Master's Thesis

Merit functions for multiobjective optimization and convergence rates analysis of multiobjective proximal gradient methods

Guidance

Associate Professor Ellen Hidemi FUKUDA
Professor Nobuo YAMASHITA

Hiroki TANABE

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University

July 2019

Abstract

Many descent algorithms for multiobjective optimization have been developed in the last two decades. However, differently from the scalar-valued optimization case, there exist few results related to the existence or the boundedness of Pareto optimal solutions. Moreover, studies about the convergence rates of these algorithms are still insufficient. In this paper, we first present two new merit functions for nonlinear multiobjective optimization, which extend the one defined for linear multiobjective optimization. These functions return zero at the solutions of the original problem and strictly positive values otherwise. Furthermore, by examining the properties of these merit functions, we show sufficient conditions for the existence of weakly Pareto optimal solutions, and for the boundedness of Pareto optimal sets. Finally, by using these functions, we analyze the convergence rates of the recently proposed multiobjective proximal gradient methods. We show that both methods with and without line searches have sublinear rate of convergence for non-convex and convex cases. We also prove that the algorithm without line searches converges linearly in the strongly convex case.

Contents

1 Introduction

Multiobjective optimization consists in minimizing (or maximizing) several objective functions at once. In this paper, we consider the following constrained multiobjective optimization problem:

$$\begin{aligned} \min \quad & F(x) \\ \text{s.t.} \quad & x \in S, \end{aligned} \tag{1}$$

where $F\colon S \to \mathbf{R}^m$ is a vector-valued functions with $F := (F_1, \ldots, F_m)^\top$, $S \subseteq \mathbf{R}^n$ is a nonempty closed convex set and $\top$ denotes transpose. We assume each F_i is a function from $\mathbf{R}^n$ to $\mathbf{R}$. Usually, this problem does not have a single point that minimizes all objective functions at once, so we use the concept of *Pareto optimality*. A point is called Pareto optimal, if there does not exist another point with the same or smaller objective function values, and with at least one objective function value being strictly smaller. Many algorithms for getting Pareto optimal solutions have been developed [7], but the research related to the existence or the boundedness of Pareto optimal set is still insufficient.

To clarify the properties of the Pareto solutions, we consider merit functions, that is, scalar-valued functions that return zero at the solutions of the original problems and strictly positive values otherwise. Merit functions for linear multiobjective optimization are proposed in [8]. Afterwards, linearized merit functions for convex multiobjective optimization [4] are also considered, and they are shown to be error bounds when the objective functions are strongly convex. In this paper, we propose nonlinear merit functions for nonlinear multiobjective optimization and prove that they are error bounds under strong convexity. Moreover, we propose a regularized and partially linearized merit function that can be used if each objective function is written as a sum of non-differentiable function and differentiable one.

Furthermore, we analyze convergence rates of multiobjective proximal gradient methods, proposed in [9]. The research about convergence rate analyses for multiobjective descent methods are relatively new. In [6], they analyze the rate of convergence of multiobjective gradient descent methods [5]. Their results seem to be appropriate since they are similar to the results of single-objective cases, but the metrics of the analyses are dependent on the gradient descent, so we cannot analyze other descent-type algorithms in the same manner. In this paper, by adopting the merit functions as the metrics, we enable any algorithms to be compared with others.

The outline of this paper is as follows. We define some basic notions and Pareto optimality in Sect. 2. In Sect. 3, we propose new merit functions for multiobjective optimization and present some of their properties. By using these properties, we show a condition for the existence and boundedness of the (weakly) Pareto optimal solutions in Sect. 4. In Sect. 5, we analyze the convergence rates of the multiobjective proximal gradient methods, and we conclude this paper in Sect. 6.

2 Preliminaries

We first present some notions that will be used in this paper. Let us denote by $\mathbf{R}$ the set of real numbers and by $\mathbf{N}$ that of positive integers. We use the symbol $\|\cdot\|$ for the Euclidean norm in $\mathbf{R}^n$. The notation $u \leq v$ ($u < v$) means that $u_i \leq v_i$ ($u_i < v_i$) for all $i \in \{1, \ldots, m\}$. Moreover, we call

$$h'(x; d) := \lim_{t \searrow 0} \frac{h(x + td) - h(x)}{t}$$

the directional derivative of $h\colon S \to \mathbf{R} \cup \{\infty\}$ at x in the direction d. Note that $h'(x; d) = \nabla h(x)^\top d$ when h is differentiable at x, where $\nabla h(x)$ stands for the gradient of h at x. The following well-known lemma shows a non-decreasing property when h is convex.

Lemma 2.1. *Assume that $h\colon S \to \mathbf{R} \cup \{\infty\}$ is a convex function and let $x \in S$ and $x + d \in S$. Then, the function $\tilde{h}\colon (0, 1] \to \mathbf{R}$ defined by*

$$\tilde{h}(\alpha) := \frac{h(x + \alpha d) - h(x)}{\alpha}$$

is non-decreasing. In particular, it follows that

$$h(x + d) - h(x) \geq \frac{h(x + \alpha d) - h(x)}{\alpha} \quad \text{for all } \alpha \in (0, 1].$$

Proof. It follows immediately from [3, Section 4.3]. □

Now, we introduce the concept of optimality for the multiobjective optimization problem (1). Recall that $x^* \in S$ is *Pareto optimal*, if there is no $x \in S$ such that $F(x) \leq F(x^*)$ and $F(x) \neq F(x^*)$. Likewise, $x^* \in S$ is *weakly Pareto optimal*, if there does not exist $x \in S$ such that $F(x) < F(x^*)$. It is known that Pareto optimal points are always weakly Pareto optimal, and the converse is not always true. We also say that $\bar{x} \in S$ is *Pareto stationary*, if and only if,

$$\max_{i \in \{1, \ldots, m\}} F_i'(\bar{x}; z - \bar{x}) \geq 0 \quad \text{for all } z \in S.$$

We state below the relation among the three concepts of Pareto optimality.

Lemma 2.2. *The following three statements hold.*

1. *If $x \in S$ is weakly Pareto optimal for (1), then x is Pareto stationary.*
2. *Let every component F_i of F be convex. If $x \in S$ is Pareto stationary for (1), then x is weakly Pareto optimal.*
3. *Let every component F_i of F be strictly convex. If $x \in S$ is Pareto stationary for (1), then x is Pareto optimal.*

Proof. It is clear from [9, Lemma 2.2]. □

3 Merit functions for multiobjective optimization

In this section, we propose some merit functions for nonlinear multiobjective optimization. A function is called merit function associated to an optimization problem if it returns zero at their solutions and strictly positive values otherwise.

3.1 A simple merit function

First, we propose a nonlinear merit function $w\colon S \to \mathbf{R} \cup \{\infty\}$ as follows:

$$w(x) \coloneqq \sup_{y \in S} \min_{i \in \{1,\ldots,m\}} \{F_i(x) - F_i(y)\}, \tag{2}$$

which is an extension of the one proposed in [8] for linear multiobjective optimization. The next theorem shows the basic property of this merit function.

Theorem 3.1. *Let w be defined by (2). Then, for all $x \in S$, we have $w(x) \geq 0$. Moreover, $x \in S$ is weakly Pareto optimal for (1) if and only if $w(x) = 0$.*

Proof. By the definition (2) of w, we get

$$w(x) \geq \min_{i \in \{1,\ldots,m\}} \{F_i(x) - F_i(x)\} = 0 \quad \text{for all } x \in S. \tag{3}$$

Now, assume that $w(x) = 0$. It follows from (2) that

$$\sup_{y \in S} \min_{i \in \{1,\ldots,m\}} \{F_i(x) - F_i(y)\} = 0 \iff \min_{i \in \{1,\ldots,m\}} \{F_i(x) - F_i(y)\} \leq 0 \text{ for all } y \in S,$$

which is equivalent to the existence of $i \in \{1, \ldots, m\}$ such that

$$F_i(x) - F_i(y) \leq 0 \text{ for all } y \in S.$$

In other words, there does not exist $y \in S$ such that

$$F_i(x) - F_i(y) > 0 \text{ for all } i \in \{1, \ldots, m\}.$$

Therefore, $x \in S$ is weakly Pareto optimal if and only if $w(x) = 0$. □

Now, we consider the following single-objective optimization problem:

$$\begin{aligned} \min \quad & w(x) \\ \text{s.t.} \quad & x \in S. \end{aligned} \tag{4}$$

If the global optimal solution x^* of (4) exists and satisfies $w(x^*) > 0$, then x^* is not weakly Pareto optimal from Theorem 3.1. However, as shown in the next theorem, the global solutions of (4) are always weakly Pareto optimal for (1). Before showing this, we need the following basic result.

