Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 20, No. 4, pp. 809–826, 2024

A PROXIMAL GRADIENT METHOD WITH BREGMAN DISTANCE IN MULTI-OBJECTIVE OPTIMIZATION*

Kangming Chen, Ellen H. Fukuda and Nobuo Yamashita†

Abstract: A multi-objective proximal gradient method was proposed recently as a suitable descent method for composite multi-objective optimization problems. However, the method solves subproblems using only Euclidean distances, and it requires the differentiable part of each objective to have a Lipschitz continuous gradient, which limits its application. In this paper, we propose an extension of this method, by using Bregman distances and requiring a less demanding assumption called relative smoothness. We also consider two stepsize strategies: the constant stepsize and the backtracking procedure. In both cases, we prove global convergence in the sense of Pareto stationarity and analyze the convergence rate through a merit function.

Key words: *multi-objective descent methods, accelerated proximal gradient methods, first-order methods, Bregman distances, relative smoothness*

Mathematics Subject Classification: *90C29, 90C30*

1 Introduction

In this paper, we consider the following unconstrained multi-objective optimization problem:

$$\begin{aligned} \min \quad & F(x) \\ \text{s.t.} \quad & x \in \mathbf{R}^n, \end{aligned} \tag{1.1}$$

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

$$F_i(x) := f_i(x) + g_i(x), \quad i = 1, \ldots, m,$$

where $f_i\colon \mathbf{R}^n \to \mathbf{R}$ is continuously differentiable, and $g_i\colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is proper, closed and convex. Problem (1.1) has many applications, including constrained multi-objective problems (i.e., when all g_i is the indicator function of a same set), machine learning [10, 27], and robust optimization [35].

A common solution strategy for multi-objective optimization is the scalarization approach [26], where the problem is converted to a single objective one using some parameters. The disadvantage of this method is that the parameters are not known in advance.

*This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation (JPMJFS2123), and the Grant-in-Aid for Scientific Research (C) (19K11840 and 21K11769) from Japan Society for the Promotion of Science.

†Corresponding author

DOI: https://doi.org/10.61208/pjo-2024-012

Another well-known method is the metaheuristics, but no theoretical guarantee of convergence exists in this case. As alternatives to these methods, in the last two decades, many multi-objective descent methods have been proposed. The descent methods generate descent directions by solving easy (i.e., strongly convex and quadratic) subproblems, and that generalize well-known single-valued optimization methods [25].

For instance, for unconstrained problems, the steepest descent and the Newton methods were proposed in [22] and [21], respectively. Also, the projected gradient [23, 24, 29], the subgradient [20], the proximal point [8, 12, 19, 28], the conditional gradient method [2, 3] and many other methods were proposed in the literature [7, 13, 32]. For problems like (1.1), which are often called composite problems, Tanabe, Fukuda, and Yamashita [35] proposed a proximal gradient method. In that paper, the search direction is computed by solving a subproblem, that considers the first-order approximation of the objectives, using the gradient only for the differentiable f_i, plus a regularization term that uses the Euclidean distance. By assuming that f_i has Lipschitz continuous gradients, they prove that every accumulation point of the generated sequence, if it exists, is a Pareto stationary point. Moreover, the rate of convergence was also proved in [36].

An intuitive extension of the former methods involves substituting the Euclidean distance with a broader, distance-like metric, thereby enhancing the method's applicability and flexibility. The notion of Bregman distance (or Bregman divergence) was first introduced in [14], which proposed an iterative algorithm to solve certain convex optimization problems involving regularization, known as the Bregman method, and re-emerged in [15]. Unlike traditional metrics such as the Euclidean distance, the Bregman distances may not be symmetric and do not necessarily satisfy the triangle inequality. However, they do adhere to a generalization of the Pythagorean theorem, allowing for the application of optimization theory techniques in a more general setting. For this reason, the Bregman distances began to be considered instead of the Euclidean one. They are frequently employed to address constrained optimization problems, especially in alleviating ill-conditioned solutions of subproblems [16, 39].

