An accelerated proximal gradient method for multiobjective optimization

Hiroki Tanabe[1], Ellen H. Fukuda[1], and Nobuo Yamashita[1]

[1]Graduate School of Informatics, Kyoto University
{tanabehiroki@amp.i.kyoto-u.ac.jp},{ellen,nobuo}@i.kyoto-u.ac.jp

Abstract

Many researchers have studied descent methods for multiobjective optimization problems in recent years. For example, Fliege and Svaiter proposed the steepest descent method for differentiable multiobjective optimization problems. Afterward, a proximal gradient method, which can solve non-differentiable problems, was also considered. However, their accelerated versions are not sufficiently studied. Recently, El Moudden and El Mouatasim proposed a natural extension of Nesterov's accelerated method for multiobjective optimization problems. They proved the global convergence rate of the algorithm ($O(1/k^2)$) under the assumption that the sequence of the Lagrangian multipliers of the subproblems is eventually fixed. However, this assumption is restrictive because it means that the method is regarded as the Nesterov's method for the weighting problem. In this paper, we propose a new accelerated algorithm, in which we solve subproblems with terms that only appear in the multiobjective case. We also show the proposed method's global convergence rate ($O(1/k^2)$) under a more natural assumption, using a merit function to measure the complexity. Moreover, we present an efficient way to solve the subproblem in the proposed method via its dual, and we confirm the validity of the proposed method through numerical experiments.

1 Introduction

Multiobjective optimization consists in minimizing (or maximizing) more than one objective function at once under possible constraints. In general, there is no single point that minimizes all objective functions simultaneously, so the concept of *Pareto optimality* becomes essential. We call a point Pareto optimal if there is no other point with the same or smaller objective function values and with at least one objective function value being strictly smaller.

One of the most popular strategies for solving multiobjective optimization problems is the *scalarization approach* [13, 14, 24]. It converts the original multiobjective optimization problem into one or several parametrized single-objective optimization problems. However, we may find it challenging to choose the appropriate parameters in advance.

In recent years, descent algorithms for multiobjective optimization problems have attracted much attention in the optimization community. For example, Fliege and Svaiter [11] proposed the steepest descent method for differentiable multiobjective optimization problems. Later, the proximal gradient method [21], which works for non-differentiable problems, was considered. We call them *first-order methods* since they only use the objective functions' first derivative. The sequences generated by these methods are known to converge to the Pareto solutions with rate $O(1/k)$ [12, 23].

On the other hand, there are many kinds of research about the acceleration of the single-objective first-order methods. After being established by Nesterov [18], researchers developed various accelerated methods. In particular, the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [2], an accelerated version of the proximal gradient method, has contributed to a wide range of research fields, such as image and signal processing.

However, studies for accelerated algorithms for multiobjective cases remain insufficient. In 2020, El Moudden and El Mouatasim [9] proposed an accelerated diagonal steepest descent method for multiobjective optimization, a natural extension of Nesterov's accelerated method for single-objective problems. They proved the global convergence rate of the algorithm ($O(1/k^2)$) under the assumption that the sequence of the Lagrangian multipliers of the subproblems is eventually fixed. Nevertheless, this assumption is restrictive because it indicates that the method is essentially the same as the Nesterov's method for minimizing a weighted sum of the objective functions.

Here, we propose a new accelerated algorithm for multiobjective optimization, in which we solve subproblems with terms that only appear in the multiobjective case. These terms vanish in the single-objective case; we can regard these methods as the generalization of the single-objective accelerated methods. Moreover, under more natural assumptions, we prove the proposed method's global convergence rate ($O(1/k^2)$) by using a merit function [22] to measure the complexity.

Furthermore, we derive a convex and differentiable dual problem for the subproblem in each iteration, which is easier to solve than the original problem when the number of objective functions is smaller than the dimension of the decision variables. We can reconstruct the original subproblem's solution from the dual problem's solution. We also implement the whole algorithm using this dual problem and confirm its effectiveness in numerical experiments.

The outline of this paper is as follows. We present some notations and concepts used in this paper in Section 2. Section 3 recalls the proximal gradient method for multiobjective optimization proposed in [21]. We present the accelerated proximal gradient method for multiobjective optimization in Section 4 and analyze its $O(1/k^2)$ convergence rate in Section 5. Section 6 introduces the efficient way to solve the subproblem via its dual problem. Finally, we report some numerical results for test problems in Section 7, demonstrating that the proposed method is faster than the existing one.

2 Preliminaries

All over this work, for any natural number d, $\mathbf{R}^d$ denotes the d-dimensional real space, $\mathbf{R}^d_+ \subseteq \mathbf{R}^d$ designates the nonnegative orthant of $\mathbf{R}^d$, i.e.,

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

and Δ^d represents the standard simplex given by

$$\Delta^d \coloneqq \left\{ \lambda \in \mathbf{R}^d_+ \;\middle|\; \sum_{i=1}^{d} \lambda_i = 1 \right\}. \tag{1}$$

Then, we can consider the partial orders induced by $\mathbf{R}^d_+$: for all $v^1, v^2 \in \mathbf{R}^d$, $v^1 \leq v^2$ (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 $v^1 < v^2$ stand for $v^1_i \leq v^2_i$ and $v^1_i < v^2_i$ for all $i = 1, \dots, d$, respectively. Moreover, let $\langle \cdot, \cdot \rangle$ be the Euclidean inner product in $\mathbf{R}^d$, i.e., $\langle u, v \rangle \coloneqq \sum_{i=1}^{d} u_i v_i$, and let $\|\cdot\|$ be the Euclidean norm, i.e., $\|u\| \coloneqq \sqrt{\langle u, u \rangle}$. Furthermore, we define ℓ_1-norm $\|\cdot\|_1$ and ℓ_∞-norm $\|\cdot\|_\infty$ by $\|u\|_1 \coloneqq \sum_{i=1}^{d} |u_i|$ and $\|u\|_\infty \coloneqq \max_{i=1,\dots,d} |u_i|$, respectively.

On the other hand, for a closed, proper and convex function $h \colon \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$, we call $\eta \in \mathbf{R}^n$ a subgradient of h at $x \in \mathbf{R}^n$ if

$$h(y) \geq h(x) + \langle \eta, y - x \rangle \quad \text{for all } y \in \mathbf{R}^n,$$

and we write $\partial h(x)$ the subdifferential of h at x, i.e., the set of all subgradients of h at x. In addition, the subdifferential for a vector-valued function is the direct product of the subdifferentials of each component. We also define the *Moreau envelope* or *Moreau-Yosida regularization* of h by

$$\mathcal{M}_h(x) \coloneqq \min_{y \in \mathbf{R}^n} \left\{ h(y) + \frac{1}{2} \|x - y\|^2 \right\}. \tag{2}$$

The minimization problem in (2) has a unique solution. We call this solution the *proximal operator* and write it as

$$\mathbf{prox}_h(x) \coloneqq \operatorname*{argmin}_{y \in \mathbf{R}^n} \left\{ h(y) + \frac{1}{2} \|x - y\|^2 \right\}. \tag{3}$$

Remark 2.1. *(i) [1, Theorem 6.24] If h is the* indicator function *of a nonempty set $S \subseteq \mathbf{R}^n$, i.e.,*

$$\chi_S(x) = \begin{cases} 0, & x \in S, \\ \infty, & x \notin S, \end{cases} \tag{4}$$

then the proximal operator reduces to the projection onto S.

(ii) [1, Theorem 6.42] The proximal operator of a closed, proper, and convex function h is non-expansive, i.e., $\|\mathbf{prox}_h(x) - \mathbf{prox}_h(y)\| \leq \|x - y\|$. In other words, $\mathbf{prox}_h$ is 1-Lipschitz continuous.

(iii) [1, Theorem 6.60] Even if a closed, proper, and convex function h is non-differentiable, its Moreau envelope $\mathcal{M}_h$ has a 1-Lipschitz continuous gradient as follows:

$$\nabla \mathcal{M}_h(x) = x - \mathbf{prox}_h(x).$$

By the way, in this paper, we focus on the following multiobjective optimization problem:

$$\min_{x \in \mathbf{R}^n} \quad F(x), \tag{5}$$

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

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

with convex and continuously differentiable functions $f_i \colon \mathbf{R}^n \to \mathbf{R}, i = 1, \dots, m$ and closed, proper and convex functions $g_i \colon \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}, i = 1, \dots, m$. We also suppose that each ∇f_i is Lipschitz continuous with constant $L_i > 0$ and define

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

From the so-called descent lemma [4, Proposition A.24], we have

$$f_i(p) - f_i(q) \leq \langle \nabla f_i(q), p - q \rangle + \frac{L}{2} \|p - q\|^2 \tag{6}$$

for all $p, q \in \mathbf{R}^n$ and $i = 1, \dots, m$, which gives

$$\begin{aligned} F_i(p) - F_i(r) &= f_i(p) - f_i(q) + g_i(p) + f_i(q) - F_i(r) \\ &\leq \langle \nabla f_i(q), p - q \rangle + g_i(p) + f_i(q) - F_i(r) + \frac{L}{2} \|p - q\|^2 \end{aligned} \tag{7}$$

for all $p, q, r \in \mathbf{R}^n$ and $i = 1, \dots, m$.