Lemma 3.1. *Let $G_i\colon S \to \mathbf{R}$ and $H_i\colon S \to \mathbf{R}$ be upper and lower semicontinuous, respectively, for all $i \in \{1, \ldots, m\}$. Then, we have*

$$\sup_{x\in S} \min_{i\in\{1,\ldots,m\}} G_i(x) - \sup_{x\in S} \min_{i\in\{1,\ldots,m\}} H_i(x) \leq \sup_{x\in S} \max_{i\in\{1,\ldots,m\}} [G_i(x) - H_i(x)].$$

Proof. Let $f\colon S \to \mathbf{R}$ and $g\colon S \to \mathbf{R}$ be upper semicontinuous. Then, it follows that

$$\sup_{x\in S}(f(x) + g(x)) \leq \sup_{y\in S} f(x) + \sup_{x\in S} g(x).$$

Now, define $h\colon S \to \mathbf{R}$ as $h \coloneqq f + g$. Then, we obtain

$$\sup_{x\in S} h(x) - \sup_{x\in S} f(x) \leq \sup_{x\in S}(h(x) - f(x)).$$

Substituting $h(x) = \min_{i\in\{1,\ldots,m\}} G_i(x)$ and $f(x) = \min_{i\ in\{1,\ldots,m\}} H_i(x)$ into the above inequality, we get

$$\sup_{x\in S} \min_{i\in\{1,\ldots,m\}} G_i(x) - \sup_{x\in S} \min_{i\in\{1,\ldots,m\}} H_i(x) \leq \sup_{x\in S} \left[\min_{i\in\{1,\ldots,m\}} G_i(x) - \min_{i\in\{1,\ldots,m\}} H_i(x) \right].$$

Now, let $j_x \in \operatorname*{argmin}_{i\in\{1,\ldots,m\}} H_i(x)$. Then, we obtain

$$\min_{i\in\{1,\ldots,m\}} H_i(x) = H_{j_x}(x).$$

This yields

$$\begin{aligned}
\sup_{x\in S} \min_{i\in\{1,\ldots,m\}} G_i(x) - \sup_{x\in S} \min_{i\in\{1,\ldots,m\}} H_i(x) &\leq \sup_{x\in S} \left[\min_{i\in\{1,\ldots,m\}} G_i(x) - H_{j_x}(x) \right] \\
&\leq \sup_{x\in S} [G_{j_x}(x) - H_{j_x}(x)] \\
&\leq \sup_{x\in S} \max_{i\in\{1,\ldots,m\}} [G_i(x) - H_i(x)],
\end{aligned}$$

where the second and third inequalties come from the definition of the minimum and the maximum. □

Theorem 3.2. *Let w be defined by* (2)*. If $x^* \in S$ is global optimal for* (4)*, then x^* is weakly Pareto optimal for* (1)*.*

Proof. Let $x^* \in S$ be a global optimal solution of (4). Then, for all $z \in S$, we have

$w(x^*) \leq w(z)$. This gives

$$\begin{aligned}
0 &\leq w(z) - w(x^*) \\
&= \sup_{y \in S} \min_{i \in \{1,\dots,m\}} \{F_i(z) - F_i(y)\} - \sup_{y \in S} \min_{i \in \{1,\dots,m\}} \{F_i(x^*) - F_i(y)\} \\
&\leq \sup_{y \in S} \max_{i \in \{1,\dots,m\}} \{(F_i(z) - F_i(y)) - (F_i(x^*) - F_i(y))\} \\
&= \max_{i \in \{1,\dots,m\}} \{F_i(z) - F_i(x^*)\},
\end{aligned}$$

where the first equality comes from the definition (2) of w, and Lemma 3.1 yields the second inequality. Therefore, x is weakly Pareto optimal for (1) by definition. □

Since the objective function of (4) is generally non-convex, the problem (4) does not necessarily have global optimal solutions. However, we can prove that every stationary point of (4) is Pareto stationary for (1).

Theorem 3.3. *Let w be defined by* (2) *and assume that there exists a directional derivative $F_i'(x; z-x)$ for all $i \in \{1,\dots,m\}$ and $x, z \in S$. If w also has an lower Dini derivative*[1] *$w'_-(x; z-x)$ for all $x, z \in S$ and is stationary for* (4)*, that is,*

$$w'_-(x; z-x) \geq 0 \quad \textit{for all } z \in S, \tag{5}$$

then x is Pareto stationary for (1)*.*

Proof. Let $x \in S$ be stationary for (4). By the definition (2) of w, we see that for all $z \in S$,

$$\begin{aligned}
&w'_-(x; z-x) \\
&= \liminf_{t \searrow 0} \frac{1}{t} \left[\sup_{y \in S} \min_{i \in \{1,\dots,m\}} \{F_i(x + t(z-x)) - F_i(y)\} - \sup_{y \in S} \min_{i \in \{1,\dots,m\}} \{F_i(x) - F_i(y)\} \right] \\
&\leq \liminf_{t \searrow 0} \frac{1}{t} \sup_{y \in S} \max_{i \in \{1,\dots,m\}} \{(F_i(x + t(z-x)) - F_i(y)\} - \{F_i(x) - F_i(y))\} \\
&= \liminf_{t \searrow 0} \max_{i \in \{1,\dots,m\}} \frac{F_i(x + t(z-x)) - F_i(x)}{t},
\end{aligned}$$

where the definition of the lower Dini derivative yields the first equality, and the inequality follows from Lemma 3.1. Now, define $h\colon S \to \mathbf{R} \cup \{\infty\}$ as

$$h(x) \coloneqq \max_{i \in \{1,\dots,m\}} \frac{F_i(x + t(z-x)) - F_i(x)}{t}.$$

[1]The lower Dini derivative of $h\colon S \to \mathbf{R} \cup \{\infty\}$ at x in the direction d is defined as

$$h'_-(x; d) \coloneqq \liminf_{t \searrow 0} \frac{h(x + td) - h(x)}{t}.$$

Since h is continuous, we have

$$\begin{aligned} w'_-(x; z-x) &= \max_{i\in\{1,\dots,m\}} \liminf_{t\searrow 0} \frac{F_i(x+t(z-x)) - F_i(x)}{t} \\ &= \max_{i\in\{1,\dots,m\}} F_i'(x; z-x), \end{aligned}$$

where the second equality comes from the definition of Dini derivative. Therefore, by (5) we get

$$\max_{i\in\{1,\dots,m\}} F_i'(x; z-x) \geq 0 \quad \text{for all } z \in S,$$

which shows that x is Pareto stationary for (1). □

Before introducing some properties of the merit function w, we define the level-boundedness of scalar-valued and vector-valued functions.

Definition 3.1. *A function $f\colon S \to \mathbf{R}$ is called* level-bounded *if the level set $\Omega_f(\alpha) \coloneqq \{x \in S \mid f(x) \leq \alpha\}$ is bounded for all $\alpha \in \mathbf{R}$. Similarly, a vector-valued function $F\colon S \to \mathbf{R}^m$ is level-bounded if the level set $\Omega_F(\zeta) \coloneqq \{x \in S \mid F(x) \leq \zeta\}$ is bounded for all $\zeta \in \mathbf{R}^m$.*

Note that if $F_i\colon S \to \mathbf{R}$ is level-bounded for all $i \in \{1, \dots, m\}$, then a vector-valued function $F \coloneqq (F_1, \dots, F_m)^\top$ is also level-bounded, but the reverse is not necessarily true. Note also that for $m = 1$, this definition coincides with the level-boundedness for scalar-valued functions. Now, we state below a sufficient condition for the level-boundedness of the merit function w.

Theorem 3.4. *Let w be defined by (2). If F_i is level-bounded for all $i \in \{1, \dots, m\}$, them w is also level-bounded.*

Proof. Suppose, contrary to our claim, that w is not level-bounded. Then, there exists $\alpha \in \mathbf{R}$ such that $\{x \in S \mid w(x) \leq \alpha\}$ is not bounded. By the definition (2) of w, the inequality $w(x) \leq \alpha$ can be written as

$$\sup_{y\in S} \min_{i\in\{1,\dots,m\}} \{F_i(x) - F_i(y)\} \leq \alpha.$$

This implies for some fixed $z \in S$ that there exists $j \in \{1, \dots, m\}$ such that

$$F_j(x) \leq F_j(z) + \alpha.$$

Therefore, it follows that

$$\{x \in S \mid w(x) \leq \alpha\} \subseteq \bigcup_{j=1}^{m} \{x \in S \mid F_j(x) \leq F_j(z) + \alpha\}.$$

Since F_i is level-bounded for all $i \in \{1, \dots, m\}$, the right-hand side must be bounded, which contradicts the unboundedness of the left-hand side. □

As indicated by the following example, even if F is level-bounded, w is not necessarily level-bounded.

Example 3.1. *Consider the bi-objective function* $F\colon \mathbf{R} \to \mathbf{R}^2$ *with each component given by*

$$F_1(x) \coloneqq x^2, \quad F_2(x) \coloneqq 0.$$

Then, the merit function w *defined by* (2) *is written as*

$$\begin{aligned} w(x) &= \sup_{y\in\mathbf{R}} \min\{F_1(x) - F_1(y), F_2(x) - F_2(y)\} \\ &= \sup_{y\in\mathbf{R}} \min\{(x^2 - y^2), 0\} = 0. \end{aligned}$$

On the other hand, F *is level-bounded because* $\lim_{\|z\|\to\infty} F_1(x) = +\infty$.