Furthermore, much research has been done, concerning optimization methods based on Bregman distances for proximal point methods [16] and proximal gradient methods [11, 38]. For multi-objective optimization, few methods that consider these distances exist, except for some proximal point methods [19, 34]. Based on this, we propose the multi-objective proximal gradient method with Bregman distance, by modifying the search direction used in [35]. Depending on the chosen Bregman distance, the subproblem can more precisely approximate the original function, potentially enhancing the accuracy of the solution and possibly reducing the total number of iterations required. In particular, the Kullback–Leibler divergence proves advantageous in optimizations over the unit simplex [6]. Moreover, a proximal gradient method with a proper Bregman distance is suitable for Poisson linear inverse problems [4], where the usual proximal gradient cannot be applied. Further examples of applications of Bregman distances can be found in [30, 39].

In this paper, we also use the concept of relative smoothness. A similar notion was proposed in [4] as a new descent lemma without Lipschitz gradient continuity, where the reference function is required to be a Legendre function. Recently, Lu et al. gave the definition of relative smoothness and relative strong convexity in [31] showing that it is less strict than the usual Lipschitz continuity of the gradients assumption. In this case, differently from [4], the reference function is not required to be strictly convex. Similarly, in this work, we will assume the less restrictive relative smoothness for the differentiable part of the objective function, making adaptations to deal with multi-objectives. Furthermore, we will consider two types of stepsizes, and for both of them, we will show convergence to

Pareto stationary points as well as convergence rates.

During the writing of this paper, we noted an unpublished paper by Chen et al. [17], which shares similarities with the current work. It is important to emphasize that our research was independently done, and grounded in the work presented in the first author's master thesis [18]. However, our papers still differ in many aspects. While the paper [17] considered vector optimization, they did not incorporate the possible nondifferentiable function g into their study. Also, our research offers a more refined global convergence since we do not require strict convexity and Lipschitz gradient continuity of the reference function. Similarly, we also found an unpublished paper by Ansary and Dutta [1], which was also done independently. In this case, they considered composite objective functions, but they require strong convexity of the reference function, and they did not discuss convergence rates.

The outline of this paper is as follows. In Section 2, we recall the proximal gradient method for multi-objective problems, the definition of the Bregman function, and some preliminary materials. In Section 3, we propose a proximal gradient method with Bregman distance for multi-objective optimization, considering both the constant stepsizes and the backtracking procedure. Section 4 contains the proof of global convergence to Pareto stationary points. In addition, we prove the convergence rates for convex and strongly convex problems.

2 Preliminaries

In this section, we will recall the multi-objective proximal gradient method proposed in [35], as well as the basic notions of Bregman distances and the so-called relative smoothness. Before that, let us first present some notations used in this paper. We denote the Euclidean inner product as $\langle \cdot, \cdot \rangle$, and the Euclidean norm as $\| \cdot \|$. The interior, the boundary and the closure of a set S are written as $\operatorname{int}(S)$, $\operatorname{bd}(S)$ and $\operatorname{cl}(S)$, respectively. We also define the relation $\preceq$ ($\prec$) in $\mathbf{R}^m$ as $u \preceq v$ ($u \prec v$) if and only if $u_i \leq v_i$ ($u_i < v_i$) for all $i = 1, \ldots, m$.

For a given scalar-valued function $q \colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ and a vector-valued function $r \colon \mathbf{R}^n \to (\mathbf{R} \cup \{+\infty\})^m$, we use $\nabla q(x) \in \mathbf{R}^n$ and $Jr(y) \in \mathbf{R}^{m \times n}$ to denote, respectively, the gradient of q at $x \in \operatorname{dom} q$, and the Jacobian matrix of r at $y \in \operatorname{dom} r$. Here, $\operatorname{dom} q := \{x \in \mathbf{R}^n \colon q(x) < +\infty\}$ and $\operatorname{dom} r := \{y \in \mathbf{R}^n \colon r(y) \prec +\infty\}$ denote the effective domain of q and r, respectively. The directional derivative of q at x in the direction $d \in \mathbf{R}^n$, if it exits, is given by

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

We denote the set of subdifferentiable points of q as $\operatorname{dom} \partial q := \{x \in \mathbf{R}^n \colon \partial q(x) \neq \emptyset\}$, where $\partial q(x)$ means the subdifferential of q at x in the sense of convex analysis, i.e.,

$$\partial q(x) := \left\{ s \in \mathbf{R}^n \colon q(y) \geq q(x) + \langle s, y - x \rangle \text{ for all } y \in \mathbf{R}^n \right\}.$$

Note that for a proper function q, we have $\partial q(x) = \emptyset$ when $x \notin \operatorname{dom} q$.