Now, we introduce some concepts used in the multiobjective optimization problem (5). Recall that

$$X^* \coloneqq \{x^* \in \mathbf{R}^n \mid \text{There does not exist } x \in \mathbf{R}^n \text{ such that } F(x) < F(x^*)\} \tag{8}$$

is the set of *weakly Pareto optimal* points for (5). We also define the effective domain of F by

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

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

$$\mathcal{L}_F(c) \coloneqq \{x \in \mathbf{R}^n \mid F(x) \leq c\}. \tag{9}$$

In addition, 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 $u_0 \colon \mathbf{R}^n \to \mathbf{R} \cup \{\infty\}$ proposed in [22]:

$$u_0(x) \coloneqq \sup_{z \in \mathbf{R}^n} \min_{i=1,\dots,m} \{F_i(x) - F_i(z)\}, \tag{10}$$

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. *[22, Theorems 3.3 and 3.4] Let u_0 be defined by* (10)*. Then, $u_0(x) \geq 0$ for all $x \in \mathbf{R}^n$. Moreover, $x \in \mathbf{R}^n$ is weakly Pareto optimal for* (5) *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. This is clearly a merit function for scalar-valued optimization.

3 Proximal gradient methods for multiobjective optimization

Let us now recall the proximal gradient method for (5), an extension of the classical proximal gradient method, proposed by Tanabe, Fukuda, and Yamashita [21]. We explain how to generate the sequence in each iteration, and afterward, we show the algorithm and its convergence rate.

For given $x \in \operatorname{dom} F$ and $\ell > 0$, we consider the following minimization problem:

$$\min_{z \in \mathbf{R}^n} \quad \varphi(z; x), \tag{11}$$

where

$$\varphi_\ell(z; x) \coloneqq \max_{i=1,\dots,m} \{\langle \nabla f_i(x), z - x \rangle + g_i(z) - g_i(x)\} + \frac{\ell}{2} \|z - x\|^2. \tag{12}$$

The convexity of g_i implies that $z \mapsto \varphi_\ell(z; x)$ is strongly convex, so the problem (11) always has a unique solution. Let us write such a solution as $p_\ell(x)$ and let $\theta_\ell(x)$ be its optimal function value, i.e.,

$$p_\ell(x) \coloneqq \operatorname*{argmin}_{z \in \mathbf{R}^n} \varphi(z; x) \quad \text{and} \quad \theta_\ell(x) \coloneqq \min_{z \in \mathbf{R}^n} \varphi(z; x). \tag{13}$$

The following proposition shows that $p_\ell(x)$ and $\theta_\ell(x)$ characterize the weak Pareto optimality of (5).

Proposition 3.1. *Let p_ℓ and θ_ℓ be defined by* (13)*. Then, the statements below hold.*

(i) The following three conditions are equivalent: (a) x is weakly Pareto optimal, (b) $p_\ell(x) = x$, and (c) $\theta_\ell(x) = 0$.

(ii) The mappings p_ℓ and θ_ℓ are both continuous.

Proof. It is clear from [21, Lemma 3.2] and the convexity of f_i. □

From Proposition 3.1, we can treat $\|p_\ell(x) - x\|_\infty < \varepsilon$ for some ε as the stopping criteria. Moreover, it is known that if $\ell > L/2$ then we have $F_i(p_\ell(x)) \leq F_i(x)$ for all $x \in \operatorname{dom} F$ and $i = 1, \ldots, m$ [23]. Now, we state below the proximal gradient method for (5).

Algorithm 1 Proximal gradient method for multiobjective optimization [21]

```
Input: $x^0 \in \operatorname{dom} F$, $\ell > L/2$, $\varepsilon > 0$
Output: $x^*$: A weakly Pareto optimal point
$k \leftarrow 0$
while $\left\|p_\ell(x^k) - x^k\right\|_\infty \geq \varepsilon$ do
    $x^{k+1} \leftarrow p_\ell(x^k)$
    $k \leftarrow k + 1$
end while
```

When $\ell \geq L$, Algorithm 1 is known to generate $\left\{x^k\right\}$ such that $\left\{u_0(x^k)\right\}$ converges to zero with rate $O(1/k)$ under the following two assumptions. Note that these assumptions are not particularly strong, as suggested in [23, Remark 5.2].

Assumption 3.1. *[23, Assumption 5.1] Let X^* and $\mathcal{L}_F$ be defined by* (8) *and* (9)*, 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*

$$R \coloneqq \sup_{F^* \in F(X^* \cap \mathcal{L}_F(F(x^0)))} \inf_{z \in F^{-1}(\{F^*\})} \left\|z - x^0\right\|^2. \tag{14}$$

Theorem 3.1. *[23, Theorem 5.2] Assume that $\ell \geq L$. Then, under Assumption 3.1, Algorithm 1 generates a sequence $\left\{x^k\right\}$ such that*

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

At the end of this section, we note some remarks about Algorithm 1.

Remark 3.1. *(i) Since $x \in \operatorname{dom} F$ implies $p_\ell(x) \in \operatorname{dom} F$, Algorithm 1 is well-defined.*

(ii) When $g_i = 0$, Algorithm 1 corresponds to the steepest descent method [11]. On the other hand, when $f_i = 0$, it matches the proximal point method [6]. Furthermore, when g_i is the indicator function (4) *of a convex set $S \subseteq \mathbf{R}^n$, it coincides with the projected gradient method [15].*

(iii) When it is difficult to estimate the Lipschitz constant L, we can set the initial value of ℓ appropriately and increase ℓ by multiplying at each iteration by some prespecified scalar until $F_i(p_\ell(x^k)) - F_i(x^k) \le \theta_\ell(x^k)$ is satisfied for all $i = 1, \dots, m$. If L is finite, the number of times that ℓ is increased is at most a constant.

4 An accelerated proximal gradient method for multiobjective optimization

This section proposes an accelerated version of the proximal gradient method for multiobjective optimization.

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

$$\min_{z \in \mathbf{R}^n} \quad \varphi_\ell^{\mathrm{acc}}(z; x, y), \tag{15}$$

where

$$\varphi_\ell^{\mathrm{acc}}(z; x, y) \coloneqq \max_{i=1,\dots,m} \left\{ \langle \nabla f_i(y), z - y \rangle + g_i(z) + f_i(y) - F_i(x) \right\} + \frac{\ell}{2} \|z - y\|^2. \tag{16}$$

Note that when $y = x$, the subproblem (15) is reduced to subproblem (11) of the proximal gradient method. Note also that when $m = 1$, the subproblem becomes

$$\min_{z \in \mathbf{R}^n} \quad \langle \nabla f_1(y), z - y \rangle + g_1(z) + \frac{\ell}{2} \|z - y\|^2.$$

Therefore, the term $F_i(x)$ is a key for acceleration in the multiobjective optimization.

Since g_i is convex for all $i = 1, \dots, m$, $z \mapsto \varphi_\ell^{\mathrm{acc}}(z; x, y)$ is strongly convex. Therefore, the subproblem (15) has a unique optimal solution $p_\ell^{\mathrm{acc}}(x, y)$ and takes the optimal function value $\theta_\ell^{\mathrm{acc}}(x, y)$, i.e.,

$$p_\ell^{\mathrm{acc}}(x, y) \coloneqq \operatorname*{argmin}_{z \in \mathbf{R}^n} \varphi_\ell^{\mathrm{acc}}(z; x, y) \quad \text{and} \quad \theta_\ell^{\mathrm{acc}}(x, y) \coloneqq \min_{z \in \mathbf{R}^n} \varphi_\ell^{\mathrm{acc}}(z; x, y). \tag{17}$$

Moreover, the optimality condition of (15) implies that for all $x \in \operatorname{dom} F$ and $y \in \mathbf{R}^n$ there exists $\eta(x, y) \in \partial g(p_\ell^{\mathrm{acc}}(x, y))$ and a Lagrange multiplier $\lambda(x, y) \in \mathbf{R}^m$ such that

$$\sum_{i=1}^m \lambda_i(x, y) \left\{ \nabla f_i(y) + \eta_i(x, y) \right\} = -\ell \left(p_\ell^{\mathrm{acc}}(x, y) - y \right) \tag{18a}$$

$$\lambda(x, y) \in \Delta^m, \quad \lambda_j(x, y) = 0 \quad \text{for all } j \notin \mathcal{I}(x, y), \tag{18b}$$

where Δ^m denotes the standard simplex (1) and

$$\mathcal{I}(x, y) \coloneqq \operatorname*{argmax}_{i=1,\dots,m} \left\{ \langle \nabla f_i(y), p_\ell^{\mathrm{acc}}(x, y) - y \rangle + g_i(p_\ell^{\mathrm{acc}}(x, y)) + f_i(y) - F_i(x) \right\}.$$

We also note that by taking $z = y$ in the objective function of (15), we have

$$\theta_\ell^{\mathrm{acc}}(x, y) \le \varphi_\ell^{\mathrm{acc}}(y; x, y) = \max_{i=1,\dots,m} \{F_i(y) - F_i(x)\} \tag{19}$$

for all $x \in \operatorname{dom} F$ and $y \in \mathbf{R}^n$. Moreover, it follows from (7) that

$$\theta_\ell^{\mathrm{acc}}(x, y) \ge \max_{i=1,\dots,m} \{F_i(p_\ell^{\mathrm{acc}}(x, y)) - F_i(x)\} \tag{20}$$

for all $x \in \operatorname{dom} F$ and $y \in \mathbf{R}^n$. We now characterize weak Pareto optimality in terms of the mappings p_ℓ^{acc} and $\theta_\ell^{\mathrm{acc}}$, corresponding to Proposition 3.1 for the proximal gradient method.