Next, we show an error bound property of the merit function (2) when each F_i is strongly convex.

Theorem 3.5. *Let* w *be defined by* (2). *Assume that* F_i *is strongly convex with modulus* $\mu_i > 0$ *for all* $i \in \{1, \dots, m\}$. *Then, we have*

$$w(x) \geq \frac{\mu}{2} \operatorname{dist}\{x, X^*\}^2 \quad \textit{for all } x \in S,$$

where $\mu \coloneqq \min_{i\in\{1,\dots,m\}} \mu_i$, $\operatorname{dist}\{x, X^*\} \coloneqq \min\{\|x - y\| \mid y \in X^*\}$ *and* $X^* \coloneqq \{x \in S \mid x \text{ is (weakly) Pareto stationary for } (1)\}$.

Proof. By the definition (2), we have

$$\begin{aligned} w(x) &= \sup_{y\in S} \min_{i\in\{1,\dots,m\}} \{F_i(x) - F_i(y)\} \\ &= -\inf_{y\in S} \max_{i\in\{1,\dots,m\}} \{F_i(y) - F_i(x)\}. \end{aligned} \tag{6}$$

Since F_i is strongly convex, there exists a unique $y_x^* \in S$ that attains infimum in (6). Now, let $\mathcal{I}_x(y) \coloneqq \{j \in \{1, \dots, m\} \mid \max_{i\in\{1,\dots,m\}}\{F_i(y) - F_i(x)\} = F_j(y) - F_j(x)\}$ for all $y \in S$. The optimality condition of (6) and the strong convexity of F_i yield

$$0 \in \partial_{y_x^*} \max_{i\in\{1,\dots,m\}} \{F_i(y_x^*) - F_i(x)\} = \operatorname{conv}\{\partial F_i(y_x^*) \mid i \in \mathcal{I}_x(y_x^*)\},$$

where the symbol ∂ stands for the subdifferential and the operator conv denotes the convex hull. Thus, there exist $\eta_i \in \partial F_i(y_x^*)$ and $\lambda_i \geq 0$ with $i \in \mathcal{I}(y_x^*)$ such that

$$\sum_{i\in\mathcal{I}_x(y_x^*)} \lambda_i \eta_i = 0, \qquad \sum_{i\in\mathcal{I}_x(y_x^*)} \lambda_i = 1. \tag{7}$$

Therefore, it follows that

$$w(x) = F_i(x) - F_i(y_x^*) \text{ for all } i \in \mathcal{I}_x(y_x^*).$$

Since F_i is strongly convex with modulus μ_i, we have

$$w(x) = F_i(x) - F_i(y_x^*) \geq \eta_i^\top (x - y_x^*) + \frac{\mu_i}{2}\|x - y_x^*\|^2 \text{ for all } i \in \mathcal{I}(y_x^*).$$

Multiplying the above inequality by λ_i and summing them up for $i \in \mathcal{I}_x(y_x^*)$, we get

$$\begin{aligned} w(x) = \sum_{i \in \mathcal{I}_x(y_x^*)} \lambda_i w(x) &\geq \sum_{i \in \mathcal{I}_x(y_x^*)} \lambda_i \left\{ \eta_i^\top (x - y_x^*) + \frac{\mu_i}{2}\|x - y_x^*\|^2 \right\} \\ &= \sum_{i \in \mathcal{I}_x(y_x^*)} \frac{\lambda_i \mu_i}{2}\|x - y_x^*\|^2 \qquad (8) \\ &\geq \frac{\mu}{2}\|x - y_x^*\|^2, \end{aligned}$$

where (7) shows the equalities, and the second inequality follows from the definition of μ and λ_i. On the other hand, the definition of y_x^* yields

$$\max_{i \in \{1,\ldots,m\}} \{F_i(y_x^*) - F_i(x)\} \leq \max_{i \in \{1,\ldots,m\}} \{F_i(y) - F_i(x)\} \quad \text{for all } y \in S.$$

This means that

$$F_j(y_x^*) - F_j(x) \leq \max_{i \in \{1,\ldots,m\}} \{F_i(y) - F_i(x)\} \text{ for all } y \in S \text{ and all } j \in \{1, \ldots, m\}.$$

Thus, there exists $i \in \mathcal{I}_x(y)$ such that

$$F_i(y_x^*) \leq F_i(y) \quad \text{for all } y \in S,$$

which shows that y_x^* is weakly Pareto stationary for (1) for any x, that is, $y_x^* \in X^*$. Therefore, we obtain $\|x - y_x^*\|^2 \geq \operatorname{dist}\{x, X^*\}^2$ for all $x \in S$, which, combined with (8), gives the assertion of the theorem. □

3.2 A regularized and partially linearized merit function

Let us now consider that each component F_i of the objective function F of (1) is defined by

$$F_i(x) \coloneqq f_i(x) + g_i(x), \quad i \in \{1, \ldots, m\}, \qquad (9)$$

where $f_i \colon S \to \mathbf{R}$ is continuously differentiable and $g_i \colon S \to \mathbf{R} \cup \{\infty\}$ is closed, proper and convex. Now, we propose a partially linearized and regularized merit function $u_\ell \colon S \to \mathbf{R}$ as follows:

$$u_\ell(x) \coloneqq \max_{y \in S} \min_{i \in \{1,\ldots,m\}} \left\{ \nabla f_i(x)^\top (x - y) + g_i(x) - g_i(y) - \frac{\ell}{2}\|x - y\|^2 \right\}, \qquad (10)$$

where $\ell \geq 0$ is a given constant. When $g_i = 0$, $u_\ell(x)$ given in (10) corresponds to the regularized and linearized merit function proposed in [4]. Note that there exists

a unique $y_\ell^* \in S$ that attains the maximum because the objective function in (10) is strongly concave with respect to y. Before stating the property of u_ℓ, we recall the so-called descent lemma.

Lemma 3.2. *[2, Proposition A.24] Let $h\colon \mathbf{R}^n \to \mathbf{R}$ be a continuously differentiable function. If ∇h is Lipschitz continuous with Lipschitz constant L, then we have*

$$h(y) - h(x) \leq \nabla h(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2.$$

The next proposition makes a connection among the merit function u_ℓ, with $\ell \geq 0$, and w defined in (2).

Proposition 3.1. *Let w and u_ℓ be defined by (2) and (10), respectively. Then, the following statements follows.*

1. *For all $x \in S$, $u_\ell(x) \geq 0$. Moreover, $x \in S$ is Pareto stationary if and only if $u_\ell(x) = 0$.*
2. *Suppose that for all $i \in \{1, \ldots, m\}$, ∇f_i is Lipschitz continuous with Lipschitz constant $L_i > 0$ and let $L_{\max} \coloneqq \max_{i\in\{1,\ldots,m\}} L_i$. Then, we have*

$$w(x) \geq u_{L_{\max}}(x) \quad \textit{for all } x \in S.$$

3. *Assume that f_i has a convexity parameter $\mu_i \geq 0$ for all $i \in \{1, \ldots, m\}$, and let $\mu \coloneqq \min_{i\in\{1,\ldots,m\}} \mu_i$. Then, we get*

$$u_\mu(x) \geq w(x) \quad \textit{for all } x \in S.$$

4. *For all $x \in S$ and $\ell > 0$, it follows that*

$$\min\{1, 1/\ell\}\, u_1(x) \leq u_\ell(x) \leq \max\{1, 1/\ell\}\, u_1(x).$$

Proof. 1. By the definition (10) of u_ℓ, we have

$$u_\ell(x) \geq \min_{i\in\{1,\ldots,m\}} \left\{ \nabla f_i(x)^\top (x - x) + g_i(x) - g_i(y) - \frac{\ell}{2}\|x - x\|^2 \right\} = 0.$$

The latter statement follows immediately from [9, Lemma 3.2].

2. Since ∇f_i is Lipschitz continuous with Lipschitz constant L_i, Lemma 3.2 yields

$$f_i(y) - f_i(x) \leq \nabla f(x)^\top (y - x) + \frac{L_i}{2}\|y - x\|^2.$$

By the definition of $L_{\max}$, we have

$$f_i(x) - f_i(y) \geq \nabla f(x)^\top (x-y) - \frac{L_{\max}}{2}\|x-y\|^2.$$

Therefore, we immediately get $w(x) \geq u_{L_{\max}}(x)$ for all $x \in S$ by the definitions of w and $u_{L_{\max}}$.

3. The convexity of f_i with parameter μ_i gives

$$\nabla f_i(x)^\top (x-y) - \frac{\mu_i}{2}\|x-y\|^2 \geq f_i(x) - f_i(y),$$

so it is clear that $u_\mu(x) \geq w(x)$ for all $x \in S$.

