. (2016). We show that the merit function converges to zero with rate $O(1/k^2)$ for any (a, b). In addition, we prove the iterates' convergence to a weak Pareto solution when $a > 0$. While the generalization of the momentum factor is an important aspect of our work, it is crucial to emphasize that our primary contribution lies in addressing the challenge of ensuring convergence in accelerated gradient methods for multi-objective optimization. As discussed in Section 4, this suggests that the proposed method might achieve finite-iteration manifold (active set) identification Sun et al. (2019) without the assumption of strong convexity.

Furthermore, we carry out numerical experiments with various (a, b) and observe that some (a, b) yield better results than the classical factors. However, it is important to note that the primary focus and significance of our work lies in the theoretical advancement of convergence assurance rather than in the specific properties of the new momentum factor.

The outline of this paper is as follows. We present some notations and definitions used in this paper in Section 2.1. Section 2.2 recalls the accelerated proximal gradient method for (1) and its associated results. We generalize the momentum factor and prove that it preserves an $O(1/k^2)$ convergence rate in Section 3, and we demonstrate the convergence of the iterates in Section 4. Finally, Section 5 provides numerical experiments and compares the numerical performances depending on the hyperparameters.

2 Preliminaries

2.1 Definitions and notations

For every natural number d, write the d-dimensional real space by $\mathbf{R}^d$, and define

$$\mathbf{R}^d_+ \coloneqq \left\{ v \in \mathbf{R}^d \;\middle|\; v_i \geq 0, i = 1, \ldots, d \right\}.$$

This induces the partial orders: for any $v^1, v^2 \in \mathbf{R}^d$, $v^1 \leq v^2$ (alternatively, $v^2 \geq v^1$) if $v^2 - v^1 \in \mathbf{R}^d_+$ and $v^1 < v^2$ (alternatively, $v^2 > v^1$) if $v^2 - v^1 \in \operatorname{int} \mathbf{R}^d_+$. In other words, $v^1 \leq v^2$ and $v^1 < v^2$ mean that $v^1_i \leq v^2_i$ and $v^1_i < v^2_i$ for all $i = 1, \ldots, d$, respectively. Furthermore, let $\langle \cdot, \cdot \rangle$ be the Euclidean inner product in $\mathbf{R}^d$, i.e., $\langle v^1, v^2 \rangle \coloneqq \sum_{i=1}^d v^1_i v^2_i$, and let $\|\cdot\|$ be the Euclidean norm, i.e., $\|v\| \coloneqq \sqrt{\langle v, v \rangle}$. Moreover, we define the ℓ_1-norm and ℓ_∞-norm by $\|v\|_1 \coloneqq \sum_{i=1}^m |v_i|$ and $\|v\|_\infty \coloneqq \max_{i=1,\ldots,d} |v_i|$, respectively.

We introduce some concepts used in the problem (1). Recall that $x^* \in \mathbf{R}^n$ is *Pareto optimal* if there is no $x \in \mathbf{R}^n$ such that $F(x) \leq F(x^*)$ and $F(x) \neq F(x^*)$. Likewise,

$$X^* \coloneqq \{x^* \in \mathbf{R}^n \mid \text{ There does not exist } x \in \mathbf{R}^n \text{ such that } F(x) < F(x^*)\} \tag{3}$$

is the set of *weakly Pareto optimal* solutions for (1). When $m = 1$, X^* reduces to the optimal solution set. It is known that all Pareto optimal points are weakly Pareto optimal, and the converse is true if every objective function is strictly convex. Moreover, define the effective domain of F by

$$\operatorname{dom} F \coloneqq \{x \in \mathbf{R}^n \mid F(x) < \infty\},$$

and write the level set of F on $c \in \mathbf{R}^m$ as

$$\mathcal{L}_F(c) \coloneqq \{x \in \mathbf{R}^n \mid F(x) \leq c\}. \tag{4}$$

Furthermore, we express the image of $A \subseteq \mathbf{R}^n$ and the inverse image of $B \subseteq (\mathbf{R} \cup \{\infty\})^m$ under F as

$$F(A) \coloneqq \{F(x) \in \mathbf{R}^m \mid x \in A\} \quad \text{and} \quad F^{-1}(B) \coloneqq \{x \in \mathbf{R}^n \mid F(x) \in B\},$$

respectively.

Finally, let us recall the merit function $u_0 \colon \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ for (1) proposed in Tanabe et al. (2023c):

$$u_0(x) \coloneqq \sup_{z \in \mathbf{R}^n} \min_{i=1,\ldots,m} [F_i(x) - F_i(z)], \tag{5}$$

which returns zero at optimal solutions and strictly positive values otherwise. The following theorem shows that u_0 is a merit function in the Pareto sense.

Theorem 2.1
(Tanabe et al., 2023c, Theorem 3.1) Let u_0 be defined by (5). Then, $u_0(x) \geq 0$ for all $x \in \mathbf{R}^n$. Moreover, $x \in \mathbf{R}^n$ is weakly Pareto optimal for (1) if and only if $u_0(x) = 0$.

Note that when $m = 1$, we have

$$u_0(x) = F_1(x) - F_1^*,$$

where F_1^* is the optimal objective value. Clearly, this is a merit function for scalar-valued optimization.

2.2 The accelerated proximal gradient method for multi-objective optimization

This subsection recalls the accelerated proximal gradient method for (1) proposed in Tanabe et al. (2023b) and its main results. Recall that each F_i is the sum of a continuously differentiable function f_i and a closed, proper, and convex function g_i, and that ∇f_i is Lipschitz continuous with Lipschitz constant $L_i > 0$. Define

$$L \coloneqq \max_{i=1,\ldots,m} L_i.$$

The method solves the following subproblem at each iteration for given $x \in \operatorname{dom} F$, $y \in \mathbf{R}^n$, and $\ell \geq L$:

$$\min_{z \in \mathbf{R}^n} \quad \varphi_\ell^{\mathrm{acc}}(z; x, y), \tag{6}$$

where

$$\varphi_\ell^{\mathrm{acc}}(z; x, y) \coloneqq \max_{i=1,\ldots,m} [\langle \nabla f_i(y), z - y \rangle + g_i(z) + f_i(y) - F_i(x)] + \frac{\ell}{2}\|z - y\|^2.$$

From the strong convexity, (6) has a unique optimal solution $p_\ell^{\mathrm{acc}}(x, y)$, i.e.,

$$p_\ell^{\mathrm{acc}}(x, y) \coloneqq \operatorname*{argmin}_{z \in \mathbf{R}^n} \varphi_\ell^{\mathrm{acc}}(z; x, y). \tag{7}$$

The following proposition characterizes weak Pareto optimality in terms of the mapping p_ℓ^{acc}.

Proposition 2.1
(Tanabe et al., 2023b, Proposition 4.1 (i)) Let $p_\ell^{\mathrm{acc}}(x, y)$ be defined by (7). Then, $y \in \mathbf{R}^n$ is weakly Pareto optimal for (1) if and only if $p_\ell^{\mathrm{acc}}(x, y) = y$ for some $x \in \mathbf{R}^n$.

This implies that using $\|p_\ell^{\mathrm{acc}}(x, y) - y\|_\infty < \varepsilon$ for some $\varepsilon > 0$ is reasonable as the stopping criteria. We state below the accelerated proximal gradient method for (1).

Algorithm 1 Accelerated proximal gradient method for (1)

Input: Set $x^0 = y^1 \in \operatorname{dom} F, \ell \geq L, \varepsilon > 0$.
Output: x^*: A weakly Pareto optimal point

```
k ← 1
t_1 ← 1
while ‖p_ℓ^acc(x^{k-1}, y^k) − y^k‖_∞ ≥ ε do
    x^k ← p_ℓ^acc(x^{k-1}, y^k)
    t_{k+1} ← √(t_k^2 + 1/4) + 1/2
    γ_k ← (t_k − 1)/t_{k+1}
    y^{k+1} ← x^k + γ_k(x^k − x^{k-1})
    k ← k + 1
end while
```

Algorithm 1 generates $\{x^k\}$ such that $\{u_0(x^k)\}$ converges to zero with rate $O(1/k^2)$ under the following assumption. This assumption is also used to analyze the proximal gradient method without acceleration Tanabe et al. (2023a) and is not particularly strong as suggested in (Tanabe et al., 2023a, Remark 5.3); it is satisfied for level bounded functions such as ℓ_1-norm, for example.

Assumption 2.1
(Tanabe et al., 2023a, Assumption 5.1) Let X^* and $\mathcal{L}_F$ be defined by (3) and (4), respectively. Then, for all $x \in \mathcal{L}_F(F(x^0))$, there exists $x^* \in X^*$ such that $F(x^*) \leq F(x)$ and

$$R \coloneqq \sup_{F^* \in F(X^* \cap \mathcal{L}_F(F(x^0))} \inf_{z \in F^{-1}(\{F^*\})} \left\|z - x^0\right\|^2 < \infty. \tag{8}$$

Theorem 2.2
(Tanabe et al., 2023b, Theorem 5.2) Under Assumption 2.1, Algorithm 1 generates $\{x^k\}$ such that

$$u_0(x^k) \leq \frac{2\ell R}{(k+1)^2} \quad \text{for all } k \geq 1,$$

where $R \geq 0$ is given by (8), and u_0 is a merit function defined by (5).

The following corollary shows the global convergence of Algorithm 1.

Corollary 2.1
(Tanabe et al., 2023b, Corollary 5.2) Suppose that Assumption 2.1 holds. Then, every accumulation point of $\{x^k\}$ generated by Algorithm 1 is weakly Pareto optimal for (1).