Proposition 4.1. *Let $p_\ell^{\mathrm{acc}}(x, y)$ and $\theta_\ell^{\mathrm{acc}}(x, y)$ be defined by* (17)*. Then, the statements below hold.*

- *(i) The following three conditions are equivalent: (a) $y \in \mathbf{R}^n$ is weakly Pareto optimal for* (5)*, (b) $p_\ell^{\mathrm{acc}}(x, y) = y$ for some $x \in \mathbf{R}^n$, and (c) $\theta_\ell^{\mathrm{acc}}(x, y) = \max_{i=1,\dots,m} \{F_i(y) - F_i(x)\}$ for some $x \in \mathbf{R}^n$.*
- *(ii) The mappings p_ℓ^{acc} and $\theta_\ell^{\mathrm{acc}}$ are both continuous.*

Proof. (i): From (19) and the uniqueness of the optimal solution for (15), the equivalence between (b) and (c) is apparent. Now, let us show that (a) and (b) are equivalent. When y is weakly Pareto optimal, we can immediately see from Proposition 3.1 that $p_\ell^{\mathrm{acc}}(x, y) = p_\ell(y) = y$ by letting $x = y$. Conversely, suppose that $p_\ell^{\mathrm{acc}}(x, y) = y$ for some $x \in \mathbf{R}^n$. Let $z \in \mathbf{R}^n$ and $\alpha \in (0, 1)$. The optimality of $p_\ell^{\mathrm{acc}}(x, y) = y$ for (15) gives

$$\begin{aligned}
\max_{i=1,\dots,m} &\{F_i(y) - F_i(x)\} \le \varphi_\ell^{\mathrm{acc}}(y + \alpha(z - y); x, y) \\
&= \max_{i=1,\dots,m} \{\langle \nabla f_i(y), \alpha(z - y)\rangle + g_i(y + \alpha(z - y)) + f_i(y) - F_i(x)\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \frac{\ell}{2}\|\alpha(z - y)\|^2.
\end{aligned}$$

Thus, from the convexity of f_i, we get

$$\begin{aligned}
\max_{i=1,\dots,m} &\{F_i(y) - F_i(x)\} \le \varphi_\ell^{\mathrm{acc}}(y + \alpha(z - y); x, y) \\
&\qquad\qquad \le \max_{i=1,\dots,m} \{F_i(y + \alpha(z - y)) - F_i(x)\} + \frac{\ell}{2}\|\alpha(z - y)\|^2.
\end{aligned}$$

Moreover, the convexity of F_i yields

$$\begin{aligned}
\max_{i=1,\dots,m} &\{F_i(y) - F_i(x)\} \le \varphi_\ell^{\mathrm{acc}}(y + \alpha(z - y); x, y) \\
&\le \max_{i=1,\dots,m} \{\alpha F_i(z) + (1 - \alpha) F_i(y) - F_i(x)\} + \frac{\ell}{2}\|\alpha(z - y)\|^2 \\
&\le \alpha \max_{i=1,\dots,m} \{F_i(z) - F_i(y)\} + \max_{i=1,\dots,m} \{F_i(y) - F_i(x)\} + \frac{\ell}{2}\|\alpha(z - y)\|^2.
\end{aligned}$$

Therefore, we get

$$\max_{i=1,\dots,m} \{F_i(z) - F_i(y)\} \geq -\frac{\ell\alpha}{2} \|z - y\|^2.$$

Taking $\alpha \searrow 0$, we obtain

$$\max_{i=1,\dots,m} \{F_i(z) - F_i(y)\} \geq 0.$$

which implies the weak Pareto optimality of y.

(ii): The objective function of (15) is continuous for x, y, and z. Thus, the optimal value function $\theta_\ell^{\mathrm{acc}}$ is also continuous from [3, Maximum Theorem]. Furthermore, since the optimal set mapping p_ℓ^{acc} is unique, p_ℓ^{acc} is continuous from [16, Corollary 8.1]. □

Proposition 4.1 implies that we can use $\|p_\ell^{\mathrm{acc}}(x, y) - y\|_\infty < \varepsilon$ for some $\varepsilon > 0$ as the stopping criteria. Now, we state below the proposed algorithm.

Algorithm 2 Accelerated proximal gradient method for multiobjective optimization

```
Input: Set $x^0 = y^1 \in \operatorname{dom} F, \ell \geq L, \varepsilon > 0$.
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:     $x^k \leftarrow p_\ell^{\mathrm{acc}}(x^{k-1}, y^k)$
5:     $t_{k+1} \leftarrow \sqrt{t_k^2 + 1/4} + 1/2$
6:     $\gamma_k \leftarrow (t_k - 1)/t_{k+1}$
7:     $y^{k+1} \leftarrow x^k + \gamma_k (x^k - x^{k-1})$
8:     $k \leftarrow k + 1$
9: end while
```

We show below the properties that $\{t_k\}$ and $\{\gamma_k\}$ satisfies.

Lemma 4.1. *Let $\{t_k\}$ and $\{\gamma_k\}$ be defined by Lines 2, 5 and 6 in Algorithm 2. Then, the following inequalities hold for all $k \geq 1$.*

(i) $t_{k+1} \geq t_k + 1/2$ *and* $t_k \geq (k+1)/2$.

(ii) $t_k^2 - t_{k+1}^2 + t_{k+1} = 0$.

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

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

$$t_{k+1} = \sqrt{t_k^2 + \frac{1}{4}} + \frac{1}{2} \geq t_k + \frac{1}{2}. \tag{21}$$

Applying the above inequality recursively, we obtain

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

(ii): An easy computation shows that

$$t_k^2 - t_{k+1}^2 + t_{k+1} = t_k^2 - \left[\sqrt{t_k^2 + \frac{1}{4}} + \frac{1}{2} \right]^2 + \sqrt{t_k^2 + \frac{1}{4}} + \frac{1}{2} = 0.$$

(iii): The statement (i) of this lemma implies that $t_{k+1} > t_k \geq 1$. Thus, the definition of γ_k leads to

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

□

We end this section by noting some remarks about the proposed algorithm.

Remark 4.1. *(i) Since $x \in \operatorname{dom} F$ implies $p_\ell^{\mathrm{acc}}(x, y) \in \operatorname{dom} F$, every x^k computed by the above algorithm is in $\operatorname{dom} F$. However, y^k is not necessarily in $\operatorname{dom} F$.*

(ii) Since $y^1 = x^0$, it follows from (19) that

$$\theta_\ell^{\mathrm{acc}}(x^0, y^1) \leq 0, \tag{22}$$

but the inequality $\theta_\ell^{\mathrm{acc}}(x^{k-1}, y^k) \leq 0$ does not necessarily hold for $k \geq 2$.

(iii) When $m = 1$, we can remove the term $f_i(y) - F_i(x)$ from the subproblem (15), so Algorithm 2 corresponds to the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [2] for single-objective optimization.

(iv) Like Remark 3.1 (ii), Algorithm 2 induces the accelerated versions of first-order algorithms such as the steepest descent [11], proximal point [6], and projected gradient methods [15].

(v) Like Remark 3.1 (iii), even if it is difficult to estimate L, we can update the constant ℓ to satisfy $F_i(p_\ell^{\mathrm{acc}}(x^{k-1}, y^k)) - F_i(x^{k-1}) \leq \theta_\ell^{\mathrm{acc}}(x^{k-1}, y^k)$ for all $i = 1, \dots, m$ in each iteration by a finite number of backtracking steps.

5 Convergence rate

This section shows that Algorithm 2 has a convergence rate of $O(1/k^2)$, which is better than Algorithm 1.

For $k \geq 0$, let $\sigma_k \colon \mathbf{R}^n \to \mathbf{R} \cup \{-\infty\}$ and $\rho_k \colon \mathbf{R}^n \to \mathbf{R}$ be defined by

$$\sigma_k(z) \coloneqq \min_{i=1,\dots,m} \left\{ F_i(x^k) - F_i(z) \right\}, \tag{23a}$$

$$\rho_k(z) \coloneqq \left\| t_{k+1} x^{k+1} - (t_{k+1} - 1) x^k - z \right\|^2, \tag{23b}$$

respectively. We present a helpful lemma on σ_k for the subsequent discussion.