4. As mentioned above, for every $\ell > 0$ there exists a unique y_ℓ^* that attains the maximum in the definition (10) of u_ℓ. Therefore, when $\ell < 1$ we have

$$\begin{aligned}
&u_\ell(x) \\
&= \min_{i\in\{1,\dots,m\}} \left\{ \nabla f_i(x)^\top (x - y_\ell^*) + g_i(x) - g_i(y_\ell^*) - \frac{\ell}{2}\|x - y_\ell^*\|^2 \right\} \\
&= \frac{1}{\ell} \min_{i\in\{1,\dots,m\}} \left\{ \nabla f_i(x)^\top \left(\ell(x - y_\ell^*)\right) + \ell(g_i(x) - g_i(y_\ell^*)) - \frac{1}{2}\left\|\ell(x - y_\ell^*)\right\|^2 \right\} \\
&\leq \frac{1}{\ell} \min_{i\in\{1,\dots,m\}} \left\{ \nabla f_i(x)^\top \left(\ell(x - y_\ell^*)\right) + g_i(x) - g_i\left(x - \ell(x - y_\ell^*)\right) - \frac{1}{2}\left\|\ell(x - y_\ell^*)\right\|^2 \right\} \\
&\leq \frac{1}{\ell} u_1(x),
\end{aligned}$$

where the first inequality follows from the convexity of g_i and the second inequality comes from the definition (10) of u_1. On the other hand, when $\ell \geq 1$ we get

$$\begin{aligned}
u_1(x) &= \min_{i\in\{1,\dots,m\}} \left\{ \nabla f_i(x)^\top (x - y_1^*) + g_i(x) - g_i(y_1^*) - \frac{1}{2}\|x - y_1^*\|^2 \right\} \\
&\geq \min_{i\in\{1,\dots,m\}} \left\{ \nabla f_i(x)^\top (x - y_\ell^*) + g_i(x) - g_i(y_\ell^*) - \frac{1}{2}\|x - y_\ell^*\|^2 \right\} \\
&\geq \min_{i\in\{1,\dots,m\}} \left\{ \nabla f_i(x)^\top (x - y_\ell^*) + g_i(x) - g_i(y_\ell^*) - \frac{\ell}{2}\|x - y_\ell^*\|^2 \right\} \\
&= u_\ell(x).
\end{aligned}$$

The above two inequalities imply $u_\ell(x) \leq \max\{1, 1/\ell\}\, u_1(x)$. Moreover, we can prove $u_\ell(x) \leq \max\{1, 1/\ell\}\, u_1(x)$ in the same manner. □

4 Existence and boundedness of the Pareto solutions

In this section, we provide sufficient conditions for the existence of weakly Pareto optimal solutions, and for the boundedness of Pareto optimal sets. First, we show a theorem about the existence of weakly Pareto optimal solutions.

Theorem 4.1. *If F is continuous and level-bounded, then* (1) *has a weakly Pareto optimal solution.*

Proof. Let F be continuous and level-bounded. Then the level set $\Omega_F(\alpha) \coloneqq \{x \in S \mid F_i(x) \leq \alpha \text{ for all } i \in \{1, \ldots, m\}\}$ is bounded for all $\alpha \in \mathbf{R}$. Now, we have

$$\Omega_F(\alpha) = \{x \in S \mid \max_{i \in \{1,\ldots,m\}} F_i(x) \leq \alpha\} = \Omega_{\max_i F_i}(\alpha),$$

so $\max_i F_i$ is also level-bounded. Moreover, since F is continuous, $\max_i F_i$ is also continuous. Thus, the problem

$$\begin{aligned} \min \quad & \max_{i \in \{1,\ldots,m\}} F_i(x) \\ \text{s.t.} \quad & x \in S \end{aligned}$$

has a global optimal solution x^*. This gives

$$\max_{i \in \{1,\ldots,m\}} F_i(x^*) \leq \max_{i \in \{1,\ldots,m\}} F_i(x) \quad \text{for all } x \in S,$$

which means that x^* is weakly Pareto optimal for (1). □

The boundedness condition of Pareto optimal sets follows immediately from Theorem 3.5, and by assuming strong convexity of the objectives.

Theorem 4.2. *If F_i is strongly convex for all $i \in \{1, \ldots, m\}$, then the (weakly) Pareto optimal set of* (1) *is bounded.*

5 Convergence rate of multiobjective proximal gradient methods

In this section, we analyze the convergence rate of multiobjective proximal gradient methods proposed in [9]. They are applicable to (1), if the problem is unconstrained, that is, $S = \mathbf{R}^n$, and each component of the objective function is defined by (9). In addition, we assume that ∇f_i is Lipschitz continuous with Lipschitz constant L_i and let $L_{\max} \coloneqq \max_{i \in \{1,\ldots,m\}} L_i$. We consider two types of proximal gradient methods, with and without line searches. Both algorithms generate some sequence $\{x^k\}$ iteratively with the following procedure:

$$x^{k+1} \coloneqq x^k + t_k d^k,$$

where $t_k > 0$ is a step size and d^k is a search direction. At every iteration k, we define the search direction d^k by solving

$$d^k = \underset{d \in \mathbf{R}^n}{\operatorname{argmin}} \left[\psi_{x^k}(d) + \frac{\ell}{2}\|d\|^2 \right], \tag{11}$$

where $\ell > 0$ is a given constant and the function $\psi_x \colon \mathbf{R}^n \to \mathbf{R}$ is defined by

$$\psi_x(d) \coloneqq \max_{i \in \{1,\ldots,m\}} \left\{ \nabla f_i(x)^\top d + g_i(x+d) - g_i(x) \right\}. \tag{12}$$

Note that we have

$$\psi_{x^k}(d^k) + \frac{\ell}{2}\|d^k\|^2 = -u_\ell(x^k), \tag{13}$$

where $u_\ell(x^k)$ is defined by (10). From now on, we suppose that an infinite sequence of iterates is generated. The next result shows an important property of ψ_x.

Lemma 5.1. *[9, Lemma 4.1] Let $\{d^k\}$ be generated by a multiobjective proximal gradient methods and recall the definition* (12) *of ψ_x. Then, we have*

$$\psi_{x^k}(d^k) \leq -\ell\|d^k\|^2 \quad \textit{for all } k.$$

5.1 Proximal gradient methods with line searches

In this section, we analyze the convergence rates of the algorithms with line searches.

(a) The non-convex case

First, we suppose that f_i is non-convex for all $i \in \{1, \ldots, m\}$. To keep the paper self-contained, we first recall the algorithm with line searches.

Algorithm 5.1 (Proximal gradient method with line searches)**.**

Step 1: Choose $\ell > 0$, $\rho \in (0,1)$, $\xi \in (0,1)$, $x^0 \in \mathbf{R}^n$ and set $k \coloneqq 0$.

Step 2: Compute d^k by solving subproblem $d^k = \underset{d \in \mathbf{R}^n}{\operatorname{argmin}} \left[\psi_{x^k}(d) + \frac{\ell}{2}\|d\|^2 \right]$.

Step 3: If $d^k = 0$, then stop.

Step 4: Compute the step length $t_k \in (0,1]$ as the maximum of

$$T_k \coloneqq \{t = \xi^j \mid j \in \mathbf{N},\ F_i(x^k + t_k d^k) \leq F_i(x^k) + t_k \rho \psi_{x^k}(d^k),\ i = 1, \ldots, m\}$$

Step 5: Set $x^{k+1} \coloneqq x^k + t_k d^k$, $k \coloneqq k+1$, and go to Step 2.

To begin with, we show the existence of a uniform lower bound on the step size t_k.

Lemma 5.2. *In Algorithm 5.1 the step size t_k satisfies the following inequality for every iteration k:*

$$t_k \geq t_{\min} \coloneqq \min\left\{\frac{2\xi(1-\rho)\ell}{L_{\max}}, 1\right\}.$$

Proof. If $t_k = 1$, then the claim is clear. Thus, we suppose that $t_k < 1$. By the definition of t_k in Step 4 of Algorithm 5.1, there exists $i \in \{1, \ldots, m\}$ such that

$$\begin{cases} F_i(x^k + \xi^{-1}t_k d^k) - F_i(x^k) > \xi^{-1}t_k \rho \psi_{x^k}(d^k) \\ 0 < \xi^{-1}t_k \leq 1. \end{cases} \tag{14}$$

On the other hand, it follows by the definition (12) of ψ_x that

$$\begin{aligned} \psi_{x^k}(d^k) &\geq \nabla f_i(x^k)^\top d^k + g_i(x^k + d^k) - g_i(x^k) \\ &\geq \frac{\xi^{-1}t_k \nabla f_i(x^k)^\top d^k + g_i(x^k + \xi^{-1}t_k d^k) - g_i(x^k)}{\xi^{-1}t_k} \\ &\geq \frac{F_i(x^k + \xi^{-1}t_k d^k) - F_i(x^k) - L_i\|\xi^{-1}t_k d^k\|^2/2}{\xi^{-1}t_k} \\ &= \frac{F_i(x^k + \xi^{-1}t_k d^k) - F_i(x^k)}{\xi^{-1}t_k} - \frac{L_i}{2}\xi^{-1}t_k\|d^k\|^2, \end{aligned} \tag{15}$$

where the second inequality comes from the convexity of g_i and Lemma 2.1, and the third one follows from the Lipschitz continuity of ∇f_i and Lemma 3.2. From (14) and (15), we have

$$\frac{F_i(x^k + \xi^{-1}t_k d^k) - F_i(x^k)}{\xi^{-1}t_k} - \frac{L_i}{2}\xi^{-1}t_k\|d^k\|^2 < \frac{1}{\rho}\frac{F_i(x^k + \xi^{-1}t_k d^k) - F_i(x^k)}{\xi^{-1}t_k}.$$