2.1 Pareto optimality and related works

Let us first introduce the concept of Pareto optimality, in particular, for the multi-objective optimization problem (1.1). Recall that $x^* \in \mathbf{R}^n$ is a *Pareto optimal* point for F, if there is no $x \in \mathbf{R}^n$ such that $F(x) \preceq F(x^*)$ and $F(x) \neq F(x^*)$. Also, $x^* \in \mathbf{R}^n$ is a *weakly Pareto optimal* point for F, if there is no $x \in \mathbf{R}^n$ such that $F(x) \prec F(x^*)$. The set of all (weakly) Pareto optimal values is called (weakly) Pareto frontier. It is known that Pareto optimal

points are always weakly Pareto optimal, but the converse is not always true. We also recall that $\bar{x}$ is *Pareto stationary* (or *critical*), if and only if,

$$\max_{i=1,\dots,m} F_i'(\bar{x}; d) \geq 0 \quad \text{for all } d \in \mathbf{R}^n.$$

The above definition generalizes the one that was given in [22] for differentiable problems. Moreover, its relation with (weakly) Pareto optimality can be seen in [35]. For the sake of completeness, we state below such relations.

Lemma 2.1 ([35, Lemma 2.2])**.** *The following assertions hold:*

1. *If x is weakly Pareto optimal for F, then x is Pareto stationary.*
2. *Assume that every component F_i of F is convex. If x is Pareto stationary for F, then x is weakly Pareto optimal.*
3. *Assume that every component F_i of F is strictly convex. If x is Pareto stationary for F, then x is Pareto optimal.*

Let us now discuss the multi-objective proximal gradient method, proposed in [35]. As many multi-objective descent methods, it generates a sequence $\{x^k\}$ iteratively with the following procedure:

$$x^{k+1} := x^k + t_k d^k,$$

where d^k is the search direction and t_k is the stepsize. At every iteration k, the direction d^k is computed by solving the following unconstrained single-objective problem:

$$d^k := \operatorname*{argmin}_{d \in \mathbf{R}^n} \left(\tilde{\psi}_{x^k}(d) + \frac{\ell}{2}\|d\|^2 \right), \tag{2.1}$$

where $\ell > 0$ and $\tilde{\psi}_{x^k} \colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is defined by

$$\tilde{\psi}_{x^k}(d) := \max_{i=1,\dots,m} \left(\nabla f_i(x^k)^\top d + g_i(x^k + d) - g_i(x^k) \right),$$

assuming $g_i(x^k) < +\infty$. In other words, the direction is defined by solving a problem, with a maximum of the first-order approximations of F_i, using the gradient only for the differentiable part f_i, plus a regularization term. In this case, the traditional Euclidean norm regularization was used, with some $\ell > 0$. We also notice that (2.1) is well-defined since the objective function of this subproblem is strongly convex.

In [35], a method with and without line searches was considered. In the first case, the stepsize was fixed as $t_k = 1$ in all iterations, while in the second case, a backtracking procedure with Armijo condition was used. For both cases, it was proved that each accumulation point of the sequence generated by the method, if it exists, is Pareto stationary [35, Theorems 4.2 and 4.3]. Moreover, the convergence rate was also established in [36]. Here, we will replace the Euclidean norm regularization used in the above subproblem with a distance-like function, called Bregman distance.

2.2 Bregman distances and relative smoothness

In this section, we review the basic properties of Bregman distances, as well as the definition of relative smoothness and relative strong convexity, that will be considered in the paper. These definitions take a function $\omega \colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ as a reference, and with the following properties: it is proper, closed, convex, and differentiable over dom $\partial\omega$. Let us start with the definition of Bregman distance [14].