Lemma 5.1. *The following inequalities hold for all $z \in \mathbf{R}^n$ and $k \geq 0$:*

$$\begin{aligned} \sigma_{k+1}(z) \leq -\frac{\ell}{2} \left\{ 2\left\langle x^{k+1} - y^{k+1}, y^{k+1} - z \right\rangle + \left\| x^{k+1} - y^{k+1} \right\|^2 \right\} \\ - \frac{\ell - L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2, \end{aligned} \tag{24}$$

$$\begin{aligned} \sigma_k(z) - \sigma_{k+1}(z) \geq \frac{\ell}{2} \left\{ 2\left\langle x^{k+1} - y^{k+1}, y^{k+1} - x^k \right\rangle + \left\| x^{k+1} - y^{k+1} \right\|^2 \right\} \\ + \frac{\ell - L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2. \end{aligned} \tag{25}$$

Proof. Suppose that $z \in \mathbf{R}^n$ and $k \geq 0$. Recall that there exist $\eta(x^k, y^{k+1}) \in \partial g(x^{k+1})$ and a Lagrange multiplier $\lambda(x^k, y^{k+1}) \in \mathbf{R}^m$ that satisfy the KKT condition (18) for the subproblem (15). From the definition (23a) of σ_{k+1}, we get

$$\sigma_{k+1}(z) = \min_{i=1,\dots,m} \left\{F_i(x^{k+1}) - F_i(z)\right\} \leq \sum_{i=1}^{m} \lambda_i(x^k, y^{k+1}) \left\{F_i(x^{k+1}) - F_i(z)\right\}.$$

where the equality follows from (18b). Taking $p = x^{k+1}, q = y^{k+1}$, and $r = z$ in (7), we have

$$\begin{aligned}\sigma_{k+1}(z) \leq \sum_{i=1}^{m} \lambda_i(x^k, y^{k+1}) \big\{\big\langle \nabla f_i(y^{k+1}), x^{k+1} - y^{k+1}\big\rangle + g_i(x^{k+1}) \\ + f_i(y^{k+1}) - F_i(z)\big\} + \frac{L}{2}\big\|x^{k+1} - y^{k+1}\big\|^2.\end{aligned}$$

Hence, the convexity of f_i and g_i yields

$$\begin{aligned}&\sigma_{k+1}(z) \\ &\leq \sum_{i=1}^{m} \lambda_i(x^k, y^{k+1}) \big\{\big\langle \nabla f_i(y^{k+1}), x^{k+1} - y^{k+1}\big\rangle + \big\langle \nabla f_i(y^{k+1}), y^{k+1} - z\big\rangle \\ &\qquad + \big\langle \eta_i(x^k, y^{k+1}), x^{k+1} - z\big\rangle\big\} + \frac{L}{2}\big\|x^{k+1} - y^{k+1}\big\|^2 \\ &= \sum_{i=1}^{m} \lambda_i(x^k, y^{k+1}) \big\langle \nabla f_i(y^{k+1}) + \eta_i(x^k, y^{k+1}), x^{k+1} - z\big\rangle + \frac{L}{2}\big\|x^{k+1} - y^{k+1}\big\|^2.\end{aligned}$$

Using (18a) with $x = x^k$ and $y = y^{k+1}$, we obtain

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

Moreover, simple calculations show that

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

which completes the proof of (24).

Now, let us show inequality (25). Again from the definition (23a) of σ_k, we obtain

$$\begin{aligned}\sigma_k(z) - \sigma_{k+1}(z) &= \min_{i=1,\dots,m} \left\{F_i(x^k) - F_i(z)\right\} - \min_{i=1,\dots,m} \left\{F_i(x^{k+1}) - F_i(z)\right\} \\ &\geq -\max_{i=1,\dots,m} \left\{F_i(x^{k+1}) - F_i(x^k)\right\}\end{aligned}.$$

Letting $p = x^{k+1}, q = x^k$, and $r = y^{k+1}$ in (7), we have

$$\begin{aligned}
&\sigma_k(z) - \sigma_{k+1}(z) \\
&\quad \geq - \max_{i=1,\dots,m} \left\{ \left\langle \nabla f_i(y^{k+1}), x^{k+1} - y^{k+1} \right\rangle + g_i(x^{k+1}) + f_i(y^{k+1}) \right. \\
&\qquad\qquad\qquad\qquad\qquad \left. - F_i(x^k) \right\} - \frac{L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 \\
&\quad = - \sum_{i=1}^{m} \lambda_i(x^k, y^{k+1}) \left\{ \left\langle \nabla f_i(y^{k+1}), x^{k+1} - y^{k+1} \right\rangle + g_i(x^{k+1}) \right. \\
&\qquad\qquad\qquad \left. + f_i(y^{k+1}) - F_i(x^k) \right\} - \frac{L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 \\
&\quad = - \sum_{i=1}^{m} \lambda_i(x^k, y^{k+1}) \left\{ \left\langle \nabla f_i(y^{k+1}), x^k - y^{k+1} \right\rangle + f_i(y^{k+1}) - f_i(x^k) \right\} \\
&\qquad - \sum_{i=1}^{m} \lambda_i(x^k, y^{k+1}) \left\{ \left\langle \nabla f_i(y^{k+1}), x^{k+1} - x^k \right\rangle + g_i(x^{k+1}) - g_i(x^k) \right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad - \frac{L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2,
\end{aligned}$$

where the first equality comes from (18b), and the second one follows by taking $x^{k+1} - y^{k+1} = (x^k - y^{k+1}) + (x^{k+1} - x^k)$. From the convexity of f_i, the first term of the above expression is nonnegative. Moreover, the convexity of g_i shows that

$$\begin{aligned}
&\sigma_k(z) - \sigma_{k+1}(z) \\
&\geq - \sum_{i=1}^{m} \lambda_i(x^k, y^{k+1}) \left\langle \nabla f_i(y^{k+1}) + \eta_i(x^k, y^{k+1}), x^{k+1} - x^k \right\rangle - \frac{L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2.
\end{aligned}$$

Thus, (18a) with $(x, y) = (x^k, y^{k+1})$ and direct calculations prove that

$$\begin{aligned}
&\sigma_k(z) - \sigma_{k+1}(z) \\
&\geq \ell \left\langle x^{k+1} - y^{k+1}, x^{k+1} - x^k \right\rangle - \frac{L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 \\
&= \frac{\ell}{2} \left\{ 2 \left\langle x^{k+1} - y^{k+1}, x^{k+1} - x^k \right\rangle - \left\| x^{k+1} - y^{k+1} \right\|^2 \right\} + \frac{\ell - L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2 \\
&= \frac{\ell}{2} \left\{ 2 \left\langle x^{k+1} - y^{k+1}, y^{k+1} - x^k \right\rangle + \left\| x^{k+1} - y^{k+1} \right\|^2 \right\} + \frac{\ell - L}{2} \left\| x^{k+1} - y^{k+1} \right\|^2.
\end{aligned}$$

□

The following corollary of Lemma 5.1 (25) is also worth mentioning.

Corollary 5.1. *Let $k_2 \geq k_1 \geq 1$. Then, we have*

$$\begin{aligned}
&\sigma_{k_1}(z) - \sigma_{k_2}(z) \\
&\qquad \geq \frac{\ell}{2} \left\{ \left\| x^{k_2} - x^{k_2 - 1} \right\|^2 - \left\| x^{k_1} - x^{k_1 - 1} \right\|^2 + \sum_{k=k_1}^{k_2 - 1} \frac{1}{t_k} \left\| x^k - x^{k-1} \right\|^2 \right\}.
\end{aligned}$$

Proof. Let $k \geq 1$. Since $\ell \geq L$, Lemma 5.1 (25) implies

$$\begin{aligned}\sigma_k(z) - \sigma_{k+1}(z) &\geq \frac{\ell}{2}\left\{2\langle x^{k+1} - y^{k+1}, y^{k+1} - x^k\rangle + \left\|x^{k+1} - y^{k+1}\right\|^2\right\}\\ &= \frac{\ell}{2}\left\{\left\|x^{k+1} - x^k\right\|^2 - \left\|y^{k+1} - x^k\right\|^2\right\}.\end{aligned}$$

Hence,the definition of y^{k+1} given in Line 7 of Algorithm 2 yields

$$\sigma_k(z) - \sigma_{k+1}(z) \geq \frac{\ell}{2}\left\{\left\|x^{k+1} - x^k\right\|^2 - \gamma_k^2\left\|x^k - x^{k-1}\right\|^2\right\}.$$

Applying this inequality repeatedly, we have

$$\begin{aligned}&\sigma_{k_1}(z) - \sigma_{k_2}(z)\\ &\geq \frac{\ell}{2}\left\{\left\|x^{k_2} - x^{k_2-1}\right\|^2 - \left\|x^{k_1} - x^{k_1-1}\right\|^2 + \sum_{k=k_1}^{k_2-1}\left(1 - \gamma_k^2\right)\left\|x^k - x^{k-1}\right\|^2\right\}.\end{aligned}$$

Using Lemma 4.1 (iii), we get the desired inequality. □

Before analyzing the convergence rate of Algorithm 2, we show that the objective function values at x^k for any $k \geq 0$ never exceed the ones at the initial point, that is, $\left\{x^k\right\}$ belongs to the level set $\mathcal{L}_F(F(x^0))$ (see (9) for the definition of $\mathcal{L}$). Note that Algorithm 2 does not guarantee the monotonically decreasing property $F(x^{k+1}) \leq F(x^k)$.