Thus, we get

$$-\frac{L_i}{2}\xi^{-1}t_k\|d^k\|^2 < \frac{1-\rho}{\rho}\frac{F_i(x^k + \xi^{-1}t_k d^k) - F_i(x^k)}{\xi^{-1}t_k}.$$

Applying (15) again gives

$$-\frac{L_i}{2}\xi^{-1}t_k\|d^k\|^2 < \frac{1-\rho}{\rho}\left(\psi_{x^k}(d^k) + \frac{L_i}{2}\xi^{-1}t_k\|d^k\|^2\right).$$

It follows from Lemma 5.1 that

$$-\frac{L_i}{2}\xi^{-1}t_k\|d^k\|^2 < \frac{1-\rho}{\rho}\left(-\ell\|d^k\|^2 + \frac{L_i}{2}\xi^{-1}t_k\|d^k\|^2\right),$$

which is equivalent to

$$t_k > \frac{2(1-\rho)\ell}{L_i\xi^{-1}}.$$

Therefore, using the definition of $L_{\max}$, we conclude that

$$t_k \geq t_{\min} := \min\left\{\frac{2\xi(1-\rho)\ell}{L_{\max}}, 1\right\}.$$

□

The next theorem shows that Algorithm 5.1 has a convergence rate of order $1/k$.

Theorem 5.1. *Suppose that there exists some nonempty set $\mathcal{J} \subseteq \{1,\ldots,m\}$ such that if $i \in \mathcal{J}$ then F_i is has a lower bound $F_i^{\min}$. Let $F^{\min} := \max_{i\in\mathcal{J}} F_i^{\min}$ and $F_0^{\max} := \max_{i\in\{1,\ldots,m\}} F_i(x^0)$. Then, the Algorithm 5.1 generates a sequence $\{x^k\}$ such that*

$$\min_{0\leq j\leq k-1} u_1(x^j) \leq \frac{F_0^{\max} - F^{\min}}{t_{\min}\rho k \min\{1, 1/\ell\}}.$$

Proof. Let $i \in \mathcal{J}$. By the definition of t_k in Step 4 of Algorithm 5.1 it follows that

$$\begin{aligned} F_i(x^k + t_k d^k) - F_i(x^k) &\leq t_k\rho\psi_{x^k}(d^k) \\ &= -t_k\rho u_\ell(x^k) - \frac{\ell t_k \rho}{2}\|d^k\|^2 \\ &\leq -t_k\rho u_\ell(x^k), \end{aligned}$$

where the equality follows from (13). Therefore, we have

$$\begin{aligned} F_i(x^k) - F_i(x^k + t_k d^k) &\geq t_k\rho u_\ell(x^k) \\ &\geq t_{\min}\rho u_\ell(x^k), \end{aligned}$$

where the second inequality comes from Lemma 5.2. Adding up the above inequality from $k = 0$ to $k = \tilde{k} - 1$, we obtain

$$\begin{aligned} F_i(x^0) - F_i(x^{\tilde{k}-1} + t_{\tilde{k}-1}d^{\tilde{k}-1}) &\geq t_{\min}\rho\sum_{j=0}^{\tilde{k}-1} u_0(x^j) \\ &\geq t_{\min}\rho\tilde{k}\min_{0\leq j\leq\tilde{k}-1} u_\ell(x^j). \end{aligned}$$

From the statement 4 of Proposition 3.1, we conclude that

$$\min_{0\leq j\leq\tilde{k}-1} u_1(x^j) \leq \frac{F_0^{\max} - F^{\min}}{t_{\min}\rho\tilde{k}\min\{1, 1/\ell\}}.$$

□

(b) The convex and strongly convex cases

When f_i is convex or strongly convex for all $i \in \{1, \ldots, m\}$, we can modify the Armijo condition and use alternatively a sufficient decrease condition (see (16)). Then, the algorithm is described as follows:

Algorithm 5.2 (Proximal gradient method with line searches (convex case))**.**

Step 1: Choose $\ell > 0,\ \gamma \in (0,1),\ \xi \in (0,1),\ x^0 \in \mathbf{R}^n$ *and set* $k \coloneqq 0$.

Step 2: Compute d^k *by solving subproblem* $d^k = \operatorname*{argmin}_{d \in \mathbf{R}^n} \left[\psi_{x^k}(d) + \frac{\ell}{2}\|d\|^2\right]$.

Step 3: If $d^k = 0$, *then stop.*

Step 4: Compute the step length $t_k \in (0,1]$ *as the maximum of*

$$T_k \coloneqq \{t = \xi^j \mid j \in \mathbf{N},\ F_i(x^k + td^k) \le F_i(x^k) + t\psi_{x^k}(d^k) + \frac{\gamma t \ell}{2}\|d^k\|^2, \quad i = 1, \ldots, m\} \tag{16}$$

Step 5: Set $x^{k+1} \coloneqq x^k + t_k d^k$, $k \coloneqq k+1$, *and go to Step 2.*

We can easily see that the step size has a lower bound in each iteration.

Lemma 5.3. *In Algorithm 5.2, the step size* t_k *satisfies the following inequality for every iteration* k*:*

$$t_k \ge t_{\min} \coloneqq \min\left\{\frac{\xi\gamma\ell}{L_{\max}}, 1\right\}.$$

Proof. Since ∇f_i is Lipschitz continuous with the Lipschitz constant L_i, we have for all $t \in (0, (\gamma\ell)/L_i]$,

$$\begin{aligned} F_i(x^k + td^k) - F_i(x^k) &\le t\nabla f_i(x^k)^\top d^k + g_i(x^k + td^k) - g_i(x^k) + \frac{L_i}{2}\|td^k\|^2 \\ &\le t\left(\nabla f_i(x^k)^\top d^k + g_i(x^k + d^k) - g_i(x^k)\right) + \frac{L_i}{2}\|td^k\|^2 \\ &\le t\psi_{x^k}(d^k) + \frac{L_i}{2}\|td^k\|^2 \\ &\le t\psi_{x^k}(d^k) + \frac{\gamma\ell}{2t}\|td^k\|^2, \end{aligned}$$

where the second inequality follows from the convexity of g_i and the third one comes from the definition (12) of ψ_x. Therefore, the condition (16) are satisfied for all $t \in (0, (\gamma\ell)/L_{\max}]$. By the definition of t_k in Step 3 of Algorithm 5.2, we get

$$t_k \ge \min\left\{\frac{\xi\gamma\ell}{L_{\max}}, 1\right\}.$$

□

Next, we prove the following lemma, which is the key to analyze the convergence rate of the method in the convex and strongly convex cases.

Lemma 5.4. *Let f_i and g_i have convexity parameters $\mu_i \in \mathbf{R}$ and $\nu_i \in \mathbf{R}$, respectively, and write $\mu \coloneqq \min_{i\in\{1,\dots,m\}} \mu_i$ and $\nu \coloneqq \min_{i\in\{1,\dots,m\}} \nu_i$. Then, Algorithm 5.2 generates a sequence $\{x^k\}$ such that for all $x \in \mathbf{R}^n$,*

$$\begin{aligned}
&\sum_{i=1}^m \lambda_i^k \left(F_i(x^{k+1}) - F_i(x)\right) \\
&\leq \frac{\ell + \nu(1-t_{\min})}{2t_{\min}} \left(\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2 \\
&\quad - (1 - t_{\min}) \left(\psi_{x^k}(d^k) + \left(\ell + \frac{\nu}{2}\right)\|d^k\|^2\right),
\end{aligned}$$

where λ_i^k satisfies the following conditions:

$$\begin{cases}
\text{There exists } \eta_i \in \partial g_i(x^k + d^k) \text{ such that } \displaystyle\sum_{i=1}^m \lambda_i^k(\nabla f_i(x^k) + \eta_i) + \ell d^k = 0, & (17) \\
\displaystyle\sum_{i=1}^m \lambda_i^k = 1, \quad \lambda_i^k \geq 0 \quad \left(i \in \mathcal{I}_{x^k}(d^k)\right), \quad \lambda_i^k = 0 \quad \left(i \notin \mathcal{I}_{x^k}(d^k)\right), & (18)
\end{cases}$$

where

$$\mathcal{I}_x(d) \coloneqq \{i \in \{1,\dots,m\} \mid \psi_x(d) = \nabla f_i(x)^\top d + g_i(x+d) - g_i(x)\}. \tag{19}$$

Proof. By the definition of t_k in Step 3 of Algorithm 5.2, it follows that for all $i \in \{1,\dots,m\}$,

$$F_i(x^{k+1}) \leq F_i(x^k) + t_k\psi_{x^k}(d^k) + \frac{\gamma t_k \ell}{2}\|d^k\|^2.$$