3 Generalization of the momentum factor and convergence rate analysis

This section generalizes the momentum factor $\{t_k\}$ used in Algorithm 1 and shows that the $O(1/k^2)$ convergence rate also holds in that case. First, we describe below the algorithm in which we replace line 5 of Algorithm 1 by a formula using given constants $a \in [0, 1)$ and $b \in [a^2/4, 1/4]$:

Algorithm 2 Accelerated proximal gradient method with general stepsizes for (1)

Input: Set $x^0 = y^1 \in \operatorname{dom} F, \ell \geq L, \varepsilon > 0, a \in [0, 1), b \in [a^2/4, 1/4]$.
Output: x^*: A weakly Pareto optimal point

1: $k \leftarrow 1$
2: $t_1 \leftarrow 1$
3: **while** $\left\| p_\ell^{\mathrm{acc}}(x^{k-1}, y^k) - y^k \right\|_\infty \geq \varepsilon$ **do**
4: $\quad x^k \leftarrow p_\ell^{\mathrm{acc}}(x^{k-1}, y^k)$
5: $\quad t_{k+1} \leftarrow \sqrt{t_k^2 - a t_k + b} + 1/2$
6: $\quad \gamma_k \leftarrow (t_k - 1)/t_{k+1}$
7: $\quad y^{k+1} \leftarrow x^k + \gamma_k (x^k - x^{k-1})$
8: $\quad k \leftarrow k + 1$
9: **end while**

The sequence $\{t_k\}$ defined in lines 2 and 5 of Algorithm 2 generalizes the well-known momentum factors in single-objective accelerated methods. For example, when $a = 0$ and $b = 1/4$, they coincide with the one in Algorithm 1 and the original FISTA Nesterov (1983); Beck & Teboulle (2009) ($t_1 = 1$ and $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$). Moreover, if $b = a^2/4$, then $\{t_k\}$ has the general term $t_k = (1 - a)k/2 + (1 + a)/2$, which corresponds to the one used in Chambolle & Dossal (2015); Su et al. (2016); Attouch & Peypouquet (2016); Attouch et al. (2018). This means that our generalization allows a finer tuning of the algorithm by varying a and b.

We present below the main theorem of this section.

Theorem 3.1
Let $\{x^k\}$ be a sequence generated by Algorithm 2 and recall that u_0 is given by (5). Then, the following two equations hold:

(i) $F_i(x^k) \leq F_i(x^0)$ *for all* $i = 1, \ldots, m$ *and* $k \geq 0$*;*

(ii) $u_0(x^k) = O(1/k^2)$ *as* $k \to \infty$ *under Assumption 2.1.*

Claim (i) means that $\{x^k\} \subseteq \mathcal{L}_F(F(x^0))$, where $\mathcal{L}_F$ denotes the level set of F (cf. (4)). Note, however, that the objective functions are generally not monotonically non-increasing. Claim (ii) also claims the global convergence rate.

Before proving Theorem 3.1, let us give several lemmas. First, we present some properties of $\{t_k\}$ and $\{\gamma_k\}$.

Lemma 3.1
Let $\{t_k\}$ and $\{\gamma_k\}$ be defined by lines 2, 5 and 6 in Algorithm 2 for arbitrary $a \in [0, 1)$ and $b \in [a^2/4, 1/4]$. Then, the following inequalities hold for all $k \geq 1$.

(i) $t_{k+1} \geq t_k + \dfrac{1-a}{2}$ *and* $t_k \geq \dfrac{1-a}{2}k + \dfrac{1+a}{2}$*;*

(ii) $t_{k+1} \leq t_k + \dfrac{1 - a + \sqrt{4b - a^2}}{2}$ *and* $t_k \leq \dfrac{1 - a + \sqrt{4b - a^2}}{2}(k-1) + 1 \leq k$*;*

(iii) $t_k^2 - t_{k+1}^2 + t_{k+1} = a t_k - b + \dfrac{1}{4} \geq a t_k$*;*

(iv) $0 \leq \gamma_k \leq \dfrac{k-1}{k+1/2}$*;*

(v) $1 - \gamma_k^2 \geq \dfrac{1}{t_k}$.

Proof. Claim (i): From the definition of $\{t_k\}$, we have

$$t_{k+1} = \sqrt{t_k^2 - at_k + b} + \frac{1}{2} = \sqrt{\left(t_k - \frac{a}{2}\right)^2 + \left(b - \frac{a^2}{4}\right)} + \frac{1}{2}. \tag{9}$$

Since $b \geq a^2/4$, we get

$$t_{k+1} \geq \left|t_k - \frac{a}{2}\right| + \frac{1}{2}.$$

Since $t_1 = 1 \geq a/2$, we can quickly see that $t_k \geq a/2$ for any k by induction. Thus, we have

$$t_{k+1} \geq t_k + \frac{1-a}{2}.$$

Applying the above inequality recursively, we obtain

$$t_k \geq \frac{1-a}{2}(k-1) + t_1 = \frac{1-a}{2}k + \frac{1+a}{2}.$$

Claim (ii): From (9) and the relation $\sqrt{\alpha+\beta} \leq \sqrt{\alpha} + \sqrt{\beta}$ with $\alpha, \beta \geq 0$, we get the first inequality. Using it recursively, it follows that

$$t_k \leq \frac{1 - a + \sqrt{4b - a^2}}{2}(k-1) + t_1 = \frac{1 - a + \sqrt{4b - a^2}}{2}(k-1) + 1.$$

Since $a \in [0, 1), b \in [a^2/4, 1/4]$, we observe that

$$\frac{1 - a + \sqrt{4b - a^2}}{2} \leq \frac{1 - a + \sqrt{1 - a^2}}{2} \leq 1.$$

Hence, the above two inequalities lead to the desired result.

Claim (iii): An easy computation shows that

$$\begin{aligned} t_k^2 - t_{k+1}^2 + t_{k+1} &= t_k^2 - \left[\sqrt{t_k^2 - at_k + b} + \frac{1}{2}\right]^2 + \sqrt{t_k^2 - at_k + b} + \frac{1}{2} \\ &= at_k - b + \frac{1}{4} \geq at_k, \end{aligned}$$

where the inequality holds since $b \leq 1/4$.

Claim (iv): The first inequlity is clear from the definition of γ_k since claim (i) yields $t_k \geq 1$. Again, the definition of γ_k and claim (i) give

$$\gamma_k = \frac{t_k - 1}{t_{k+1}} \leq \frac{t_k - 1}{t_k + (1-a)/2} = 1 - \frac{3-a}{2t_k + 1 - a}.$$

Combining with claim (ii), we get

$$\begin{aligned} \gamma_k &\leq 1 - \frac{3-a}{\left(1 - a + \sqrt{4b - a^2}\right)(k-1) + 3 - a} \\ &= \frac{\left(1 - a + \sqrt{4b - a^2}\right)(k-1)}{\left(1 - a + \sqrt{4b - a^2}\right)(k-1) + 3 - a} \\ &= \frac{k-1}{k - 1 + (3-a)/\left(1 - a + \sqrt{4b - a^2}\right)}. \end{aligned} \tag{10}$$

On the other hand, it follows that

$$\min_{a\in[0,1),b\in[a^2/4,1/4]} \frac{3-a}{1-a+\sqrt{4b-a^2}} = \min_{a\in[0,1)} \frac{3-a}{1-a+\sqrt{1-a^2}} = \frac{3}{2}, \tag{11}$$

where the second equality follows from the monotonic non-decreasing property implied by

$$\frac{\mathrm{d}}{\mathrm{d}a}\left(\frac{3-a}{1-a+\sqrt{1-a^2}}\right) = \frac{2\sqrt{1-a^2}+3a-1}{\left(\sqrt{1-a^2}-a+1\right)^2\sqrt{1-a^2}} > 0 \quad \text{for all } a\in[0,1).$$

Combining (10) and (11), we obtain $\gamma_k \le (k-1)/(k+1/2)$.

Claim (v): claim (i) implies that $t_{k+1} > t_k \ge 1$. Thus, the definition of γ_k implies that

$$1-\gamma_k^2 = 1-\left(\frac{t_k-1}{t_{k+1}}\right)^2 \ge 1-\left(\frac{t_k-1}{t_k}\right)^2 = \frac{2t_k-1}{t_k^2} \ge \frac{2t_k-t_k}{t_k^2} = \frac{1}{t_k}.$$

□

As in Tanabe et al. (2023b), we also introduce $\sigma_k\colon \mathbf{R}^n \to \mathbf{R}\cup\{-\infty\}$ and $\rho_k\colon \mathbf{R}^n \to \mathbf{R}$ for $k \ge 0$ as follows, which assist the analysis:

$$\begin{aligned} \sigma_k(z) &\coloneqq \min_{i=1,\dots,m}\left[F_i(x^k)-F_i(z)\right], \\ \rho_k(z) &\coloneqq \left\|t_{k+1}x^{k+1}-(t_{k+1}-1)x^k-z\right\|^2. \end{aligned} \tag{12}$$

The following lemma on σ_k is helpful in the subsequent discussions.

Lemma 3.2

(Tanabe et al., 2023b, Lemma 5.1) Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 2. Then, the following inequalities hold for all $z\in\mathbf{R}^n$ and $k\ge 0$:

(i) $\sigma_{k+1}(z) \le -\frac{\ell}{2}\left(2\langle x^{k+1}-y^{k+1}, y^{k+1}-z\rangle + \left\|x^{k+1}-y^{k+1}\right\|^2\right) - \frac{\ell-L}{2}\left\|x^{k+1}-y^{k+1}\right\|^2;$

(ii) $\sigma_k(z)-\sigma_{k+1}(z) \ge \frac{\ell}{2}\left(2\langle x^{k+1}-y^{k+1}, y^{k+1}-x^k\rangle + \left\|x^{k+1}-y^{k+1}\right\|^2\right) + \frac{\ell-L}{2}\left\|x^{k+1}-y^{k+1}\right\|^2.$

Therefore, from Lemma 3.1 (v), we can obtain the following result quickly in the same way as in the proof of (Tanabe et al., 2023b, Corollary 5.1).

Lemma 3.3

Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 2. Then, we have

$$\sigma_{k_1}(z)-\sigma_{k_2}(z) \ge \frac{\ell}{2}\left(\left\|x^{k_2}-x^{k_2-1}\right\|^2 - \left\|x^{k_1}-x^{k_1-1}\right\|^2 + \sum_{k=k_1}^{k_2-1}\frac{1}{t_k}\left\|x^k-x^{k-1}\right\|^2\right)$$

for any $k_2 \ge k_1 \ge 1$.

We can now show the first part of Theorem 3.1.