Theorem 5.1. *Algorithm 2 generates a sequence $\left\{x^k\right\}$ such that*

$$F_i(x^k) \leq F_i(x^0) \quad \text{for all } i = 1, \dots, m, k \geq 0.$$

Proof. Let $i = 1, \dots, m$ and $p \geq 1$. Then, we have

$$F_i(x^p) - F_i(x^{p+1}) \geq -\max_{i=1,\dots,m}\left\{F_i(x^{p+1}) - F_i(x^p)\right\}.$$

With similar arguments used in the proof of (25) in Lemma 5.1, we obtain

$$\begin{aligned}F_i(x^p) - F_i(x^{p+1}) \geq \frac{\ell}{2}&\left\{2\langle x^{p+1} - y^{p+1}, y^{p+1} - x^p\rangle + \left\|x^{p+1} - y^{p+1}\right\|^2\right\}\\ &+ \frac{\ell - L}{2}\left\|x^{p+1} - y^{p+1}\right\|^2. \quad (26)\end{aligned}$$

Note that this inequality also holds for $p = 0$. Again, in the same way as in the proof of Corollary 5.1, we get

$$F_i(x^k) - F_i(x^1) \geq \frac{\ell}{2}\left\{\left\|x^k - x^{k-1}\right\|^2 - \left\|x^1 - x^0\right\|^2 + \sum_{p=1}^{k-1}\frac{1}{t_p}\left\|x^p - x^{p-1}\right\|^2\right\}.$$

Since $t_1 = 1$, the above inequality reduces to

$$F_i(x^k) - F_i(x^1) \geq \frac{\ell}{2} \left\{ \left\| x^k - x^{k-1} \right\|^2 + \sum_{p=2}^{k-1} \frac{1}{t_p} \left\| x^p - x^{p-1} \right\|^2 \right\}.$$

Moreover, (26) with $p = 0$ and the fact that $y^1 = x^0$ imply $F_i(x^1) \leq F_i(x^0)$, so we can show that $F_i(x^k) \leq F_i(x^0)$. □

The following result provides the fundamental relation for our convergence rate analysis.

Lemma 5.2. *Let σ_k and ρ_k be defined by* (23). *Then, we have*

$$t_{k+1}^2 \sigma_{k+1}(z) + \frac{\ell}{2} \rho_k(z) + \frac{\ell - L}{2} \sum_{p=1}^{k} t_{p+1}^2 \left\| x^{p+1} - y^{p+1} \right\|^2 \leq \frac{\ell}{2} \left\| x^0 - z \right\|^2$$

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

Proof. Let $p \geq 1$ and $z \in \mathbf{R}^n$. Recall from Lemma 5.1 that

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

To get a relation between $\sigma_p(z)$ and $\sigma_{p+1}(z)$, we multiply the first inequality above by $(t_{p+1} - 1)$ and add it to the second one:

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

Multiplying this inequality by t_{p+1} and using the relation $t_p^2 = t_{p+1}^2 - t_{p+1}$ (cf. Lemma 4.1 (ii)), we get

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

Applying the usual Pythagoras relation

$$\|b-a\|^2 + 2\langle b-a, a-c\rangle = \|b-c\|^2 - \|a-c\|^2$$

to the right-hand side of the last inequality with

$$a \coloneqq t_{p+1}y^{p+1}, \quad b \coloneqq t_{p+1}x^{p+1}, \quad c \coloneqq (t_{p+1}-1)x^p + z,$$

we get

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

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

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

Now, let $k \geq 0$. Adding the above inequality from $p = 0$ to $p = k$ and using $t_1 = 1$ and $\rho_0(z) = \left\|x^1 - z\right\|^2$, we have

$$\sigma_1(z) - t_{k+1}^2\sigma_{k+1}(z) \geq \frac{\ell}{2}\left\{\rho_k(z) - \left\|x^1 - z\right\|^2\right\} + \frac{\ell - L}{2}\sum_{p=1}^{k} t_{k+1}^2\left\|x^{k+1} - y^{k+1}\right\|^2. \tag{27}$$

Since Lemma 4.1 (24) with $k = 0$ and $y^1 = x^0$ lead to

$$\begin{aligned} \sigma_1(z) &\leq -\frac{\ell}{2}\left\{\left\|x^1 - z\right\|^2 - \left\|x^0 - z\right\|^2\right\} - \frac{\ell - L}{2}\left\|x^1 - y^1\right\|^2 \\ &\leq -\frac{\ell}{2}\left\{\left\|x^1 - z\right\|^2 - \left\|x^0 - z\right\|^2\right\}, \end{aligned}$$

where the second inequality follows since $L \geq \ell$. From the above two inequalities, we can derive the desired inequality. □

Using Lemma 5.2, we can evaluate the convergence rate of Algorithm 2 with the following theorem:

Theorem 5.2. *Under Assumption 3.1, Algorithm 2 generates a sequence* $\left\{x^k\right\}$ *such that*

$$u_0(x^k) \leq \frac{2\ell R}{(k+1)^2},$$

where $R \geq 0$ *is given in* (14)*, and* u_0 *is a merit function defined by* (10)*.*

Proof. Let $k \geq 0$. Since $\rho_k(z) \geq 0$, Lemma 5.2 gives

$$t_{k+1}^2 \sigma_{k+1}(z) \leq \frac{\ell}{2} \left\| x^0 - z \right\|^2.$$

It follows from Lemma 4.1 (i), we have

$$\sigma_{k+1}(z) \leq \frac{2\ell}{(k+2)^2} \left\| x^0 - z \right\|^2.$$

With similar arguments used in the proof of Theorem 3.1 (see [23, Theorem 5.2]), we get the desired inequality □

6 Efficient computation of the subproblem via its dual

This section discusses the way of computing the subproblem (15). Define

$$\psi_i(z; x, y) \coloneqq \langle \nabla f_i(y), z - y \rangle + g_i(z) + f_i(y) - F_i(x) + \frac{\ell}{2} \|z - y\|^2 \tag{28}$$

for all $i = 1, \ldots, m$. Then, we can rewrite the objective function $\varphi_\ell^{\mathrm{acc}}(z; x, y)$ of (15) as

$$\varphi_\ell^{\mathrm{acc}}(z; x, y) = \max_{i=1,\ldots,m} \psi_i(z; x, y).$$

Recall that $\Delta^m \subseteq \mathbf{R}^m$ represents the standard simplex (1). Since $\max_{i=1,\ldots,m} q_i = \max_{\lambda \in \Delta^m} \sum_{i=1}^m \lambda_i q_i$ for any $q \in \mathbf{R}^m$, we get

$$\varphi_\ell^{\mathrm{acc}}(z; x, y) = \max_{\lambda \in \Delta^m} \sum_{i=1}^m \lambda_i \psi_i(z; x, y).$$

Then, the subproblem (15) reduces to the following minimax problem:

$$\min_{z \in \mathbf{R}^n} \max_{\lambda \in \Delta^m} \quad \sum_{i=1}^m \lambda_i \psi_i(z; x, y).$$

We can see that $\mathbf{R}^n$ is convex, Δ^m is compact and convex, and $\sum_{i=1}^m \lambda_i \psi_i(z; x, y)$ is convex for z and concave for λ. Therefore, Sion's minimax theorem [20] shows that the above problem is equivalent to

$$\max_{\lambda \in \Delta^m} \min_{z \in \mathbf{R}^n} \quad \sum_{i=1}^m \lambda_i \psi_i(z; x, y).$$

The definition (28) of ψ_i yields

$$\begin{aligned}\min_{z\in\mathbf{R}^n}\sum_{i=1}^m \lambda_i\psi_i(z;x,y) = {} & \min_{z\in\mathbf{R}^n}\left\{\sum_{i=1}^m \lambda_i g_i(z) + \frac{\ell}{2}\left\| z - y + \frac{1}{\ell}\sum_{i=1}^m \lambda_i \nabla f_i(y)\right\|^2\right\} \\ & - \frac{1}{2\ell}\left\|\sum_{i=1}^m \lambda_i\nabla f_i(y)\right\|^2 + \sum_{i=1}^m \lambda_i\left\{f_i(y) - F_i(x)\right\} \\ = {} & \ell\,\mathcal{M}_{\frac{1}{\ell}\sum\limits_{i=1}^m \lambda_i g_i}\left(y - \frac{1}{\ell}\sum_{i=1}^m \lambda_i\nabla f_i(y)\right) \\ & - \frac{1}{2\ell}\left\|\sum_{i=1}^m \lambda_i\nabla f_i(y)\right\|^2 + \sum_{i=1}^m \lambda_i\left\{f_i(y) - F_i(x)\right\},\end{aligned}$$

where $\mathcal{M}$ is the Moreau envelope (2). Based on the discussion above, we obtain the dual problem of (15) as follows:

$$\begin{aligned} \max_{\lambda\in\mathbf{R}^m} \quad & \omega(\lambda) \\ \text{s.t.} \quad & \lambda \geq 0 \quad \text{and} \quad \sum_{i=1}^m \lambda_i = 1, \end{aligned} \tag{29}$$

where

$$\begin{aligned}\omega(\lambda) := {} & \ell\,\mathcal{M}_{\frac{1}{\ell}\sum\limits_{i=1}^m \lambda_i g_i}\left(y - \frac{1}{\ell}\sum_{i=1}^m \lambda_i\nabla f_i(y)\right) \\ & - \frac{1}{2\ell}\left\|\sum_{i=1}^m \lambda_i\nabla f_i(y)\right\|^2 + \sum_{i=1}^m \lambda_i\left\{f_i(y) - F_i(x)\right\}.\end{aligned} \tag{30}$$