Definition 2.2. The *Bregman distance* associated with ω is the function $B_\omega\colon \operatorname{dom}\omega \times \operatorname{dom}\partial\omega \to \mathbf{R}$ given by

$$B_\omega(x,y) = \omega(x) - \omega(y) - \langle \nabla\omega(y), x-y\rangle \quad \text{for all } x \in \operatorname{dom}\omega,\ y \in \operatorname{dom}\partial\omega.$$

Note that the convexity of ω implies

$$B_\omega(x,y) \geq 0 \quad \text{for all } x \in \operatorname{dom}\omega, y \in \operatorname{dom}\partial\omega.$$

Moreover, the definition of Bregman distance give

$$x = y \quad \Longrightarrow \quad B_\omega(x,y) = 0. \tag{2.2}$$

It should be noted that, under the assumption of strict convexity for ω, the converse implication holds as well.

Example 2.3 (Euclidean distance)**.** If $\omega(x) = \frac{1}{2}\|x\|^2$, then $B_\omega(x,y) = \frac{1}{2}\|x-y\|^2$. In this case, $B_\omega(x,y) = B_\omega(y,x)$, which is not necessarily true in general.

Example 2.4 (Negative entropy (or Kullback–Leibler divergence))**.** If $\omega(x) = \sum_{i=1}^n x_i \ln x_i$ when $x \succeq 0$, and $\omega(x) = +\infty$ otherwise (with the convention $0\ln 0 = 0$), then $B_\omega(x,y) = \sum_{i=1}^n x_i \ln \frac{x_i}{y_i} - \sum_{i=1}^n (x_i - y_i)$.

We also list some lemmas that will be used in the paper. Lemma 2.6 is essential in analyzing the convergence of the proximal gradient methods with Bregman distance, which can be also proved based on Lemma 2.5.

Lemma 2.5 (Three-points identity)**.** *[16, Lemma 3.1] Take $a, b \in \operatorname{dom}\partial\omega$ and $c \in \operatorname{dom}\omega$. Then the following equality holds:*

$$\langle \nabla\omega(b) - \nabla\omega(a), c-a\rangle = B_\omega(c,a) + B_\omega(a,b) - B_\omega(c,b).$$

Lemma 2.6 ([16, Lemma 3.2])**.** *For any proper closed convex function $\theta : \mathbf{R}^n \to \mathbf{R}\cup\{+\infty\}$ and any $z \in \operatorname{dom}\partial\omega$, if ω is differentiable at $z_+ = \operatorname*{argmin}_{x\in\operatorname{dom}\omega}(\theta(x) + B_\omega(x,z))$, then*

$$\theta(x) + B_\omega(x,z) \geq \theta(z_+) + B_\omega(z_+,z) + B_\omega(x,z_+) \quad \textit{for all } x \in \operatorname{dom}\omega.$$

Many optimization methods, including the multi-objective proximal gradient method proposed in [35], assume the gradients of the objectives to be Lipschitz continuous. For a scalar-valued function q, it means that there exists some $\tilde{L}_q$ such that

$$\|\nabla q(x) - \nabla q(y)\| \leq \tilde{L}_q\|x-y\| \quad \text{for all } x,y \in \operatorname{int}(\operatorname{dom} q).$$

However, this is a rather strict condition. We refer to [31] for examples of convex differentiable functions that do not satisfy such a condition. Moreover, even if it is satisfied, the Lipschitz constant may be too large, making its practical usage difficult. Compared to the gradient Lipschitz condition, the following notion of relative smoothness, using a function ω as a reference function, is shown to be less restrictive.

Definition 2.7 ((Relative smoothness) [31, Definition 1.1])**.** Let ω be convex and differentiable on $\operatorname{dom}\omega$. A function q is called *L_q-smooth relative* to ω on $\operatorname{dom}\omega$ if for any $x, y \in \operatorname{int}(\operatorname{dom}\omega)$, there exists a scalar L_q such that

$$q(x) \leq q(y) + \langle \nabla q(y), x-y\rangle + L_q B_\omega(x,y).$$

Definition 2.8 ((Relative strongly convexity) [31, Definition 1.2])**.** Let ω be convex and differentiable on $\operatorname{dom}\omega$. A function q is called μ_q-*strongly convex relative* to ω on $\operatorname{dom}\omega$ if for any $x, y \in \operatorname{int}(\operatorname{dom}\omega)$, there exists a scalar $\mu_q \geq 0$ such that

$$q(x) \geq q(y) + \langle \nabla q(y), x - y\rangle + \mu_q B_\omega(x, y).$$

Note that the definition of relative smoothness gives an upper approximation of q that is similar to the so-called descent lemma. Moreover, in [31, Section 2], many classes of optimization problems were listed, showing constructions of reference functions ω that make the objective function smooth relative to an easily determined constant.