Theorem 3.1 (i). From Lemma 3.3, we can prove this part with similar arguments used in the proof of (Tanabe et al., 2023b, Theorem 5.1). □

The next step is to prepare the proof of Theorem 3.1 (ii). First, we mention the following relation, used frequently hereafter:

$$\left\|v^2-v^1\right\|^2 + 2\langle v^2-v^1, v^1-v^3\rangle = \left\|v^2-v^3\right\|^2 - \left\|v^1-v^3\right\|^2, \tag{13}$$

$$\sum_{s=1}^{r}\sum_{p=1}^{s} A_p = \sum_{p=1}^{r}\sum_{s=p}^{r} A_p \tag{14}$$

for any vectors v^1, v^2, v^3 and sequence $\{A_p\}$. With these, we show the lemma below, which is similar to (Tanabe et al., 2023b, Lemma 5.2) but more complex due to the generalization of $\{t_k\}$.

Lemma 3.4
Let $\{x^k\}$ and $\{y^k\}$ be sequences generated by Algorithm 2. Also, let σ_k and ρ_k be defined by (12). Then, we have

$$\begin{aligned}
&\frac{\ell}{2}\|x^0 - z\|^2 \\
&\quad \geq \frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4} - b\right)k\right]\sigma_{k+1}(z) \\
&\qquad + \frac{\ell}{2(1-a)}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right]\|x^{k+1} - x^k\|^2 \\
&\qquad + \frac{\ell}{2(1-a)}\sum_{p=1}^{k}\left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right]\|x^p - x^{p-1}\|^2 \\
&\qquad + \frac{\ell}{2}\rho_k(z) + \frac{\ell - L}{2}\sum_{p=1}^{k} t_{p+1}^2\|x^{p+1} - y^{p+1}\|^2
\end{aligned}$$

for all $k \geq 0$ and $z \in \mathbf{R}^n$.

Proof. Let $p \geq 1$ and $z \in \mathbf{R}^n$. Recall that Lemma 3.2 gives

$$\begin{aligned}
-\sigma_{p+1}(z) &\geq \frac{\ell}{2}\left[2\langle x^{p+1} - y^{p+1}, y^{p+1} - z\rangle + \|x^{p+1} - y^{p+1}\|^2\right] \\
&\qquad + \frac{\ell - L}{2}\|x^{p+1} - y^{p+1}\|^2, \\
\sigma_p(z) - \sigma_{p+1}(z) &\geq \frac{\ell}{2}\left[2\langle x^{p+1} - y^{p+1}, y^{p+1} - x^p\rangle + \|x^{p+1} - y^{p+1}\|^2\right] \\
&\qquad + \frac{\ell - L}{2}\|x^{p+1} - y^{p+1}\|^2.
\end{aligned}$$

We then multiply the second inequality above by $(t_{p+1} - 1)$ and add it to the first one:

$$\begin{aligned}
&(t_{p+1} - 1)\sigma_p(z) - t_{p+1}\sigma_{p+1}(z) \\
&\quad \geq \frac{\ell}{2}\left[t_{p+1}\|x^{p+1} - y^{p+1}\|^2 + 2\langle x^{p+1} - y^{p+1}, t_{p+1}y^{p+1} - (t_{p+1} - 1)x^p - z\rangle\right] \\
&\qquad + \frac{\ell - L}{2}t_{p+1}\|x^{p+1} - y^{p+1}\|^2.
\end{aligned}$$

Multiplying this inequality by t_{p+1} and using the relation $t_p^2 = t_{p+1}^2 - t_{p+1} + (at_p - b + 1/4)$ (cf. Lemma 3.1 (iii)), we get

$$\begin{aligned}
t_p^2\sigma_p(z) - t_{p+1}^2\sigma_{p+1}(z) &\geq \frac{\ell}{2}\Big[\|t_{p+1}(x^{p+1} - y^{p+1})\|^2 \\
&\qquad + 2t_{p+1}\langle x^{p+1} - y^{p+1}, t_{p+1}y^{p+1} - (t_{p+1} - 1)x^p - z\rangle\Big] \\
&\qquad + \frac{\ell - L}{2}t_{p+1}^2\|x^{p+1} - y^{p+1}\|^2 + \left(at_p - b + \frac{1}{4}\right)\sigma_p(z).
\end{aligned}$$

Applying (13) to the right-hand side of the last inequality with

$$v^1 := t_{p+1}y^{p+1}, \quad v^2 := t_{p+1}x^{p+1}, \quad v^3 := (t_{p+1} - 1)x^p + z.$$

we get

$$\begin{aligned}
&t_p^2\sigma_p(z) - t_{p+1}^2\sigma_{p+1}(z) \\
&\quad \geq \frac{\ell}{2}\left[\|t_{p+1}x^{p+1} - (t_{p+1} - 1)x^p - z\|^2 - \|t_{p+1}y^{p+1} - (t_{p+1} - 1)x^p - z\|^2\right] \\
&\qquad + \frac{\ell - L}{2}t_{p+1}^2\|x^{p+1} - y^{p+1}\|^2 + \left(at_p - b + \frac{1}{4}\right)\sigma_p(z).
\end{aligned}$$

Recall that $\rho_p(z) \coloneqq \left\|t_{p+1}x^{p+1} - (t_{p+1}-1)x^p - z\right\|^2$. Then, considering the definition of y^p given in line 7 of Algorithm 2, we obtain

$$t_p^2\sigma_p(z) - t_{p+1}^2\sigma_{p+1}(z) \geq \frac{\ell}{2}\left[\rho_p(z) - \rho_{p-1}(z)\right] + \frac{\ell - L}{2}t_{p+1}^2\left\|x^{p+1} - y^{p+1}\right\|^2 + \left(at_p - b + \frac{1}{4}\right)\sigma_p(z).$$

Now, let $k \geq 0$. Lemma 3.3 with $(k_1, k_2) = (p, k+1)$ implies

$$\begin{aligned} t_p^2\sigma_p(z) - t_{p+1}^2\sigma_{p+1}(z) \geq{}& \frac{\ell}{2}\left[\rho_p(z) - \rho_{p-1}(z)\right] + \frac{\ell - L}{2}t_{p+1}^2\left\|x^{p+1} - y^{p+1}\right\|^2 \\ &+ \left(at_p - b + \frac{1}{4}\right)\left[\sigma_{k+1}(z) + \frac{\ell}{2}\left(\left\|x^{k+1} - x^k\right\|^2 - \left\|x^p - x^{p-1}\right\|^2 + \sum_{r=p}^{k}\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|^2\right)\right]. \end{aligned}$$

Adding up the above inequality from $p = 1$ to $p = k$, the fact that $t_1 = 1$ and $\rho_0(z) = \left\|x^1 - z\right\|^2$ leads to

$$\begin{aligned} \sigma_1(z) - t_{k+1}^2\sigma_{k+1}(z) \geq{}& \frac{\ell}{2}\left[\rho_k(z) - \left\|x^1 - z\right\|^2\right] + \frac{\ell - L}{2}\sum_{p=1}^{k}t_{p+1}^2\left\|x^{p+1} - y^{p+1}\right\|^2 \\ &+ \left(a\sum_{p=1}^{k}t_p + \left(\frac{1}{4} - b\right)k\right)\left[\sigma_{k+1}(z) + \frac{\ell}{2}\left\|x^{k+1} - x^k\right\|^2\right] \\ &- \frac{\ell}{2}\sum_{p=1}^{k}\left(at_p - b + \frac{1}{4}\right)\left\|x^p - x^{p+1}\right\|^2 \\ &+ \frac{\ell}{2}\sum_{p=1}^{k}\left(at_p - b + \frac{1}{4}\right)\sum_{r=p}^{k}\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|^2. \end{aligned} \tag{15}$$

Let us write the last two terms of the right-hand side for (15) as S_1 and S_2, respectively. Equation (14) yields

$$\begin{aligned} S_2 &= \frac{\ell}{2}\sum_{r=1}^{k}\sum_{p=1}^{r}\left(at_p - b + \frac{1}{4}\right)\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|^2 \\ &= \frac{\ell}{2}\sum_{p=1}^{k}\sum_{r=1}^{p}\left(at_r - b + \frac{1}{4}\right)\frac{1}{t_p}\left\|x^p - x^{p-1}\right\|^2. \end{aligned}$$

Hence, it follows that

$$\begin{aligned} S_1 + S_2 &= \frac{\ell}{2}\sum_{p=1}^{k}\left[\frac{1}{t_p}\sum_{r=1}^{p}\left(at_r - b + \frac{1}{4}\right) - \left(at_p - b + \frac{1}{4}\right)\right]\left\|x^p - x^{p-1}\right\|^2 \\ &= \frac{\ell}{2}\sum_{p=1}^{k}\frac{1}{t_p}\left[a\left(\sum_{r=1}^{p-1}t_r - t_p^2 + t_p\right) + \left(\frac{1}{4} - b\right)(p - t_p)\right]\left\|x^p - x^{p-1}\right\|^2. \end{aligned} \tag{16}$$

Again $t_1 = 1$ gives

$$\begin{aligned} -t_p^2 + t_p &= \sum_{r=1}^{p-1}(-t_{r+1}^2 + t_{r+1} + t_r^2 - t_r) = \sum_{r=1}^{p-1}\left(-(1-a)t_r - b + \frac{1}{4}\right) \\ &= -(1-a)\sum_{r=1}^{p-1}t_r + \left(\frac{1}{4} - b\right)(p-1), \end{aligned}$$

where the second equality comes from Lemma 3.1 (iii). Thus, we get

$$\sum_{r=1}^{p-1} t_r = \frac{t_p^2 - t_p}{1-a} + \left(\frac{1}{4} - b\right)\frac{p-1}{1-a}. \tag{17}$$

Substituting this into (16), it follows that

$$S_1 + S_2 = \frac{\ell}{2(1-a)} \sum_{p=1}^{k} \left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right] \left\|x^p - x^{p-1}\right\|^2.$$

Combined with (15) and (17), we have

$$\begin{aligned}
&\sigma_1(z) - t_{k+1}^2 \sigma_{k+1}(z) \\
&\geq \frac{\ell}{2}\left[\rho_k(z) - \left\|x^1 - z\right\|^2\right] + \frac{\ell - L}{2}\sum_{p=1}^{k} t_{p+1}^2 \left\|x^{k+1} - y^{k+1}\right\|^2 \\
&\quad + \frac{1}{1-a}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right]\left[\sigma_{k+1}(z) + \frac{\ell}{2}\left\|x^{k+1} - x^k\right\|^2\right] \\
&\quad + \frac{\ell}{2(1-a)}\sum_{p=1}^{k}\left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right]\left\|x^p - x^{p-1}\right\|^2.
\end{aligned}$$