Since $\gamma < 1$, we get

$$F_i(x^{k+1}) \leq F_i(x^k) + t_k\psi_{x^k}(d^k) + \frac{t_k \ell}{2}\|d^k\|^2.$$

Now, the convexity of f_i with modulus $\mu_i \in \mathbf{R}$ yields that for all $x \in \mathbf{R}^n$,

$$\begin{aligned}
F_i(x^{k+1}) &\leq F_i(x) + \nabla f_i(x^k)^\top(x^k - x) - \frac{\mu_i}{2}\|x^k - x\|^2 + g_i(x^k) - g_i(x) \\
&\quad + t_k\psi_{x^k}(d^k) + \frac{t_k\ell}{2}\|d^k\|^2 \\
&\leq F_i(x) + \nabla f_i(x^k)^\top(x^k - x) - \frac{\mu}{2}\|x^k - x\|^2 + g_i(x^k) - g_i(x) \\
&\quad + t_k\psi_{x^k}(d^k) + \frac{t_k\ell}{2}\|d^k\|^2,
\end{aligned}$$

where the second inequality follows from the definition of μ. Multiplying by λ_i^k and

summing for all $i \in \{1, \ldots, m\}$, we have

$$
\begin{aligned}
&\sum_{i=1}^{m} \lambda_i^k \Big(F_i(x^{k+1}) - F_i(x) \Big) \\
\leq &\sum_{i=1}^{m} \lambda_i^k \Big(\nabla f_i(x^k)^\top (x^k - x) - \frac{\mu}{2}\|x^k - x\|^2 + g_i(x^k) - g_i(x) \\
&+ t_k \psi_{x^k}(d^k) + \frac{t_k \ell}{2}\|d^k\|^2 \Big) \\
= &\sum_{i=1}^{m} \lambda_i^k \Bigg(\nabla f_i(x^k)^\top (x^k + d^k - x) - \frac{\mu}{2}\|x^k - x\|^2 + g_i(x^k + d^k) - g_i(x) \\
&- \Big(\nabla f_i(x^k)^\top d^k + g_i(x^k + d^k) - g_i(x^k) \Big) + t_k \psi_{x^k}(d^k) + \frac{t_k \ell}{2}\|d^k\|^2 \Bigg) \\
= &\sum_{i=1}^{m} \lambda_i^k \Big(\nabla f_i(x^k)^\top (x^k + d^k - x) + g_i(x^k + d^k) - g_i(x) \Big) \\
&- (1 - t_k)\psi_{x^k}(d^k) - \frac{\mu}{2}\|x^k - x\|^2 + \frac{t_k \ell}{2}\|d^k\|^2,
\end{aligned}
$$

where the second equality follows from (18) and (19). Now, let $\eta_i \in \partial g_i(x^k + d^k)$. Then, from the convexity of g_i with modulus $\nu_i \in \mathbf{R}$ we get

$$
\begin{aligned}
&\sum_{i=1}^{m} \lambda_i^k \Big(F_i(x^{k+1}) - F_i(x) \Big) \\
\leq &\sum_{i=1}^{m} \lambda_i^k \left[\Big(\nabla f_i(x^k) + \eta_i \Big)^\top (x^k + d^k - x) - \frac{\nu_i}{2}\|x^k + d^k - x\|^2 \right] \\
&- (1 - t_k)\psi_{x^k}(d^k) - \frac{\mu}{2}\|x^k - x\|^2 + \frac{t_k \ell}{2}\|d^k\|^2 \\
\leq &\sum_{i=1}^{m} \lambda_i^k \Big(\nabla f_i(x^k) + \eta_i \Big)^\top (x^k + d^k - x) \\
&- (1 - t_k)\psi_{x^k}(d^k) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^k + d^k - x\|^2 + \frac{t_k \ell}{2}\|d^k\|^2,
\end{aligned}
$$

where the definition of ν and (18) yield the second inequality. Now, the condition (17)

gives

$$
\begin{aligned}
&\sum_{i=1}^{m} \lambda_i^k \left(F_i(x^{k+1}) - F_i(x) \right) \\
\leq & -\ell (d^k)^\top \left(x^k + d^k - x \right) - (1-t_k)\psi_{x^k}(d^k) - \frac{\mu}{2}\|x^k - x\|^2 \\
& - \frac{\nu}{2}\|x^k + d^k - x\|^2 + \frac{t_k \ell}{2}\|d^k\|^2 \\
\leq & -\ell (d^k)^\top \left(x^k + t_k d^k - x \right) - \ell(1-t_k)\|d^k\|^2 - (1-t_k)\psi_{x^k}(d^k) - \frac{\mu}{2}\|x^k - x\|^2 \\
& - \frac{\nu}{2}\left(\|x^k + t_k d^k - x\|^2 + 2(1-t_k)(d^k)^\top (x^k + t_k d^k - x) + (1-t_k)^2 \|d^k\|^2 \right) + \frac{t_k \ell}{2}\|d^k\|^2 \\
\leq & -(\ell + \nu(1-t_k))(d^k)^\top \left(x^k + t_k d^k - x \right) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2 \\
& - (1-t_k)\psi_{x^k}(d^k) - (1-t_k)\left(\ell + \frac{\nu}{2} \right)\|d^k\|^2 + \frac{t_k(\ell + \nu(1-t_k))}{2}\|d^k\|^2 \\
\leq & -\frac{\ell + \nu(1-t_k)}{2t_k}\left(2(x^k - x)^\top (t_k d^k) + \|t_k d^k\|^2 \right) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2 \\
& - (1-t_k)\left(\psi_{x^k}(d^k) + \left(\ell + \frac{\nu}{2} \right)\|d^k\|^2 \right).
\end{aligned}
$$

Finally, the definition of $t_{\min}$ yields

$$
\begin{aligned}
&\sum_{i=1}^{m} \lambda_i^k \left(F_i(x^{k+1}) - F_i(x) \right) \\
\leq & \frac{\ell + \nu(1-t_{\min})}{2t_{\min}}\left(\|x^k - x\|^2 - \|x^{k+1} - x\|^2 \right) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2 \\
& - (1-t_{\min})\left(\psi_{x^k}(d^k) + \left(\ell + \frac{\nu}{2} \right)\|d^k\|^2 \right).
\end{aligned}
$$

where the second inequality comes from (17). □

The next theorem shows that the proximal gradient method in the convex case described in Algorithm 5.2 has a convergence rate of order $1/k$.

Theorem 5.2. *Suppose that there exists some nonempty set $\mathcal{J} \subseteq \{1, \ldots, m\}$ such that if $i \in \mathcal{J}$ then F_i is has a lower bound $F_i^{\min}$. Let $F^{\min} := \max_{i \in \mathcal{J}} F_i^{\min}$ and $F_0^{\max} := \max_{i \in \{1,\ldots,m\}} F_i(x^0)$. Let F_i be convex for all $i \in \{1, \ldots, m\}$. If the level set $\Omega_F(x^0) := \{x \in \mathbf{R}^n \mid F(x) \leq F(x^0)\}$ is bounded, then Algorithm 5.2 generates a sequence $\{x^k\}$ such that*

$$
w(x^k) \leq \frac{R}{2t_{\min}k},
$$

where $R := \ell \sup_{x \in \Omega_F(x^0)} \|x - x^0\| + (1 - t_{\min})[2(F_0^{\max} - F^{\min}) - \ell t_{\min} \sum_{j=0}^{k-1} \|d^k\|^2] < \infty$.

Proof. From Lemma 5.4, we have for all $x \in \mathbf{R}^n$,

$$\begin{aligned}
&\sum_{i=1}^{m} \lambda_i^k \left(F_i(x^{j+1}) - F_i(x)\right) \\
\leq &\frac{\ell}{2t_{\min}} \left(\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right) - (1 - t_{\min}) \left(\psi_{x^k}(d^k) + \ell\|d^k\|^2\right).
\end{aligned}$$

Adding up the above inequality from $j = 0$ to $j = k - 1$, we obtain

$$\begin{aligned}
&\sum_{j=0}^{k-1} \sum_{i=1}^{m} \lambda_i^j \left(F_i(x^j) - F_i(x)\right) \\
\leq &\frac{\ell}{2t_{\min}} \left(\|x^0 - x\|^2 - \|x^k - x\|^2\right) - (1 - t_{\min}) \sum_{j=0}^{k-1} \left(\psi_{x^j}(d^j) + \ell\|d^j\|^2\right).
\end{aligned}$$