If we can find the global optimal solution λ^* of this dual problem (29), we ca n construct the optimal solution z^* of the original subproblem (15) as

$$z^* = \mathbf{prox}_{\frac{1}{\ell}\sum\limits_{i=1}^m \lambda_i^* g_i}\left(y - \frac{1}{\ell}\sum_{i=1}^m \lambda_i^*\nabla f_i(y)\right),$$

where **prox** denotes the proximal operator (3). Since $\sum_{i=1}^m \lambda_i\psi_i(z;x,y)$ is concave for λ, it is clear that $\omega(\lambda) = \min_{z\in\mathbf{R}^n}\lambda_i\psi_i(z;x,y)$ is concave. Furthermore, ω is differentiable, as the following theorem shows.

Theorem 6.1. *The function $\omega\colon \mathbf{R}^m \to \mathbf{R}$ defined by* (30) *is continuously differentiable at every $\lambda \in \mathbf{R}^m$ and*

$$\begin{aligned}\nabla\omega(\lambda) = {} & g\left(\mathbf{prox}_{\frac{1}{\ell}\sum\limits_{i=1}^m \lambda_i g_i}\left(y - \frac{1}{\ell}\sum_{i=1}^m \lambda_i\nabla f_i(y)\right)\right) \\ & + J_f(y)\left(\mathbf{prox}_{\frac{1}{\ell}\sum\limits_{i=1}^m \lambda_i g_i}\left(y - \frac{1}{\ell}\sum_{i=1}^m \lambda_i\nabla f_i(y)\right) - y\right) + f(y) - F(x),\end{aligned}$$

where **prox** *is the proximal operator* (3)*, and* $J_f(y)$ *is the Jacobian matrix given by*

$$J_f(y) \coloneqq (\nabla f_1(y), \dots, \nabla f_m(y))^\top .$$

Proof. Define

$$h(z, \lambda) \coloneqq \sum_{i=1}^m \lambda_i g_i(z) + \frac{\ell}{2} \left\| z - y + \frac{1}{\ell} \sum_{i=1}^m \lambda_i \nabla f_i(y) \right\|^2 .$$

Clearly, h is continuous on $\mathbf{R}^n \times \mathbf{R}^m$. Moreover, $h_x(\cdot) \coloneqq h(z, \lambda)$ is continuously differentiable and

$$\nabla_\lambda h_x(\lambda) = g(z) + J_f(y) \left(z - y + \frac{1}{\ell} \sum_{i=1}^m \lambda_i \nabla f_i(y) \right) .$$

Furthermore,

$$\mathbf{prox}_{\frac{1}{\ell} \sum\limits_{i=1}^m \lambda_i g_i} \left(y - \frac{1}{\ell} \sum_{i=1}^m \lambda_i \nabla f_i(y) \right) = \operatorname*{argmin}_{z \in \mathbf{R}^n} h(z, \lambda)$$

is also continuous at every $\lambda \in \mathbf{R}^m$ (cf. [19, Theorem 2.26 and Exercise 7.38]). Therefore, the well-known result in first order differentiability analysis of the optimal value function [5, Theorem 4.13] gives

$$\begin{aligned}
&\nabla_\lambda \left[\ell \mathcal{M}_{\frac{1}{\ell} \sum\limits_{i=1}^m \lambda_i g_i} \left(y - \frac{1}{\ell} \sum_{i=1}^m \lambda_i \nabla f_i(y) \right) \right] \\
&\quad = g \left(\mathbf{prox}_{\frac{1}{\ell} \sum\limits_{i=1}^m \lambda_i g_i} \left(y - \frac{1}{\ell} \sum_{i=1}^m \lambda_i \nabla f_i(y) \right) \right) \\
&\qquad + J_f(y) \left(\mathbf{prox}_{\frac{1}{\ell} \sum\limits_{i=1}^m \lambda_i g_i} \left(y - \frac{1}{\ell} \sum_{i=1}^m \lambda_i \nabla f_i(y) \right) - y + \frac{1}{\ell} \sum_{i=1}^m \lambda_i \nabla f_i(y) \right) .
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&\nabla_\lambda \left[-\frac{1}{2\ell} \left\| \sum_{i=1}^m \lambda_i \nabla f_i(y) \right\|^2 + \sum_{i=1}^m \lambda_i \left\{ f_i(y) - F_i(x) \right\} \right] \\
&\qquad = -\frac{1}{\ell} J_f(y) \sum_{i=1}^m \lambda_i \nabla f_i(y) + f(y) - F(x).
\end{aligned}$$

Adding the above two equalities, we get the desired result. □

This theorem shows that the dual problem (29) is an m-dimensional differentiable convex optimization problem. Thus, if we can compute the proximal operator of $\sum_{i=1}^m \lambda_i g_i$ quickly, then we can solve (29) using convex optimization techniques such as the interior point method [7]; for cases where $n \gg m$, the computational cost is much lower than solving the subproblem (15) directly. Note, for example, that if $g_i(x) = g_1(x)$ for all $i = 1, \dots, m$, or if $g_i(x) = g_i(x_{I_i})$ and the index sets I_i do not overlap each other, then we can evaluate the proximal operator of $\sum_{i=1}^m \lambda_i g_i$ from the proximal operator of each g_i.

7 Numerical experiments

This section illustrates the proposed methods' performance compared to the proximal gradient methods without acceleration [21]. The experiments are carried out on a machine with a 2.3 GHz Intel Core i7 CPU and 32 GB memory, and we implement all codes in Python 3.9.9. We solve all test problems using Algorithms 1 (Proximal Gradient Method; PGM) and 2 (Accelerated Proximal Gradient Method; Acc-PGM) with backtracking. In both algorithms, we converted the subproblem into a dual problem as discussed in Section 6 and solved it using the trust-region interior point method [8] with the scientific library SciPy. We set $\varepsilon = 10^{-5}$ for the stopping criteria in each experiment. Moreover, we chose 1000 initial points, commonly for both algorithms, and randomly with a uniform distribution between $\underline{b}$ and $\bar{b}$.

Experiment 1 (bi-objective)

In the first experiment, we solve the test problem from [17] with the following objective functions:

$$f_1(x) = \frac{1}{n}\|x\|^2, f_2(x) = \frac{1}{n}\|x-2\|^2, g_1(x) = g_2(x) = 0. \tag{31}$$

We set the dimension of x to be $n = 50$ and the lower and upper bounds of the initial points to be $\underline{b} = (-2, \dots, -2)^\top$ and $\bar{b} = (4, \dots, 4)^\top$, respectively. Note that since $\mathbf{prox}_{\lambda_1 g_1 + \lambda_2 g_2}(x) = x$ for all $x \in \mathbf{R}^n$ and $(\lambda_1, \lambda_2) \in \mathbf{R}^2$, we can solve the subproblems quickly by using Theorem 6.1.

Figure 1 plots the objective function values for $k = 0$ (i.e., at the initial points), $k = 10$, and the terminal points of each algorithm, respectively. We can see from Figure 1 that both algorithms bring a wide range of Pareto solutions. However, the objective function values at $k = 10$ are smaller when using Algorithm 2, which shows the efficiency of the proposed acceleration. Moreover, from Table 1, we see that Algorithm 2 converges faster than Algorithm 1.