Easy calculations give

$$\begin{aligned}
&\sigma_1(z) + \frac{\ell}{2}\left\|x^1 - z\right\|^2 \\
&\geq \frac{1}{1-a}\left[t_{k+1}^2 - a t_{k+1} + \left(\frac{1}{4} - b\right)k\right]\sigma_{k+1}(z) \\
&\quad + \frac{\ell}{2(1-a)}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right]\left\|x^{k+1} - x^k\right\|^2 \\
&\quad + \frac{\ell}{2(1-a)}\sum_{p=1}^{k}\left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right]\left\|x^p - x^{p-1}\right\|^2 \\
&\quad + \frac{\ell}{2}\rho_k(z) + \frac{\ell - L}{2}\sum_{p=1}^{k} t_{p+1}^2 \left\|x^{k+1} - y^{k+1}\right\|^2.
\end{aligned}$$

Lemma 3.2 (i) with $k = 0$ and $y^1 = x^0$ and (13) with $(v^1, v^2, v^3) = (x^0, x^1, z)$ lead to

$$\sigma_1(z) \leq -\frac{\ell}{2}\left[\left\|x^1 - z\right\|^2 - \left\|x^0 - z\right\|^2\right] - \frac{\ell - L}{2}\left\|x^1 - y^1\right\|^2.$$

From the above two inequalities and the fact that $\ell \geq L$, we can derive the desired inequality. □

Let us define the linear function $P\colon \mathbf{R} \to \mathbf{R}$ and quadratic ones $Q_1\colon \mathbf{R} \to \mathbf{R}$, $Q_2\colon \mathbf{R} \to \mathbf{R}$, and $Q_3\colon \mathbf{R} \to \mathbf{R}$ by

$$\begin{aligned}
P(\alpha) &\coloneqq \frac{a^2(\alpha - 1)}{2}, \\
Q_1(\alpha) &\coloneqq \frac{1-a}{4}\alpha^2 + \left[1 - \frac{a}{2} + \frac{1 - 4b}{4(1-a)}\right]\alpha + 1, \\
Q_2(\alpha) &\coloneqq \frac{a(1-a)}{4}\alpha^2 + \left[\frac{a}{2} + \frac{1 - 4b}{4(1-a)}\right]\alpha, \\
Q_3(\alpha) &\coloneqq \left(\frac{1-a}{2}\alpha + 1\right)^2.
\end{aligned} \tag{18}$$

The following lemma provides the key relation to evaluate the convergence rate of Algorithm 2.

Lemma 3.5
Under Assumption 2.1, Algorithm 2 generates a sequence $\{x^k\}$ such that

$$\frac{\ell R}{2} \geq Q_1(k)u_0(x^{k+1}) + \frac{\ell}{2}Q_2(k)\left\|x^{k+1} - x^k\right\|^2 + \frac{\ell}{2}\sum_{p=1}^{k} P(p)\left\|x^p - x^{p-1}\right\|^2 + \frac{\ell - L}{2}\sum_{p=1}^{k} Q_3(p)\left\|x^{p+1} - y^{p+1}\right\|^2$$

for all $k \geq 0$, where $R \geq 0$ and $P, Q_1, Q_2, Q_3 \colon \mathbf{R} \to \mathbf{R}$ are given in (8) and (18), respectively, and u_0 is a merit function defined by (5).

Proof. Let $k \geq 0$. With similar arguments used in the proof of Theorem 2.2 (see (Tanabe et al., 2023b, Theorem 5.2)), we get

$$\sup_{F^* \in F(X^* \cap \mathcal{L}_F(F(x^0)))} \inf_{z \in F^{-1}(\{F^*\})} \sigma_{k+1}(z) = u_0(x^{k+1}).$$

Since $\rho_k(z) \geq 0$, Lemma 3.4 and the above equality lead to

$$\begin{aligned}
\frac{\ell R}{2} \geq{}& \frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4} - b\right)k\right] u_0(x^{k+1}) \\
&+ \frac{\ell}{2(1-a)}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right]\left\|x^{k+1} - x^k\right\|^2 \\
&+ \frac{\ell}{2(1-a)}\sum_{p=1}^{k}\left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right]\left\|x^p - x^{p-1}\right\|^2 \\
&+ \frac{\ell - L}{2}\sum_{p=1}^{k} t_{p+1}^2\left\|x^{p+1} - y^{p+1}\right\|^2.
\end{aligned}$$

We now show that the coefficients of the four terms on the right-hand side can be bounded from below by the polynomials given in (18). First, by using the relation

$$t_{k+1} \geq \frac{1-a}{2}k + 1 \tag{19}$$

obtained from Lemma 3.1 (i) and $a \in [0, 1)$, we have

$$\begin{aligned}
\frac{1}{1-a}\left[t_{k+1}^2 - at_{k+1} + \left(\frac{1}{4} - b\right)k\right] &= \frac{1}{1-a}\left[t_{k+1}(t_{k+1} - a) + \left(\frac{1}{4} - b\right)k\right] \\
&\geq \frac{1}{1-a}\left[\left(\frac{1-a}{2}k + 1\right)\left(\frac{1-a}{2}k + 1 - a\right) + \left(\frac{1}{4} - b\right)k\right] = Q_1(k).
\end{aligned}$$

Again, (19) gives

$$\begin{aligned}
\frac{1}{1-a}\left[a(t_{k+1}^2 - t_{k+1}) + \left(\frac{1}{4} - b\right)k\right] &= \frac{a}{1-a}t_{k+1}(t_{k+1} - 1) + \frac{1-4b}{4(1-a)}k \\
&\geq \frac{a}{1-a}\left(\frac{1-a}{2}k + 1\right)\left(\frac{1-a}{2}k\right) + \frac{1-4b}{4(1-a)}k = Q_2(k).
\end{aligned}$$

Moreover, since $t_p \leq p$ (cf. Lemma 3.1 (ii)), $t_k \geq 1$ (cf. Lemma 3.1 (i)), and $b \in (a^2/4, 1/4]$, we obtain

$$\frac{1}{1-a}\left[a^2(t_p - 1) + \left(\frac{1}{4} - b\right)\frac{p - t_p + a(t_p - 1)}{t_p}\right] \geq \frac{a^2}{1-a}(t_p - 1) \geq P(p).$$

It is also clear from (19) that

$$t_{p+1}^2 \geq Q_3(p).$$

Thus, combining the above five inequalities, we get the desired inequality. □

Then, we can finally prove the main theorem.

Theorem 3.1 (ii). It is clear from Lemma 3.5 and $Q_1(k) = O(k^2)$ as $k \to \infty$. □

Remark 3.1
Lemma 3.5 also implies the following other claims than Theorem 3.1 (ii):

- $O(1/k^2)$ *convergence rate of* $\left\{\left\|x^{k+1} - x^k\right\|^2\right\}$ *when* $a > 0$*;*
- *the absolute convergence of* $\left\{k\left\|x^{k+1} - x^k\right\|^2\right\}$ *when* $a > 0$*;*
- *the absolute convergence of* $\left\{k^2\left\|x^k - y^k\right\|^2\right\}$ *when* $\ell > L$*.*

Note that the second one generalize (Chambolle & Dossal, 2015, Corollary 3.2) for single-objective problems.

4 Convergence of the iterates

While the last section shows that Algorithm 2 has an $O(1/k^2)$ convergence rate like Algorithm 1, this section proves the following theorem, which is more strict than Corollary 2.1 related to Algorithm 1:

Theorem 4.1
Let $\left\{x^k\right\}$ *be generated by Algorithm 2 with* $a > 0$*. Then, under Assumption 2.1, the following two properties hold:*

(i) $\left\{x^k\right\}$ *is bounded, and it has an accumulation point;*

(ii) $\left\{x^k\right\}$ *converges to a weak Pareto optimum for (1).*

The latter claim is also significant in application. For example, finite-time manifold (active set) identification, which detects the low-dimensional manifold where the optimal solution belongs, essentially requires only the convergence of the generated sequence to a unique point rather than the strong convexity of the objective functions Sun et al. (2019).

Again, we will prove Theorem 4.1 after showing some lemmas. First, we mention the following result, obvious from Assumption 2.1 and Theorem 3.1 (i).

Lemma 4.1
Let $\left\{x^k\right\}$ *be generated by Algorithm 2. Then, for any* $k \geq 0$*, there exists* $z \in X^* \cap \mathcal{L}_F(F(x^0))$ *(see (3) and (4) for the definitions of* X^* *and* $\mathcal{L}_F$*) such that*

$$\sigma_k(z) \geq 0 \quad \text{and} \quad \left\|z - x^0\right\|^2 \leq R,$$

where $R \geq 0$ *is given by (8).*

The following lemma also contributes strongly to the proof of the main theorem.