From the condition (16) of the step size t_k, we have for all $p \in \mathcal{J}$

$$\begin{aligned}
&\sum_{j=0}^{k-1} \sum_{i=1}^{m} \lambda_i^j \left(F_i(x^j) - F_i(x)\right) \\
\leq &\frac{\ell}{2t_{\min}} \|x^0 - x\|^2 - (1 - t_{\min}) \sum_{j=0}^{k-1} \left(\frac{F_p(x^{j+1}) - F_p(x^j)}{t_j} + \frac{\ell}{2}\|d^j\|^2\right) \\
\leq &\frac{\ell}{2t_{\min}} \|x^0 - x\|^2 - (1 - t_{\min}) \sum_{j=0}^{k-1} \left(\frac{F_p(x^{j+1}) - F_p(x^j)}{t_{\min}} + \frac{\ell}{2}\|d^j\|^2\right) \\
= &\frac{\ell}{2t_{\min}} \|x^0 - x\|^2 - (1 - t_{\min}) \left(\frac{F_p(x^k) - F_p(x^0)}{t_{\min}} + \frac{\ell}{2} \sum_{j=0}^{k-1} \|d^j\|^2\right) \\
\leq &\frac{\ell}{2t_{\min}} \|x^0 - x\|^2 + (1 - t_{\min}) \left(\frac{F_0^{\max} - F_p^{\min}}{t_{\min}} - \frac{\ell}{2} \sum_{j=0}^{k-1} \|d^j\|^2\right) \\
\leq &\frac{\ell}{2t_{\min}} \|x^0 - x\|^2 + (1 - t_{\min}) \left(\frac{F_0^{\max} - F^{\min}}{t_{\min}} - \frac{\ell}{2} \sum_{j=0}^{k-1} \|d^j\|^2\right).
\end{aligned}$$

where the third inequality follows from the definition of $t_{\min}$. Let $\bar{\lambda}_i^{k-1} := \sum_{j=0}^{k-1} \lambda_i^j / k$. Then it follows that

$$\sum_{i=1}^{m} \bar{\lambda}_i^{k-1} \left(F_i(x^k) - F_i(x)\right) \leq \frac{\ell}{2t_{\min} k} \|x^0 - x\|^2 + \frac{1 - t_{\min}}{k} \left(\frac{F_0^{\max} - F^{\min}}{t_{\min}} - \frac{\ell}{2} \sum_{j=0}^{k-1} \|d^j\|^2\right).$$

Since $\bar{\lambda}_i^{k-1} \geq 0, \sum_{i=1}^m \bar{\lambda}_i^{k-1} = 1$, we see that

$$\min_{i\in\{1,\ldots,m\}} \Big(F_i(x^k) - F_i(x)\Big) \leq \frac{\ell}{2t_{\min}k}\|x^0 - x\|^2 + \frac{1-t_{\min}}{k}\left(\frac{F_0^{\max} - F^{\min}}{t_{\min}} - \frac{\ell}{2}\sum_{j=0}^{k-1}\|d^j\|^2\right).$$

Taking the supremum in the level set $\Omega_F(x^0)$ yields

$$\sup_{x\in\Omega_F(x^0)} \min_{i\in\{1,\ldots,m\}} \Big(F_i(x^k) - F_i(x)\Big) \leq \frac{R}{2t_{\min}k}.$$

Now, Lemma 5.1 and (16) imply that $\{F_i(x^k)\}$ is decreasing, so we conclude that

$$w(x^k) \leq \frac{R}{2t_{\min}k}.$$

□

5.2 Proximal gradient methods without line searches

When we set $\ell > L_{\max}$, we can fix the step size $t_k = 1$ for each iteration k. The algorithm without line searches is described as follows:

Algorithm 5.3 (The proximal gradient method without line searches)**.**

Step 1: Choose $\ell > L_{\max}/2$, $x^0 \in \mathbf{R}^n$ *and set* $k \coloneqq 0$.

Step 2: Compute d^k *by solving subproblem* $d^k = \operatorname{argmin}_{d\in\mathbf{R}^n} \left[\psi_{x^k}(d) + \frac{\ell}{2}\|d\|^2\right]$.

Step 3: If $d^k = 0$, *then stop.*

Step 4: Set $x^{k+1} \coloneqq x^k + d^k$, $k \coloneqq k+1$, *and go to Step 2.*

(a) The non-convex case

First, we analyze the convergence rate when f_i is non-convex for all $i \in \{1,\ldots,m\}$. The next theorem shows that Algorithm 5.3 has a convergence rate of order $1/k$.

Theorem 5.3. *Suppose that there exists some nonempty set* $\mathcal{J} \subseteq \{1,\ldots,m\}$ *such that if* $i \in \mathcal{J}$ *then* F_i *has a lower bound* $F_i^{\min}$. *Let* $F^{\min} \coloneqq \max_{i\in\mathcal{J}} F_i^{\min}$ *and* $F_0^{\max} \coloneqq \max_{i\in\{1,\ldots,m\}} F_i(x^0)$. *Then, the Algorithm 5.3 generates a sequence* $\{x^k\}$ *such that*

$$\min_{0\leq j\leq k-1} u_1(x^j) \leq \frac{F_0^{\max} - F^{\min}}{k\min\left\{1, \frac{1}{L_{\max}}\right\}}.$$

Proof. Let $i \in \mathcal{J}$. From the Lipschitz continuity of ∇f_i, we have

$$\begin{aligned}
F_i(x^{k+1}) - F_i(x^k) &\leq \nabla f_i(x^k)^\top d^k + g_i(x^{k+1}) - g_i(x^k) + \frac{L_i}{2}\|d^k\|^2 \\
&\leq \nabla f_i(x^k)^\top d^k + g_i(x^{k+1}) - g_i(x^k) + \frac{L_{\max}}{2}\|d^k\|^2 \\
&\leq \max_{i\in\{1,\ldots,m\}} \left\{ \nabla f_i(x^k)^\top d^k + g_i(x^{k+1}) - g_i(x^k) + \frac{L_{\max}}{2}\|d^k\|^2 \right\} \\
&= -u_{L_{\max}}(x^k) \\
&\leq -\min\left\{1, \frac{1}{L_{\max}}\right\} u_1(x^k),
\end{aligned}$$

where the equality follows from the definition (10) of $u_{L_{\max}}$, and the last inequality comes from the statement 3 of Proposition 3.1. Adding up the above inequality from $k = 0$ to $k = \tilde{k} - 1$ yields that

$$\begin{aligned}
F_i(x^{\tilde{k}}) - F_i(x^0) &\leq -\sum_{k=0}^{\tilde{k}-1} \left\{1, \frac{1}{L_{\max}}\right\} u_1(x^k) \\
&\leq -\tilde{k}\min\left\{1, \frac{1}{L_{\max}}\right\} \min_{0\leq k\leq \tilde{k}-1} u_1(x^j).
\end{aligned}$$

Thus, we get

$$\min_{0\leq k\leq \tilde{k}-1} u_1(x^k) \leq \frac{F_0^{\max} - F^{\min}}{\tilde{k}\min\left\{1, \frac{1}{L_{\max}}\right\}}.$$

□

(b) The convex and strongly convex cases

We start with proving the following lemma. Note that we add an assumption $\ell > L_{\max}$.

Lemma 5.5. *Let f_i and g_i have convexity parameters $\mu_i \in \mathbf{R}$ and $\nu_i \in \mathbf{R}$, respectively, and write $\mu \coloneqq \min_{i\in\{1,\ldots,m\}} \mu_i$ and $\nu \coloneqq \min_{i\in\{1,\ldots,m\}} \nu_i$. If $\ell > L_{\max}$, then for all $x \in \mathbf{R}^n$ it follows that*

$$\sum_{i=1}^m \lambda_i^k \left(F_i(x^{k+1}) - F_i(x)\right) \leq \frac{\ell}{2}\left(\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right) - \frac{\nu}{2}\|x^{k+1} - x\|^2 - \frac{\mu}{2}\|x^k - x\|^2,$$

where λ_i^k satisfies the following conditions:

$$\begin{cases} \text{There exists } \eta_i \in \partial g_i(x^k + d^k) \text{ such that } \displaystyle\sum_{i=1}^m \lambda_i^k (\nabla f_i(x^k) + \eta_i) + \ell d^k = 0, & (20) \\ \displaystyle\sum_{i=1}^m \lambda_i^k = 1, \quad \lambda_i^k \geq 0 \quad \left(i \in \mathcal{I}_{x^k}(d^k)\right), \quad \lambda_i^k = 0 \quad \left(i \notin \mathcal{I}_{x^k}(d^k)\right), & (21) \end{cases}$$

where

$$\mathcal{I}_x(d) \coloneqq \{i \in \{1, \dots, m\} \mid \psi_x(d) = \nabla f_i(x)^\top d + g_i(x + d) - g_i(x)\}.$$

Proof. Since $\ell > L_{\max}$, ∇f_i is Lipschitz continuous with Lipschitz constant ℓ for each $i \in \{1, \dots, m\}$. Therefore, from Lemma 3.2 we have

$$F_i(x^{k+1}) - F_i(x^k) \leq \nabla f_i(x^k)^\top (x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\|d^k\|^2.$$