Table 1: Average computational costs for solving (31)

	PGM (Algorithm 1)	Acc-PGM (Algorithm 2)
Execution time	10.51 s	2.84 s
Iteration counts	232.0	65.0

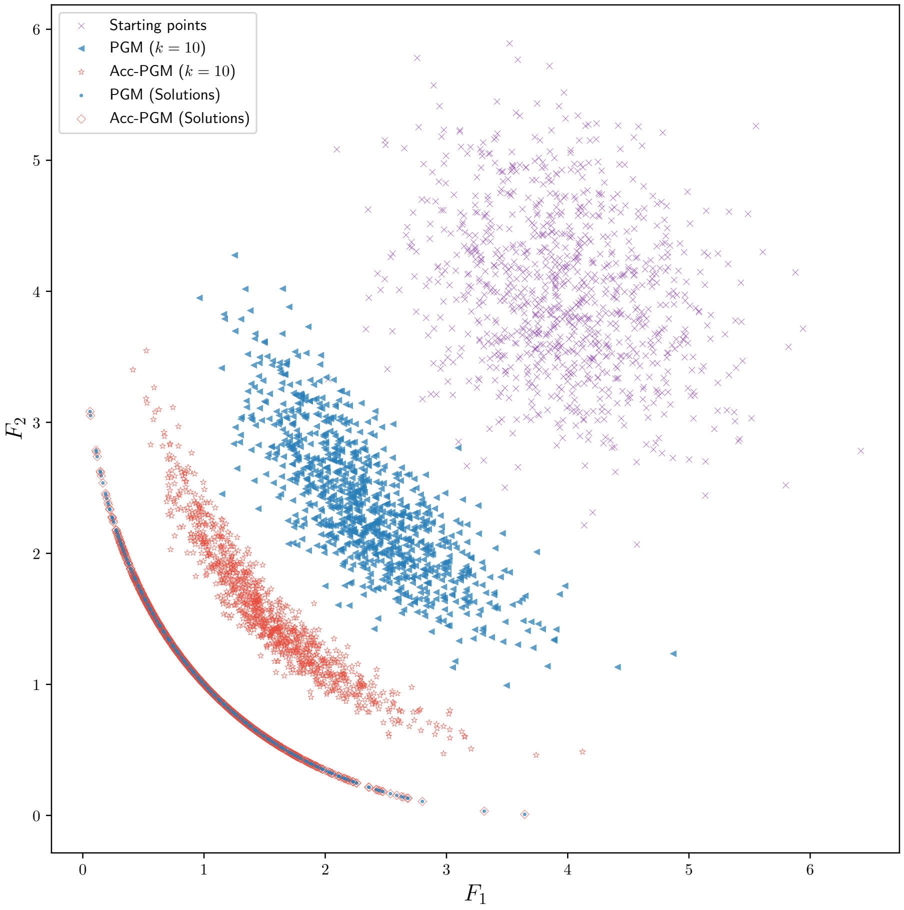

Figure 1: Objective function values obtained by Algorithm 1 (PGM) and 2 (Acc-PGM) for (31)

Moreover, using the same settings, we solve the following problem with additional non-differentiable terms to (31):

$$f_1(x) = \frac{1}{n}\|x\|^2, f_2(x) = \frac{1}{n}\|x-2\|^2, g_1(x) = \frac{1}{n}\|x\|_1, g_2(x) = \frac{1}{2n}\|x-1\|_1. \tag{32}$$

For this problem, we can also express the proximal operator of $\lambda_1 g_1 + \lambda_2 g_2$ for $\lambda \in R^2$ explicitly as

$$\mathbf{prox}_{\lambda_1 g_1 + \lambda_2 g_2}(x) = \mathcal{O}_{\lambda_2/(2n)}\left(\mathcal{O}_{\lambda_1/n}\left(x + \frac{\lambda_2}{2n}\right) - \frac{\lambda_2}{2n} - 1\right) + 1,$$

where $\mathcal{O}_\tau$ with $\tau > 0$ is the soft-thresholding operator defined by

$$\mathcal{O}_\tau(x) = \begin{cases} x - \tau & x \geq \tau, \\ 0 & -\tau < x < \tau, \\ x + \tau & x \leq -\tau. \end{cases}$$

Like (31), we plot the objective function values for $k = 0, 10$ and the terminal points in Figure 2 and summarize the average computational costs in Table 2. Then, we can still observe that Algorithm 2 achieves faster convergence to the Pareto frontier.

Table 2: Average computational costs for solving (32)

	PGM (Algorithm 1)	Acc-PGM (Algorithm 2)
Execution time	9.90 s	7.24 s
Iteration counts	219.0	161.2

Experiment 2 (tri-objective)

The second experiment deals with the tri-objective problem from [10] with the following objective functions:

$$\begin{gathered} f_1(x) = \frac{1}{n^2}\sum_{i=1}^n i(x_i - i)^4, f_2(x) = \exp\left(\sum_{i=1}^n \frac{x_i}{n}\right) + \|x\|^2, \\ f_3(x) = \frac{1}{n(n+1)}\sum_{i=1}^n i(n-i+1)\exp(-x_i), \\ g_1(x) = g_2(x) = g_3(x) = 0. \end{gathered} \tag{33}$$

We set $n = 50, \underline{b} = (-2, \dots, -2)^\top$ and $\bar{b} = (2, \dots, 2)^\top$. Recall that since $g_1 = g_2 = g_3 = 0$, we have $\mathbf{prox}_{\lambda_1 g_1 + \lambda_2 g_2 + \lambda_3 g_3}(x) = x$. Figure 3 and Table 3 show that Algorithm 2 reaches as diverse Pareto frontier as Algorithm 1 but needs

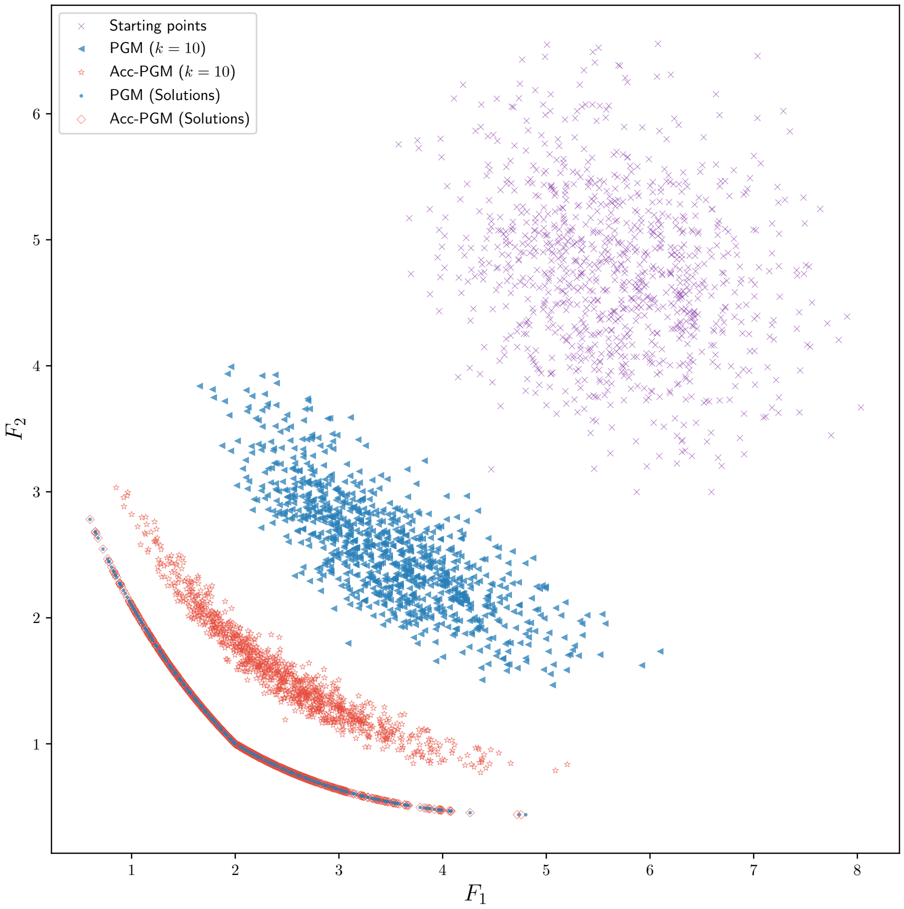

Figure 2: Objective function values obtained by Algorithms 1 (PGM) and 2 (Acc-PGM) for (32)

Table 3: Average computational costs for solving (33)

	PGM (Algorithm 1)	Acc-PGM (Algorithm 2)
Execution time	79.17 s	29.27 s
Iteration counts	639.9	247.1

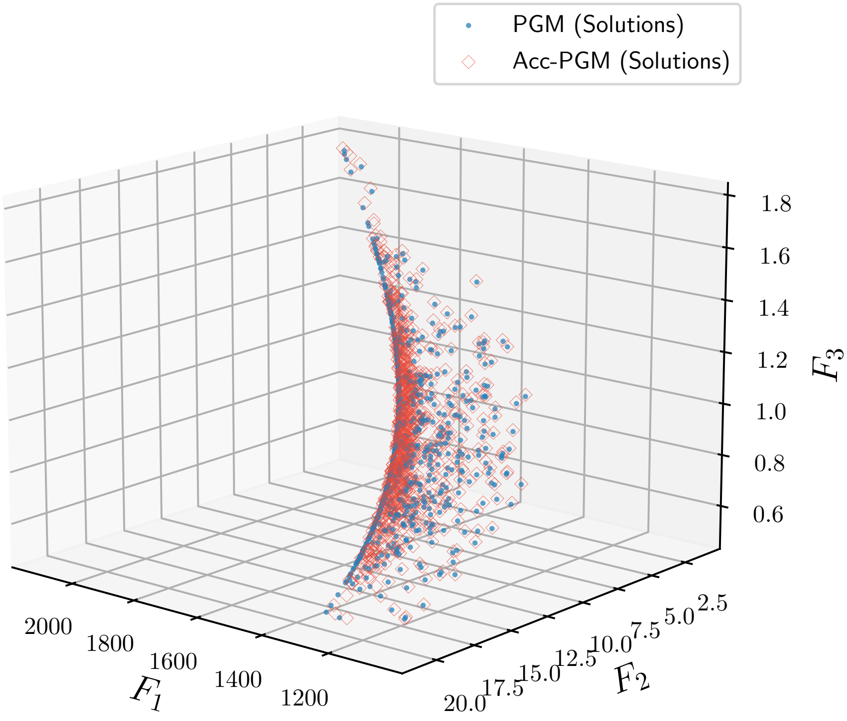

Figure 3: Objective function values obtained by Algorithm 1 (PGM) and 2 (Acc-PGM) for (33)

less computational cost. Note that, unlike Experiment 1, we plot only the final solutions in Figure 3 to improve visibility.