Lemma 4.2
Let $\{\gamma_q\}$ *be defined by line 6 in Algorithm 2. Then, we have*

$$\sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q \leq 2(s-1) \quad \text{for all } s, r \geq 1.$$

Proof. By using Lemma 3.1 (iv), we see that

$$\prod_{q=s}^{p} \gamma_q \leq \prod_{q=s}^{p} \frac{q-1}{q+1/2}.$$

Let Γ and B denote the gamma and beta functions defined by

$$\Gamma(\alpha) := \int_0^\infty \tau^{\alpha-1}\exp(-\tau)\,\mathrm{d}\tau \quad \text{and} \quad \mathrm{B}(\alpha,\beta) := \int_0^1 \tau^{\alpha-1}(1-\tau)^{\beta-1}\,\mathrm{d}\tau, \tag{20}$$

respectively. Applying the well-known properties:

$$\Gamma(\alpha) = (\alpha-1)!, \quad \Gamma(\alpha+1) = \alpha\Gamma(\alpha), \quad \text{and} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{21}$$

we get

$$\prod_{q=s}^{p}\gamma_q \le \frac{\Gamma(p)/\Gamma(s-1)}{\Gamma(p+3/2)/\Gamma(s+1/2)} = \frac{B(p,3/2)}{B(s-1,3/2)}.$$

This implies

$$\sum_{p=s}^{r}\prod_{q=s}^{p}\gamma_q \le \sum_{p=1}^{r} B(p,3/2)/B(s-1,3/2).$$

Then, it follows from the definition (20) of B that

$$\begin{aligned}
\sum_{p=s}^{r}\prod_{q=s}^{p}\gamma_q &\le \sum_{p=s}^{r}\int_0^1 \tau^{p-1}(1-\tau)^{1/2}\,\mathrm{d}\tau/B(s-1,3/2)\\
&= \int_0^1 \sum_{p=s}^{r}\tau^{p-1}(1-\tau)^{1/2}\,\mathrm{d}\tau/B(s-1,3/2)\\
&= \int_0^1 \frac{\tau^{s-1}-\tau^{r}}{1-\tau}(1-\tau)^{1/2}\,\mathrm{d}\tau/B(s-1,3/2)\\
&= \frac{B(s,1/2)-B(r+1,1/2)}{B(s-1,3/2)} \le \frac{B(s,1/2)}{B(s-1,3/2)}.
\end{aligned}$$

Using again (21), we conclude that

$$\sum_{p=s}^{r}\prod_{q=s}^{p}\gamma_q \le \frac{\Gamma(s)\Gamma(1/2)/\Gamma(s+1/2)}{\Gamma(s-1)\Gamma(3/2)/\Gamma(s+1/2)} = 2(s-1).$$

□

Now, we introduce two functions $\omega_k\colon \mathbf{R}^n \to \mathbf{R}$ and $\nu_k\colon \mathbf{R}^n \to \mathbf{R}$ for any $k \ge 1$, which will help our analysis, by

$$\omega_k(z) := \max\left(0, \left\|x^k - z\right\|^2 - \left\|x^{k-1} - z\right\|^2\right), \tag{22}$$

$$\nu_k(z) := \left\|x^k - z\right\|^2 - \sum_{s=1}^{k}\omega_s(z). \tag{23}$$

The lemma below describes the properties of ω_k and ν_k.

Lemma 4.3

Let $\{x^k\}$ be generated by Algorithm 2 and recall that $X^*, \mathcal{L}_F, \omega_k$, and ν_k are defined by (3), (4), (22) and (23), respectively. Moreover, suppose that Assumption 2.1 holds and that $z \in X^* \cap \mathcal{L}_F(F(x^0))$ satisfies the statement of Lemma 4.1 for some $k \ge 1$. Then, it follows for all $r = 1, \ldots, k$ that

(i) $\displaystyle\sum_{s=1}^{r}\omega_s(z) \le \sum_{s=1}^{r}(6s-5)\left\|x^s - x^{s-1}\right\|^2;$

(ii) $\nu_{r+1}(z) \le \nu_r(z).$

Proof. Claim (i): Let $k \geq p \geq 1$. From the definition of y^{p+1} given in line 7 of Algorithm 2, we have

$$\begin{aligned}
&\left\|x^{p+1} - z\right\|^2 - \left\|x^p - z\right\|^2 \\
&= -\left\|x^{p+1} - x^p\right\|^2 + 2\left\langle x^{p+1} - y^{p+1}, x^{p+1} - z\right\rangle + 2\gamma_p\left\langle x^p - x^{p-1}, x^{p+1} - z\right\rangle \\
&= -\left\|x^{p+1} - x^p\right\|^2 + 2\left\langle x^{p+1} - y^{p+1}, y^{p+1} - z\right\rangle + 2\left\|x^{p+1} - y^{p+1}\right\|^2 \\
&\quad + 2\gamma_p\left\langle x^p - x^{p-1}, x^{p+1} - z\right\rangle.
\end{aligned}$$

On the other hand, Lemma 3.2 (i) gives

$$2\left\langle x^{p+1} - y^{p+1}, y^{p+1} - z\right\rangle \leq -\frac{2}{\ell}\sigma_{p+1}(z) - \frac{2\ell - L}{\ell}\left\|x^{p+1} - y^{p+1}\right\|^2.$$

Moreover, Lemma 3.3 with $(k_1, k_2) = (p+1, k+1)$ implies

$$\begin{aligned}
-\frac{2}{\ell}\sigma_{p+1}(z) &\leq -\frac{2}{\ell}\sigma_{k+1}(z) - \left\|x^{k+1} - x^k\right\|^2 + \left\|x^{p+1} - x^p\right\|^2 - \sum_{r=p+1}^{k}\frac{1}{t_r}\left\|x^r - x^{r-1}\right\|^2 \\
&\leq \left\|x^{p+1} - x^p\right\|^2,
\end{aligned}$$

where the second inequality comes from the assumption on z. Combining the above three inequalities, we get

$$\begin{aligned}
&\left\|x^{p+1} - z\right\|^2 - \left\|x^p - z\right\|^2 \leq \frac{L}{\ell}\left\|x^{p+1} - y^{p+1}\right\|^2 + 2\gamma_p\left\langle x^p - x^{p-1}, x^{p+1} - z\right\rangle \\
&= \frac{L}{\ell}\left\|x^{p+1} - y^{p+1}\right\|^2 + \gamma_p\Big(\left\|x^p - z\right\|^2 - \left\|x^{p-1} - z\right\|^2 + \left\|x^p - x^{p-1}\right\|^2 + 2\left\langle x^p - x^{p-1}, x^{p+1} - x^p\right\rangle\Big).
\end{aligned}$$

Using the relation $\left\|x^{p+1} - y^{p+1}\right\|^2 + 2\gamma_p\left\langle x^p - x^{p-1}, x^{p+1} - x^p\right\rangle = \left\|x^{p+1} - x^p\right\|^2 + \gamma_p^2\left\|x^p - x^{p-1}\right\|^2$, which holds from the definition of y^k, we have

$$\begin{aligned}
\left\|x^{p+1} - z\right\|^2 - \left\|x^p - z\right\|^2 \leq -\frac{\ell - L}{\ell}\left\|x^{p+1} - y^{p+1}\right\|^2 + \left\|x^{p+1} - x^p\right\|^2 \\
+ \gamma_p\left(\left\|x^p - z\right\|^2 - \left\|x^{p-1} - z\right\|^2\right) + (\gamma_p + \gamma_p^2)\left\|x^p - x^{p-1}\right\|^2.
\end{aligned}$$

Since $0 \leq \gamma_p \leq 1$ from Lemma 3.1 (iv) and $\ell \geq L$, we obtain

$$\begin{aligned}
\left\|x^{p+1} - z\right\|^2 - \left\|x^p - z\right\|^2 &\leq \gamma_p\left(\left\|x^p - z\right\|^2 - \left\|x^{p-1} - z\right\|^2 + 2\left\|x^p - x^{p-1}\right\|^2\right) + \left\|x^{p+1} - x^p\right\|^2 \\
&\leq \gamma_p\left(\omega_p(z) + 2\left\|x^p - x^{p-1}\right\|^2\right) + \left\|x^{p+1} - x^p\right\|^2,
\end{aligned}$$

where the second inequality follows from the definition (22) of ω_p. Since the right-hand side is nonnegative, (22) again gives

$$\omega_{p+1}(z) \leq \gamma_p\left(\omega_p(z) + 2\left\|x^p - x^{p-1}\right\|^2\right) + \left\|x^{p+1} - x^p\right\|^2.$$

Let $s \leq k$. Applying the above inequality recursively and using $\gamma_1 = 0$, we get

$$\begin{aligned}
\omega_s(z) &\leq 3\sum_{p=2}^{s}\prod_{q=p}^{s}\gamma_q\left\|x^p - x^{p-1}\right\|^2 + 2\prod_{q=1}^{s}\gamma_q\left\|x^1 - x^0\right\|^2 + \left\|x^s - x^{s-1}\right\|^2 \\
&\leq 3\sum_{p=2}^{s}\prod_{q=p}^{s}\gamma_q\left\|x^p - x^{p-1}\right\|^2 + \left\|x^s - x^{s-1}\right\|^2.
\end{aligned}$$

Adding up the above inequality from $s = 1$ to $s = r \leq k$, we have

$$\begin{aligned}\sum_{s=1}^{r} \omega_s(z) &\leq 3 \sum_{s=1}^{r} \sum_{p=1}^{s} \prod_{q=p}^{s} \gamma_q \left\|x^p - x^{p-1}\right\|^2 + \sum_{s=1}^{r} \left\|x^s - x^{s-1}\right\|^2 \\ &= 3 \sum_{p=1}^{r} \sum_{s=p}^{r} \prod_{q=p}^{s} \gamma_q \left\|x^p - x^{p-1}\right\|^2 + \sum_{s=1}^{r} \left\|x^s - x^{s-1}\right\|^2 \\ &= \sum_{s=1}^{r} \left(3 \sum_{p=s}^{r} \prod_{q=s}^{p} \gamma_q + 1 \right) \left\|x^s - x^{s-1}\right\|^2,\end{aligned}$$

where the first equality follows from (14). Thus, Lemma 4.2 implies

$$\sum_{s=1}^{r} \omega_s(z) \leq \sum_{s=1}^{r} (6s-5) \left\|x^s - x^{s-1}\right\|^2.$$

Claim (ii): Equation (23) yields

$$\begin{aligned}\nu_{r+1}(z) &= \left\|x^{r+1} - z\right\|^2 - \omega_{r+1}(z) - \sum_{s=1}^{r} \omega_s(z) \\ &= \left\|x^{r+1} - z\right\|^2 - \max\left(0, \left\|x^{r+1} - z\right\|^2 - \left\|x^r - z\right\|^2\right) - \sum_{s=1}^{r} \omega_s(z) \\ &\leq \left\|x^{r+1} - z\right\|^2 - \left(\left\|x^{r+1} - z\right\|^2 - \left\|x^r - z\right\|^2\right) - \sum_{s=1}^{r} \omega_s(z) \\ &= \left\|x^r - z\right\|^2 - \sum_{s=1}^{r} \omega_s(z) = \nu_r(z),\end{aligned}$$

where the second and third equalities come from the definitions (22) and (23) of ω_{r+1} and ν_r, respectively. □

Let us now prove the first part of the main theorem.