Let $\eta_i \in \partial g_i(x^k + d^k)$. The convexity of f_i with modulus μ_i gives

$$\begin{aligned} &F_i(x^{k+1}) - F_i(x) \\ \leq &\nabla f_i(x^k)^\top (x^k - x) - \frac{\mu_i}{2}\|x^k - x\|^2 + g_i(x^k) - g_i(x) \\ &+ \nabla f_i(x^k)^\top (x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k) + \frac{\ell}{2}\|d^k\|^2 \\ \leq &\nabla f_i(x^k)^\top (x^k + d^k - x) + g_i(x^k + d^k) - g_i(x) - \frac{\mu}{2}\|x^k - x\|^2 + \frac{\ell}{2}\|d^k\|^2 \\ \leq &\nabla f_i(x^k)^\top (x^k + d^k - x) + \eta_i^\top (x^k + d^k - x) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2 + \frac{\ell}{2}\|d^k\|^2, \end{aligned}$$

where the equality follows from the definition of μ and the fact that $x^{k+1} = x^k + d^k$, and the last inequality comes from the convexity of g_i. Multiplying the above inequality by λ_i^k and summing for all $i \in \{1, \dots, m\}$, we obtain

$$\begin{aligned} &\sum_{i=1}^m \lambda_i^k \left(F_i(x^{k+1}) - F_i(x)\right) \\ \leq &\sum_{i=1}^m \lambda_i^k \Big(\nabla f_i(x^k)^\top (x^k + d^k - x) + \eta_i^\top (x^k + d^k - x) \\ &- \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2 + \frac{\ell}{2}\|d^k\|^2\Big) \\ = &-\ell (d^k)^\top (x^k + d^k - x) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2 + \frac{\ell}{2}\|d^k\|^2 \\ = &-\frac{\ell}{2}\left(2(d^k)^\top (x^k - x) + \|d^k\|^2\right) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2 \\ = &\frac{\ell}{2}\left(\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2, \end{aligned}$$

where (20) and (21) give the first inequality. □

The next theorem shows that the proximal gradient method without line search described in Algorithm 5.3 has a convergence rate of order $1/k$ in the convex case.

Theorem 5.4. *Let F_i be convex for all $i \in \{1, \ldots, m\}$. If the level set $\Omega_F(x^0) \coloneqq \{x \in S \mid F(x) \leq F(x^0)\}$ is bounded and $\ell > L_{\max}$, then Algorithm 5.3 generates a sequence $\{x^k\}$ such that*

$$w(x^k) \leq \frac{\tilde{R}}{2k},$$

where $\tilde{R} \coloneqq \ell \sup_{x \in \Omega_F(x^0)} \|x - x^0\| < \infty$.

Proof. From Lemma 5.5, we have for all $x \in \mathbf{R}^n$

$$\sum_{i=1}^{m} \lambda_i^k \left(F_i(x^{k+1}) - F_i(x)\right) \leq \frac{\ell}{2} \left(\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right).$$

Adding up the above inequality from $j = 0$ to $j = k - 1$, we obtain

$$\begin{aligned}\sum_{j=1}^{k-1} \sum_{i=1}^{m} \lambda_i^j \left(F_i(x^{k+1}) - F_i(x)\right) &\leq \frac{\ell}{2} \left(\|x^0 - x\|^2 - \|x^{k+1} - x\|^2\right) \\ &\leq \frac{\ell}{2} \|x^0 - x\|^2.\end{aligned}$$

Let $\bar{\lambda}_i^{k-1} \coloneqq \sum_{j=0}^{k-1} \lambda_i^j / k$. Then, it follows that

$$\sum_{i=1}^{m} \bar{\lambda}_i^{k-1} \left(F_i(x^k) - F_i(x)\right) \leq \frac{\ell}{2k} \|x^0 - x\|^2.$$

Since $\bar{\lambda}_i^{k-1} \geq 0, \sum_{i=1}^{m} \bar{\lambda}_i^{k-1} = 1$, we see that

$$\min_{i \in \{1, \ldots, m\}} \left(F_i(x^k) - F_i(x)\right) \leq \frac{\ell}{2k} \|x^0 - x\|^2.$$

Taking the supremum in the level set $\Omega_{F(x^0)}$ yields

$$\sup_{x \in \Omega_{F(x^0}} \left(F_i(x^k) - F_i(x)\right) \leq \frac{\tilde{R}}{2k}.$$

Since $\{F_i(x^k)\}$ is decreasing, we have

$$w(x^k) \leq \frac{\tilde{R}}{2k}.$$

□

Moreover, if we assume strong convexity, then Algorithm 5.3 converges linearly.

Theorem 5.5. *Let f_i and g_i have convexity parameters $\mu_i \in \mathbf{R}$ and $\nu_i \in \mathbf{R}$, respectively, and write $\mu \coloneqq \min_{i\in\{1,\dots,m\}} \mu_i$ and $\nu \coloneqq \min_{i\in\{1,\dots,m\}} \nu_i$. If $\ell > L_{\max}$, then there exists a Pareto optimal point $x^* \in \mathbf{R}^n$ such that for all iteration k,*

$$\|x^{k+1} - x^*\| \leq \sqrt{\frac{\ell - \mu}{\ell + \nu}}\|x^k - x^*\|.$$

Thus, we have

$$\|x^k - x^*\| \leq \left(\sqrt{\frac{\ell - \mu}{\ell + \nu}}\right)^k \|x^0 - x^*\|.$$

Proof. Because each F_i is strongly convex, $\{x^k\}$ has an accumulation point $x^* \in \mathbf{R}^n$. Note that x^* is a Pareto optimal point [9, Lemma 2.2 and Theorem 4.2]. Now, from Lemma 5.5, we have

$$\sum_{i=1}^m \lambda_i^k \left(F_i(x^{k+1}) - F_i(x^*)\right) \leq \frac{\ell}{2}\left(\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2.$$

Since the left hand side is nonnegative because of Lemma 5.1 and (16), we obtain

$$0 \leq \frac{\ell}{2}\left(\|x^k - x\|^2 - \|x^{k+1} - x\|^2\right) - \frac{\mu}{2}\|x^k - x\|^2 - \frac{\nu}{2}\|x^{k+1} - x\|^2,$$

which is equivalent to

$$\|x^{k+1} - x^*\| \leq \sqrt{\frac{\ell - \mu}{\ell + \nu}}\|x^k - x^*\|.$$

□

6 Conclusion

We proposed two new merit functions and clarified the associate properties of error bounds and level-boundedness. By using them, we showed a sufficient condition on the existence and boundedness of the weakly Pareto optimal set. Moreover, by considering these merit functions we analyzed the convergence rates of proximal gradient methods. We showed that the algorithms have convergence rates of order $1/k$, $1/k$, and r^k for some $r \in (0, 1)$, respectively for non-convex, convex and strongly convex cases.

In single-objective optimization, for convex cases we can obtain convergence rate of $O(1/k^2)$ with accelerated methods such as [1]. For future research, we may consider extending these topics to multiobjective optimization.

Acknowledgments

First of all, the author would like to express his sincere gratitude to Associate Professor Ellen Hidemi Fukuda for her kind guidance and invaluable discussions in this study, and constructive criticisms in the writing of the manuscript. Also, the author wishes to tender his acknowledgement to Professor Nobuo Yamashita for his kind guidance and helpful advices. In addition, the author is grateful to Program-Specific Associate Professor Hiroyuki Sato for his comments and support. Finally, the author thanks to all members of Yamashita's Laboratory, my friends, and my family for their encouragements.

References

[1] 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.

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

[3] Bertsekas, D. P.: *Convex Analysis and Optimization*, Athena Scientific, Belmont, Mass., 2003.

[4] Dutta, J., Kesarwani, P. and Gupta, S.: Gap functions and error bounds for nonsmooth convex vector optimization problem, *A Journal of Mathematical Programming and Operations Research*, Vol. 66 (2017), 1807–1836.

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

[6] Fliege, J., Vaz, A. I. F. and Vicente, L. N.: Complexity of gradient descent for multiobjective optimization, to appear in *Optimization Methods and Software*.

[7] Fukuda, E. H. and Graña Drummond, L. M.: A survey on multiobjective descent methods, *Pesquisa Operacional*, Vol. 34 (2014), 585–620.

[8] Liu, C. G., Ng, K. F. and Yang, W. H.: Merit functions in vector optimization, *Mathematical Programming*, Vol. 119 (2009), 215–237.

[9] 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.

Master's Thesis

Merit functions for multiobjective optimization and convergence rates analysis of multiobjective proximal gradient methods

Guidance

Associate Professor Ellen Hidemi FUKUDA
Professor Nobuo YAMASHITA

Hiroki TANABE

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University

July 2019

Merit functions for multiobjective optimization and convergence rates analysis of multiobjective proximal gradient methods

Hiroki TANABE

Abstract

Many descent algorithms for multiobjective optimization have been developed in the last two decades. However, differently from the scalar-valued optimization case, there exist few results related to the existence or the boundedness of Pareto optimal solutions. Moreover, studies about the convergence rates of these algorithms are still insufficient. In this paper, we first present two new merit functions for nonlinear multiobjective optimization, which extend the one defined for linear multiobjective optimization. These functions return zero at the solutions of the original problem and strictly positive values otherwise. Furthermore, by examining the properties of these merit functions, we show sufficient conditions for the existence of weakly Pareto optimal solutions, and for the boundedness of Pareto optimal sets. Finally, by using these functions, we analyze the convergence rates of the recently proposed multiobjective proximal gradient methods. We show that both methods with and without line searches have sublinear rate of convergence for non-convex and convex cases. We also prove that the algorithm without line searches converges linearly in the strongly convex case.