To demonstrate the effectiveness for constrained problems, under the setting $n = 50, \underline{b} = (0, \dots, 0)^\top$, and $\bar{b} = (2, \dots, 2)^\top$, we solve the following modified problem:

$$\begin{gathered} f_1(x) = \frac{1}{n^2} \sum_{i=1}^{n} i(x_i - i)^4, f_2(x) = \exp\left(\sum_{i=1}^{n} \frac{x_i}{n}\right) + \|x\|^2, \\ f_3(x) = \frac{1}{n(n+1)} \sum_{i=1}^{n} i(n - i + 1) \exp(-x_i), \\ g_1(x) = g_2(x) = g_3(x) = \chi_{\mathbf{R}^n_+}(x), \end{gathered} \tag{34}$$

where $\chi_{\mathbf{R}^n_+}$ denotes the indicator function (4) of the nonnegative orthant. As we mentioned in Remark 2.1 (i), the proximal operator of $\lambda_1 g_1 + \lambda_2 g_2 + \lambda_3 g_3 = \chi_{\mathbf{R}^n_+}$ reduces to the projection onto $\mathbf{R}^n_+$, i.e.,

$$\mathbf{prox}_{\lambda_1 g_1 + \lambda_2 g_2 + \lambda_3 g_3}(x) = \max\left\{x, 0\right\},$$

where the max operator is taken componentwise. As we can observe from Figure 4 and Table 4, Algorithm 2 still obtains part of the Pareto frontier and is faster.

Table 4: Average computational costs for solving (34)

	PGM (Algorithm 1)	Acc-PGM (Algorithm 2)
Execution time	152.70 s	36.84 s
Iteration counts	1066.2	275.4

8 Conclusion

By putting the information of the previous points into the subproblem, we have successfully accelerated the gradient method for multiobjective optimization and proved its convergence rate under natural assumptions, which was an open problem. Moreover, we showed the efficient way of computing the subproblem via its dual. As suggested by experiments, the proposed methods are also effective from the numerical point of view.

Since many methods for single-objective optimization had been developed following the idea of Nesterov's acceleration technique, this research may also contribute to the development of various multiobjective optimization methods. Such extensions will be subjects of future works.

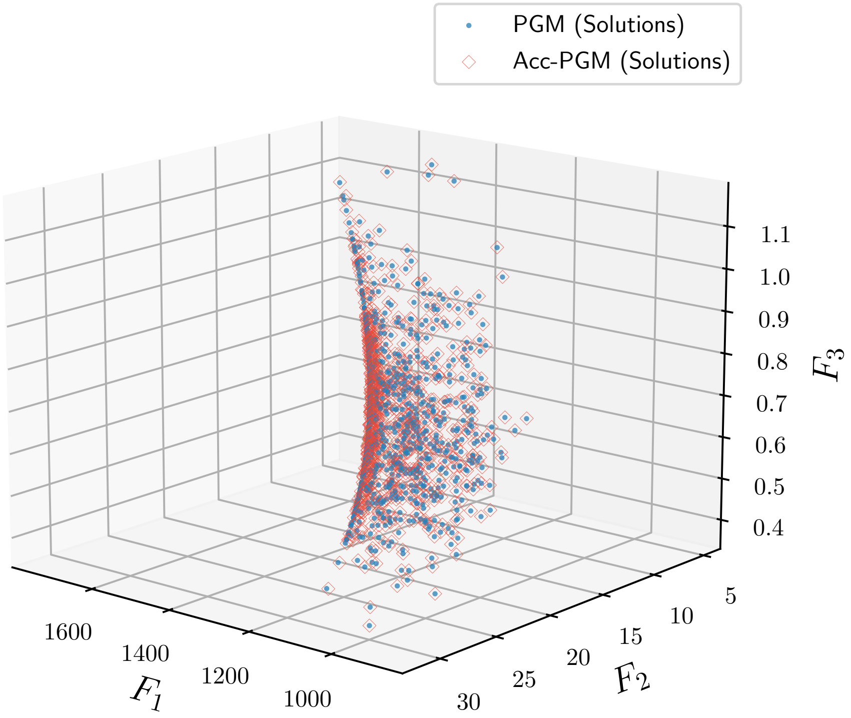

Figure 4: Objective function values obtained by Algorithms 1 (PGM) and 2 (Acc-PGM) for (34)

Acknowledgements

This work was supported by the Grant-in-Aid for Scientific Research (C) (17K00032 and 19K11840) and Grant-in-Aid for JSPS Fellows (20J21961) from the Japan Society for the Promotion of Science.

References

[1] Beck, A.: *First-Order Methods in Optimization*, Society for Industrial and Applied Mathematics, 2017.

[2] Beck, A. and Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences*, Vol. 2 (2009), 183–202.

[3] Berge, C.: *Topological spaces*, Dover Publications, Edinburgh, 1963.

[4] Bertsekas, D. P.: Nonlinear Programming, Athena Scientific, Belmont, Mass., second edition, 1999.

[5] Bonnans, J. F. and Shapiro, A.: *Perturbation Analysis of Optimization Problems*, Springer New York, 2000.

[6] Bonnel, H., Iusem, A. N. and Svaiter, B. F.: Proximal methods in vector optimization, *SIAM Journal on Optimization*, Vol. 15 (2005), 953–970.

[7] Boyd, S. and Vandenberghe, L.: *Convex Optimization*, Cambridge University Press, 2017.

[8] Byrd, R. H., Hribar, M. E. and Nocedal, J.: An interior point algorithm for large-scale nonlinear programming, *SIAM Journal on Optimization*, Vol. 9 (1999), 877–900.

[9] El Moudden, M. and El Mouatasim, A.: Accelerated diagonal steepest descent method for unconstrained multiobjective optimization, *Journal of Optimization Theory and Applications*, Vol. 188 (2021), 220–242.

[10] Fliege, J., Graña Drummond, L. M. and Svaiter, B. F.: Newton's method for multiobjective optimization, *SIAM Journal on Optimization*, Vol. 20 (2009), 602–626.

[11] Fliege, J. and Svaiter, B. F.: Steepest descent methods for multicriteria optimization, *Mathematical Methods of Operations Research*, Vol. 51 (2000), 479–494.

[12] Fliege, J., Vaz, A. I. and Vicente, L. N.: Complexity of gradient descent for multiobjective optimization, *Optimization Methods and Software*, Vol. 34 (2019), 949–959.

[13] Gass, S. and Saaty, T.: The computational algorithm for the parametric objective function, *Naval Research Logistics Quarterly*, Vol. 2 (1955), 39–45.

[14] Geoffrion, A. M.: Proper efficiency and the theory of vector maximization, *Journal of Mathematical Analysis and Applications*, Vol. 22 (1968), 618–630.

[15] Graña Drummond, L. M. and Iusem, A. N.: A projected gradient method for vector optimization problems, *Computational Optimization and Applications*, Vol. 28 (2004), 5–29.

[16] Hogan, W. W.: Point-to-Set Maps in Mathematical Programming, *SIAM Review*, Vol. 15 (1973), 591–603.

[17] Jin, Y., Olhofer, M. and Sendhoff, B.: 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, San Francisco, CA, USA, 2001, Morgan Kaufmann Publishers Inc.

[18] Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$, *Dokl. Akad. Nauk SSSR*, Vol. 269 (1983), 543–547.

[19] Rockafellar, R. T. and Wets, R. J. B.: *Variational Analysis*, Vol. 317 of *Grundlehren der mathematischen Wissenschaften*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[20] Sion, M.: On general minimax theorems, *Pacific Journal of Mathematics*, Vol. 8 (1958), 171–176.

[21] Tanabe, H., Fukuda, E. H. and Yamashita, N.: Proximal gradient methods for multiobjective optimization and their applications, *Computational Optimization and Applications*, Vol. 72 (2019), 339–361.

[22] Tanabe, H., Fukuda, E. H. and Yamashita, N.: New merit functions and error bounds for non-convex multiobjective optimization, arXiv: 2010.09333, 2020.

[23] Tanabe, H., Fukuda, E. H. and Yamashita, N.: Convergence rates analysis of a multiobjective proximal gradient methods, arXiv: 2010.08217, 2021.

[24] Zadeh, L. A.: Optimality and Non-Scalar-Valued Performance Criteria, *IEEE Transactions on Automatic Control*, Vol. 8 (1963), 59–60.