Theorem 4.1 (i). Let $k \geq 1$ and suppose that $z \in X^* \cap \mathcal{L}_F(F(x^0))$ satisfies the statement of Lemma 4.1, where X^* and $\mathcal{L}_F$ are given by (3) and (4), respectively. Then, Lemma 4.3 (ii) gives

$$\begin{aligned}\nu_k(z) \leq \nu_1(z) &= \left\|x^1 - z\right\|^2 - \omega_1(z) \\ &= \left\|x^1 - z\right\|^2 - \max\left(0, \left\|x^1 - z\right\|^2 - \left\|x^0 - z\right\|^2\right) \\ &\leq \left\|x^1 - z\right\|^2 - \left(\left\|x^1 - z\right\|^2 - \left\|x^0 - z\right\|^2\right) = \left\|x^0 - z\right\|^2,\end{aligned}$$

where the second equality follows from the definition (22) of ω_1. Considering the definition (23) of ν_k, we obtain

$$\left\|x^k - z\right\|^2 \leq \left\|x^0 - z\right\|^2 + \sum_{s=1}^{k} \omega_s(z).$$

Taking the square root of both sides and using (22), we get

$$\left\|x^k - z\right\| \leq \sqrt{\left\|x^0 - z\right\|^2 + \sum_{s=1}^{k} (6s-5) \left\|x^s - x^{s-1}\right\|^2}.$$

Applying the reverse triangle inequality $\|x^k - x^0\| - \|x^0 - z\| \leq \|x^k - z\|$ to the left-hand side leads to

$$\begin{aligned}\|x^k - x^0\| &\leq \|x^0 - z\| + \sqrt{\|x^0 - z\|^2 + \sum_{s=1}^{k}(6s-5)\|x^s - x^{s-1}\|^2} \\ &\leq \sqrt{R} + \sqrt{R + \sum_{s=1}^{k}(6s-5)\|x^s - x^{s-1}\|^2},\end{aligned}$$

where the second inequality comes from the assumption on z. Moreover, since $a > 0$, the right-hand side is bounded from above according to Lemma 3.5. This implies that $\{x^k\}$ is bounded, and so it has accumulation points. □

Before proving Theorem 4.1 (ii), we show the following lemma.

Lemma 4.4
Let $\{x^k\}$ be generated by Algorithm 2 with $a > 0$ and suppose that Assumption 2.1 holds. Then, if $\bar{z}$ is an accumulation point of $\{x^k\}$, then $\{\|x^k - \bar{z}\|\}$ is convergent.

Proof. Assume that $\{x^{k_j}\} \subseteq \{x^k\}$ converges to $\bar{z}$. Then, we have $\sigma_{k_j}(\bar{z}) \to 0$ by the definition (12) of σ_{k_j}. Therefore, we can regard $\bar{z}$ to satisfy the statement of Lemma 4.1 at $k = \infty$, and thus the inequalities of Lemma 4.3 hold for any $r \geq 1$ and $\bar{z}$. This means $\{\nu_k(\bar{z})\}$ is non-increasing and bounded, i.e., convergent. Hence $\{\|x^k - \bar{z}\|\}$ is convergent. □

Finally, we finish the proof of the main theorem.

Theorem 4.1 (ii). Suppose that $\{x^{k_j^1}\}$ and $\{x^{k_j^2}\}$ converges to $\bar{z}^1$ and $\bar{z}^2$, respectively. From Lemma 4.4, we see that

$$\lim_{j\to\infty}\left(\left\|x^{k_j^2} - \bar{z}^1\right\|^2 - \left\|x^{k_j^2} - \bar{z}^2\right\|^2\right) = \lim_{j\to\infty}\left(\left\|x^{k_j^1} - \bar{z}^1\right\|^2 - \left\|x^{k_j^1} - \bar{z}^2\right\|^2\right).$$

This yields that $\|\bar{z}^1 - \bar{z}^2\|^2 = -\|\bar{z}^1 - \bar{z}^2\|^2$, and so $\|\bar{z}^1 - \bar{z}^2\|^2 = 0$, i.e., $\{x^k\}$ is convergent. Let $x^k \to x^*$. Since $\|x^{k+1} - x^k\|^2 \to 0$, $\{y^k\}$ is also convergent to x^*. Therefore, Proposition 2.1 shows that x^* is weakly Pareto optimal for (1). □

5 Numerical experiments

This section compares the performance between Algorithm 2 with various a and b and Algorithm 1 ($a = 0, b = 1/4$) through numerical experiments. Our newly introduced generalized momentum factor, while not primarily focused on improving convergence rates, serves to provide a theoretical link between different accelerated gradient methods. The primary goal of the numerical experiments is to confirm that our proposed method performs as theoretically expected. At the same time, it suggests that some momentum factors may potentially lead to better results. We run all experiments in Python 3.9.9 on a machine with 2.3 GHz Intel Core i7 CPU and 32 GB memory. For each example, we test 15 different hyperparameters combining $a = 0, 1/6, 1/4, 1/2, 3/4$ and $b = a^2/4, (a^2+1)/8, 1/4$, i.e.,

$$(a,b) = \left\{\begin{array}{c}(0,0),(0,1/8),(0,1/4),\\ (1/6,1/144),(1/6,37/288),(1/6,1/4),\\ (1/4,1/64),(1/4,17/128),(1/4,1/4),\\ (1/2,1/16),(1/2,5/32),(1/2,1/4),\\ (3/4,9/64),(3/4,25/128),(3/4,1/4)\end{array}\right\},$$

and we set $\varepsilon = 10^{-5}$ for the stopping criteria.

5.1 Artificial test problems (bi-objective and tri-objective)

First, we focus on solving multi-objective test problems, which are generally formulated as in problem (1). Specifically, we use the test problems of Tanabe et al. (2023b), whose objective functions are defined by

$$f_1(x) = \frac{1}{n}\|x\|^2, f_2(x) = \frac{1}{n}\|x-2\|^2, g_1(x) = g_2(x) = 0, \tag{JOS1}$$

$$f_1(x) = \frac{1}{n}\|x\|^2, f_2(x) = \frac{1}{n}\|x-2\|^2, g_1(x) = \frac{1}{n}\|x\|_1, g_2(x) = \frac{1}{2n}\|x-1\|_1, \tag{JOS1-L1}$$

$$\begin{cases} f_1(x) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} i(x_i - i)^4, f_2(x) = \exp\left(\sum_{i=1}^{n}\frac{x_i}{n}\right) + \|x\|^2, \\ f_3(x) = \dfrac{1}{n(n+1)}\displaystyle\sum_{i=1}^{n} i(n-i+1)\exp(-x_i), g_1(x) = g_2(x) = g_3(x) = 0, \end{cases} \tag{FDS}$$

$$\begin{cases} f_1(x) = \dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} i(x_i - i)^4, f_2(x) = \exp\left(\sum_{i=1}^{n}\frac{x_i}{n}\right) + \|x\|^2, \\ f_3(x) = \dfrac{1}{n(n+1)}\displaystyle\sum_{i=1}^{n} i(n-i+1)\exp(-x_i), g_1(x) = g_2(x) = g_3(x) = \chi_{\mathbf{R}^n_+}(x), \end{cases} \tag{FDS-CON}$$

where $x \in \mathbf{R}^n, n = 50$ and $\chi_{\mathbf{R}^n_+}$ is an indicator function (2) of the nonnegative orthant. These problems include modifications inspired by Jin et al. (2001); Fliege et al. (2009). We choose 1000 initial points, commonly for all pairs (a, b), and randomly with a uniform distribution between $\underline{c}$ and $\overline{c}$, where $\underline{c} = (-2, \dots, -2)^\top$ and $\overline{c} = (4, \dots, 4)^\top$ for (JOS1) and (JOS1-L1), $\underline{c} = (-2, \dots, -2)^\top$ and $\overline{c} = (2, \dots, 2)^\top$ for (FDS), and $\underline{c} = (0, \dots, 0)^\top$ and $\overline{c} = (2, \dots, 2)^\top$ for (FDS-CON). Moreover, we use backtracking for updating ℓ, with 1 as the initial value of ℓ and 2 as the constant multiplied into ℓ at each iteration (cf. (Tanabe et al., 2023b, Remark 4.1 (v))). Furthermore, at each iteration, we transform the subproblem (6) into their dual as suggested in Tanabe et al. (2023b) and solve them with the trust-region interior point method Byrd et al. (1999) using the scientific library SciPy.

Figure 1 and Table 1 present the experimental results. Figure 1 plots the solutions only for the cases $(a, b) = (0, 1/4), (3/4, 1/4)$, but other combinations also yield similar plots, including a wide range of Pareto solutions. Table 1 shows that the new momentum factors are fast enough to compete with the existing ones ($(a, b) = (0, 1/4)$ or $b = a^2/4$) and better than them in some cases.

5.2 Image deblurring (single-objective)

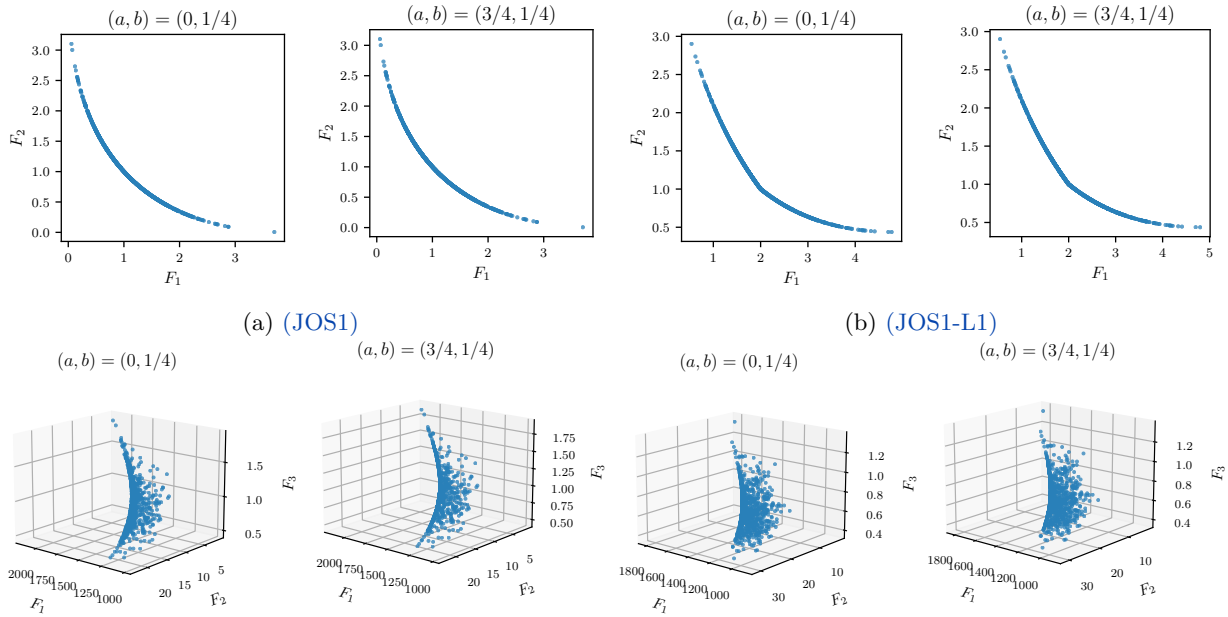
Since our proposed momentum factor is also new in the single-objective context, we also tackle deblurring the cameraman test image via a single-objective ℓ_2-ℓ_1 minimization, inspired by Beck & Teboulle (2009). This experiment also aims to show that our momentum coefficients, which combine existing well-known momentum coefficients while ensuring convergence of the point sequence, perform comparably well for application tasks. In detail, as shown in Figure 2, to a 256×256 cameraman test image with each pixel scaled to $[0, 1]$, we generate an observed image by applying a Gaussian blur of size 9×9 and standard deviation 4 and adding a zero-mean white Gaussian noise with standard deviation 10^{-3}.

Letting θ, B, and W be the observed image, the blur matrix, and the inverse of the Haar wavelet transform Haar (1910), respectively, consider the single-objective problem (1) with $m = 1$ and

$$f_1(x) \coloneqq \|BWx - \theta\|^2 \quad \text{and} \quad g_1(x) = \lambda\|x\|_1,$$

where $\lambda \coloneqq 2 \times 10^{-5}$ is a regularization parameter. Unlike in the previous subsection, we can compute ∇f's Lipschitz constant by calculating $(BW)^\top(BW)$'s eigenvalues using the two-dimensional cosine transform Hansen et al. (2006), so we use it constantly as ℓ. Moreover, we use the observed image's Wavelet transform as the initial point.

Figure 3 shows the reconstructed image from the obtained solution. Although there are some quirks in the way images are deblurred, such as the way stripes remain depending on the hyperparameters, it can be



(a) (JOS1)

(b) (JOS1-L1)

(c) (FDS)

(d) (FDS-CON)

Figure 1: Pareto solutions obtained with some (a, b)

(a) Original

(b) Blurred and noisy

Figure 2: Deblurring of the cameraman

Table 1: Average computational costs to solve the multi-objective examples

(a) (JOS1)

a	b	Time [s]	Iterations
0	0	6.442	97.0
0	1/8	5.158	81.217
0	1/4	4.207	65.0
1/6	1/144	4.244	67.0
1/6	37/288	5.182	82.0
1/6	1/4	4.268	66.0
1/4	1/64	6.224	99.0
1/4	17/128	7.239	113.566
1/4	1/4	3.205	51.0
1/2	1/16	4.51	72.0
1/2	5/32	4.562	71.0
1/2	1/4	4.466	70.0
3/4	9/64	4.323	67.998
3/4	25/128	3.104	49.0
3/4	1/4	3.741	47.0

(b) (JOS1-L1)

a	b	Time [s]	Iterations
0	0	10.733	157.512
0	1/8	11.054	161.065
0	1/4	11.122	161.734
1/6	1/144	9.85	141.731
1/6	37/288	9.994	144.863
1/6	1/4	10.399	150.592
1/4	1/64	9.271	135.804
1/4	17/128	9.463	137.108
1/4	1/4	9.662	139.848
1/2	1/16	7.439	109.082
1/2	5/32	7.642	110.204
1/2	1/4	7.723	111.599
3/4	9/64	5.253	77.366
3/4	25/128	5.39	79.425
3/4	1/4	5.678	82.37

(c) (FDS)

a	b	Time [s]	Iterations
0	0	29.24	204.438
0	1/8	29.797	210.595
0	1/4	30.565	214.934
1/6	1/144	24.964	174.393
1/6	37/288	25.375	177.944
1/6	1/4	26.065	182.398
1/4	1/64	22.94	159.737
1/4	17/128	23.311	162.629
1/4	1/4	23.976	166.918
1/2	1/16	17.909	122.653
1/2	5/32	18.14	123.96
1/2	1/4	18.221	125.697
3/4	9/64	13.584	94.176
3/4	25/128	13.674	94.705
3/4	1/4	13.795	94.868

(d) (FDS-CON)

a	b	Time [s]	Iterations
0	0	37.345	259.508
0	1/8	37.439	261.522
0	1/4	37.94	263.911
1/6	1/144	32.463	227.063
1/6	37/288	38.265	229.736
1/6	1/4	45.661	231.958
1/4	1/64	41.434	209.35
1/4	17/128	33.664	211.69
1/4	1/4	30.772	213.811
1/2	1/16	22.92	158.448
1/2	5/32	23.1	159.685
1/2	1/4	23.539	162.226
3/4	9/64	17.092	118.616
3/4	25/128	17.123	118.063
3/4	1/4	17.115	118.844

observed that deblurring is generally successful for all parameters. Moreover, we summarize the numerical performance in Table 2. Like the last subsection, this example also suggests that some of our new momentum factors may occasionally improve the algorithm's performance even for single-objective problems.

Table 2: Computational costs for the image deblurring

a	b	Total time [s]	Iteration counts	$F_1(x^{200})$
0	0	85.391	517	10.285
0	1/8	85.037	517	10.367
0	1/4	85.128	517	10.456
1/6	1/144	80.692	480	8.867
1/6	37/288	80.833	480	8.88
1/6	1/4	81.449	480	8.904
1/4	1/64	71.583	417	8.491
1/4	17/128	71.165	417	8.459
1/4	1/4	48.997	416	8.442
1/2	1/16	39.447	319	9.63
1/2	5/32	41.76	318	9.351
1/2	1/4	41.122	318	9.125
3/4	9/64	47.621	399	23.558
3/4	25/128	43.671	393	21.832
3/4	1/4	40.17	388	20.493

6 Conclusion

We have generalized the momentum factor of the multi-objective accelerated proximal gradient algorithm Tanabe et al. (2023b) in a form that is even new in the single-objective context and includes the known FISTA momentum factors Beck & Teboulle (2009); Chambolle & Dossal (2015). Furthermore, with the proposed momentum factor, we proved under reasonable assumptions that the algorithm has an $O(1/k^2)$ convergence rate and that the iterates converge to Pareto solutions. Moreover, the numerical results reinforced these theoretical properties and suggested the potential for our new momentum factor to improve the performance.

As we mentioned in Section 4, our proposed method has the potential to achieve finite-time manifold (active set) identification Sun et al. (2019) without the assumption of the strong convexity (or its generalizations such as PL conditions or error bounds Karimi et al. (2016)). Moreover, we took a single update rule of t_k for all iterations in this work, but the adaptive change of the strategy in each iteration is conceivable. It might also be interesting to estimate the Lipschitz constant simultaneously with that change, like in Scheinberg et al. (2014). In addition, an extension to the inexact scheme like Villa et al. (2013) would be significant. Furthermore, in single-objective optimization, non-convex objectives for FISTA have been proposed Li & Lin (2015), and extending this approach to multi-objective optimization remains an open problem. Regarding the application of our method in settings where only stochastic gradients are available, adapting our approach to such scenarios is an interesting direction for future research. Recent studies in multi-objective optimization with stochastic gradients Liu & Vicente (2021); Zhou et al. (2022) provide valuable insights and foundations for such an adaptation. This is an open area for exploration, possibly in conjunction with techniques such as dual averaging Xiao (2010). Those are issues to be addressed in the future.

Figure 3: Deblurred image

Acknowledgements

This work was supported by the Grant-in-Aid for Scientific Research (C) (21K11769 and 19K11840) and Grant-in-Aid for JSPS Fellows (20J21961) from the Japan Society for the Promotion of Science.

References

Hedy Attouch and Juan Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, sep 2016. ISSN 10526234. doi: 10.1137/15M1046095. URL https://doi.org/10.1137/15M1046095.

Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1):123–175, mar 2018. ISSN 1436-4646. doi: 10.1007/S10107-016-0992-8. URL https://doi.org/10.1007/s10107-016-0992-8.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, jan 2009. ISSN 19364954. doi: 10.1137/080716542. URL https://doi.org/10.1137/080716542.

Jose Yunier Bello Cruz, Jefferson G. Melo, and Ray V.G. Serra. A proximal gradient splitting method for solving convex vector optimization problems. *Optimization*, 71(1):33–53, jan 2022. ISSN 0233-1934. doi: 10.1080/02331934.2020.1800699. URL https://doi.org/10.1080/02331934.2020.1800699.

Radu Ioan Boţ and Sorin Mihai Grad. Inertial forward-backward methods for solving vector optimization problems. *Optimization*, 67(7):959–974, jul 2018. ISSN 10294945. doi: 10.1080/02331934.2018.1440553. URL https://doi.org/10.1080/02331934.2018.1440553.

Henri Bonnel, Alfredo Noel Iusem, and Benar Fux Svaiter. Proximal methods in vector optimization. *SIAM Journal on Optimization*, 15(4):953–970, jan 2005. ISSN 1052-6234. doi: 10.1137/S1052623403429093. URL https://doi.org/10.1137/S1052623403429093.

Richard H. Byrd, Mary E. Hribar, and Jorge Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, jan 1999. ISSN 1052-6234. doi: 10.1137/S1052623497325107. URL https://doi.org/10.1137/S1052623497325107.

Gabriel A. Carrizo, Pablo A. Lotito, and María C. Maciel. Trust region globalization strategy for the nonconvex unconstrained multiobjective optimization problem. *Mathematical Programming*, 159(1-2):339–369, sep 2016. ISSN 0025-5610. doi: 10.1007/s10107-015-0962-6. URL https://doi.org/10.1007/s10107-015-0962-6.

Emilio Carrizosa and J. B. G. Frenk. Dominating sets for convex functions with some applications. *Journal of Optimization Theory and Applications*, 96(2):281–295, feb 1998. ISSN 0022-3239. doi: 10.1023/A:1022614029984. URL https://doi.org/10.1023/A:1022614029984.

Antonin Chambolle and Charles Dossal. On the convergence of the iterates of the Fast Iterative Shrinkage/Thresholding Algorithm. *Journal of Optimization Theory and Applications*, 166(3):968–982, may 2015. ISSN 1573-2878. doi: 10.1007/S10957-015-0746-4. URL https://doi.org/10.1007/s10957-015-0746-4.

Jean Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, mar 2012. ISSN 1631-073X. doi: 10.1016/J.CRMA.2012.03.014. URL https://doi.org/10.1016/J.CRMA.2012.03.014.

Jin-Dong Dong, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, and Min Sun. DPP-Net: Device-aware progressive search for Pareto-optimal neural architectures. In Vittori Ferrari, Hebert Martial, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 540–555. Springer Cham, Munich, first edition, 2018. ISBN 9783030012519. doi: 10.1007/978-3-030-01252-6_32. URL https://doi.org/10.1007/978-3-030-01252-6_32.

Thomas Elsken, Frank Hutter, and Jan Hendrik Metzen. Efficient multi-objective neural architecture search via Lamarckian evolution. In *7th International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByME42AqK7.

Hans Eschenauer, Juhani Koski, and Andrzej Osyczka. *Multicriteria Design Optimization*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1990. ISBN 978-3-642-48699-9. doi: 10.1007/978-3-642-48697-5. URL https://doi.org/10.1007/978-3-642-48697-5.

Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494, aug 2000. ISSN 1432-2994. doi: 10.1007/s001860000043. URL https://doi.org/10.1007/s001860000043.

Jörg Fliege, Luis Mauricio Graña Drummond, and Benar Fux Svaiter. Newton's method for multiobjective optimization. *SIAM Journal on Optimization*, 20(2):602–626, jan 2009. ISSN 10526234. doi: 10.1137/08071692X. URL https://doi.org/10.1137/08071692X.

Jörg Fliege, A. Ismael F. Vaz, and Luis Nunes Vicente. Complexity of gradient descent for multiobjective optimization. *Optimization Methods and Software*, 34(5):949–959, aug 2019. ISSN 10294937. doi: 10.1080/10556788.2018.1510928. URL https://doi.org/10.1080/10556788.2018.1510928.

Ellen Hidemi Fukuda and Luis Mauricio Graña Drummond. Inexact projected gradient method for vector optimization. *Computational Optimization and Applications*, 54(3):473–493, apr 2013. ISSN 09266003. doi: 10.1007/s10589-012-9501-z. URL http://doi.org/10.1007/s10589-012-9501-z.

Ellen Hidemi Fukuda and Luis Mauricio Graña Drummond. A survey on multiobjective descent methods. *Pesquisa Operacional*, 34(3):585–620, dec 2014. ISSN 16785142. doi: 10.1590/0101-7438.2014.034.03.0585. URL https://doi.org/10.1590/0101-7438.2014.034.03.0585.

Xavier Gandibleux, Marc Sevaux, Kenneth Sörensen, and Vincent T'kindt. *Metaheuristics for Multiobjective Optimisation*, volume 535 of *Lecture Notes in Economics and Mathematical Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-20637-8. doi: 10.1007/978-3-642-17144-4. URL http://doi.org/10.1007/978-3-642-17144-4.

Saul Gass and Thomas Saaty. The computational algorithm for the parametric objective function. *Naval Research Logistics Quarterly*, 2(1-2):39–45, 1955. ISSN 00281441. doi: 10.1002/nav.3800020106. URL https://doi.org/10.1002/nav.3800020106.

Arthur M. Geoffrion. Proper efficiency and the theory of vector maximization. *Journal of Mathematical Analysis and Applications*, 22(3):618–630, jun 1968. ISSN 0022247X. doi: 10.1016/0022-247X(68)90201-1. URL https://doi.org/10.1016/0022-247X(68)90201-1.

Max Leandro Nobre Gonçalves, Fernando S. Lima, and Leandro F. Prudente. Globally convergent Newton-type methods for multiobjective optimization. *Computational Optimization and Applications*, 83(2):403–434, nov 2022. ISSN 0926-6003. doi: 10.1007/s10589-022-00414-7. URL http://www.optimization-online.org/DB_HTML/2020/08/7955.htmlhttps://link.springer.com/10.1007/s10589-022-00414-7.

Alfred Haar. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69(3):331–371, sep 1910. ISSN 0025-5831. doi: 10.1007/BF01456326. URL http://link.springer.com/10.1007/BF01456326.

Per Christian Hansen, James G. Nagy, and Dianne P. O'Leary. *Deblurring Images: Matrices, Spectra, and Filtering*. Society for Industrial and Applied Mathematics, jan 2006. ISBN 978-0-89871-618-4. doi: 10.1137/1.9780898718874. URL https://doi.org/10.1137/1.9780898718874.

Yaochu Jin, Markus Olhofer, and Bernhard Sendhoff. Dynamic weighted aggregation for evolutionary multi-objective optimization: Why does it work and how? In *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, GECCO'01, pp. 1042–1049, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607749. doi: 10.5555/2955239.2955427. URL https://dl.acm.org/doi/10.5555/2955239.2955427.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 795–811, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46128-1. doi: 10.1007/978-3-319-46128-1_50. URL https://doi.org/10.1007/978-3-319-46128-1_50.

Ye-Hoon Kim, Bhargava Reddy, Sojung Yun, and Chanwon Seo. NEMO: Neuro-evolution with multiobjective optimization of deep neural network for speed and accuracy. In *ICML'17 AutoML Workshop*, 2017. URL https://www.semanticscholar.org/paper/0a9c6947a0b6f79526e537cb83925ef60df674e8.

Huan Li and Zhouchen Lin. Accelerated Proximal Gradient Methods for Nonconvex Programming. In C Cortes, N Lawrence, D Lee, M Sugiyama, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qingfu Zhang, and Sam Kwong. Pareto multi-Task learning. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 12060–12070, dec 2019. doi: 10.5555/3454287.3455367. URL https://dl.acm.org/doi/10.5555/3454287.3455367.

Suyun Liu and Luis Nunes Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. *Annals of Operations Research*, mar 2021. ISSN 0254-5330. doi: 10.1007/s10479-021-04033-z. URL http://link.springer.com/10.1007/s10479-021-04033-z.

Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. *Computational Management Science*, 19(3):513–537, jul 2022. ISSN 1619-697X. doi: 10.1007/s10287-022-00425-z. URL https://link.springer.com/10.1007/s10287-022-00425-z.

L. R. Lucambio Pérez and L. F. Prudente. Nonlinear conjugate gradient methods for vector optimization. *SIAM Journal on Optimization*, 28(3):2690–2720, jan 2018. ISSN 1052-6234. doi: 10.1137/17M1126588. URL https://doi.org/10.1137/17M1126588.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269:543–547, 1983. URL http://mi.mathnet.ru/eng/dan/v269/i3/p543.

Katya Scheinberg, Donald Goldfarb, and Xi Bai. Fast first-order methods for composite convex optimization with backtracking. *Foundations of Computational Mathematics*, 14(3):389–417, jun 2014. ISSN 1615-3375. doi: 10.1007/s10208-014-9189-9. URL https://doi.org/10.1007/s10208-014-9189-9.

Ozan Sener. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 525–536, Red Hook, NY, USA, 2018. Curran Associates Inc. doi: 10.5555/3326943.3326992. URL https://dl.acm.org/doi/10.5555/3326943.3326992.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016. URL https://jmlr.org/papers/v17/15-084.html.

Yifan Sun, Halyun Jeong, Julie Nutini, and Mark Schmidt. Are we there yet? Manifold identification of gradient-related proximal methods. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR*, pp. 1110–1119, 2019. URL http://proceedings.mlr.press/v89/sun19a.html.

Hiroki Tanabe, Ellen Hidemi Fukuda, and Nobuo Yamashita. Proximal gradient methods for multi-objective optimization and their applications. *Computational Optimization and Applications*, 72(2):339–361, mar 2019. ISSN 15732894. doi: 10.1007/s10589-018-0043-x. URL https://doi.org/10.1007/s10589-018-0043-x.

Hiroki Tanabe, Ellen H. Fukuda, and Nobuo Yamashita. Convergence rates analysis of a multiobjective proximal gradient method. *Optimization Letters*, 17(2):333–350, apr 2023a. ISSN 18624480. doi: 10.1007/s11590-022-01877-7. URL https://doi.org/10.1007/s11590-022-01877-7.

Hiroki Tanabe, Ellen Hidemi Fukuda, and Nobuo Yamashita. An accelerated proximal gradient method for multiobjective optimization. *Computational Optimization and Applications*, 86(2):421–455, nov 2023b. ISSN 0926-6003. doi: 10.1007/s10589-023-00497-w. URL https://doi.org/10.48550/arXiv.2202.10994https://link.springer.com/10.1007/s10589-023-00497-w.

Hiroki Tanabe, Ellen Hidemi Fukuda, and Nobuo Yamashita. New merit functions for multiobjective optimization and their properties. *Optimization*, pp. 1–38, jul 2023c. ISSN 0233-1934. doi: 10.1080/02331934.2023.2232794. URL https://doi.org/10.48550/arXiv.2010.09333http://arxiv.org/abs/2010.09333https://www.tandfonline.com/doi/full/10.1080/02331934.2023.2232794.

Silvia Villa, Saverio Salzo, Luca Baldassarre, and Alessandro Verri. Accelerated and inexact forward-backward algorithms. *SIAM Journal on Optimization*, 23(3):1607–1633, aug 2013. ISSN 10526234. doi: 10.1137/110844805. URL https://doi.org/10.1137/110844805.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010. URL http://jmlr.org/papers/v11/xiao10a.html.

L. A. Zadeh. Optimality and non-scalar-valued performance criteria. *IEEE Transactions on Automatic Control*, 8(1):59–60, 1963. ISSN 15582523. doi: 10.1109/TAC.1963.1105511. URL https://doi.org/10.1109/TAC.1963.1105511.

Shiji Zhou, Wenpeng Zhang, Jiyan Jiang, Wenliang Zhong, Jinjie GU, and Wenwu Zhu. On the Convergence of Stochastic Multi-Objective Gradient Manipulation and Beyond. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38103–38115. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f91bd64a3620aad8e70a27ad9cb3ca57-Paper-Conference.pdf.