3 The Multi-Objective Proximal Gradient Method with Bregman Distance

In this section, we explain in detail the proposed proximal gradient method with Bregman distance. From now on, we suppose that the following assumption holds.

Assumption 3.1. Let $\omega\colon \mathbf{R}^n \to \mathbf{R}\cup\{+\infty\}$ be a reference function that is proper, closed, convex, and differentiable over $\operatorname{dom}\partial\omega$. We assume that f_i is L^i-smooth relative to ω and that $\operatorname{dom} g \subseteq \operatorname{dom}\omega$. Also, the level set $\{x \in \operatorname{dom} g \mid \omega(x) \leq \alpha\}$ is compact for all $\alpha \in \mathbf{R}$.

We recall that in many works, ω is assumed to be a Legendre function, i.e., it is proper, closed, essentially smooth and strictly convex. Also, essentially smooth means $\operatorname{dom}\partial\omega \neq \emptyset$ or, equivalently, $\operatorname{int}(\operatorname{dom}\omega) \neq \emptyset$ with ω differentiable on $\operatorname{int}(\operatorname{dom}\omega)$ and $\|\nabla\omega(x^k)\| \to +\infty$ when $\{x^k\} \subset \operatorname{int}(\operatorname{dom}\omega)$, $x^k \to x \in \operatorname{bd}(\operatorname{dom}\omega)$ (see [33, Theorem 26.1]). In this work, we do not require strict convexity of ω unless it is explicitly specified. Observe also that compactness of the defined level set is guaranteed, for instance, when $\operatorname{dom} g$ is compact, or ω is level bounded.

Using the relatively smooth constants, we also define

$$L := \max_{i=1,\ldots,m} L^i. \tag{3.1}$$

Now, let us notice that when the stepsize t_k is equal to 1, then $d^k = x^{k+1} - x^k$. Thus, in this case, the subproblem (2.1) can be written as

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathbf{R}^n} \left(\psi_{x^k}(x) + \frac{\ell}{2}\|x - x^k\|^2\right), \tag{3.2}$$

where $\ell > 0$ and $\psi_{x^k}\colon \mathbf{R}^n \to \mathbf{R}\cup\{+\infty\}$ is defined by

$$\psi_{x^k}(x) := \max_{i=1,\ldots,m} \Big(\nabla f_i(x^k)^\top\big(x - x^k\big) + g_i(x) - g_i(x^k)\Big). \tag{3.3}$$

Based on this, in each iteration k of our proposed method, we consider the following subproblem:

$$x^{k+1} = p_{L_k}(x^k) \in \operatorname*{argmin}_{x\in\mathbf{R}^n}\Big(\psi_{x^k}(x) + L_k B_\omega\big(x, x^k\big)\Big), \tag{3.4}$$

where ψ_{x^k} is defined in (3.3), $L_k > 0$, and B_ω is the Bregman distance associated with ω, supposing Assumption 3.1. Clearly, this subproblem is equivalent to (3.2) when $\omega(x) = \|x\|^2/2$ (if $L_k = \ell$ for all k). We further define its optimal value as follows:

$$\theta(x^k) := \min_{x\in\mathbf{R}^n}\Big(\psi_{x^k}(x) + L_k B_\omega\big(x, x^k\big)\Big) = \psi_{x^k}(p_{L_k}(x^k)) + L_k B_\omega\big(p_{L_k}(x^k), x^k\big). \tag{3.5}$$

Now, we observe that the definition of ψ_{x^k} gives $\psi_{x^k}(x^k) = 0$, and $B_\omega\big(x^k, x^k\big) = 0$ holds. This means that

$$\theta(x^k) \leq 0 \quad \text{for all } k.$$

We now prove that $p_{L_k}(x^k)$ is in fact well-defined.

Proposition 3.1. *The term $p_{L_k}(x^k)$ given in (3.4) is well-defined for all k and $p_{L_k}(x^k) \in \operatorname{dom} \partial\omega$. If in addition ω is strongly convex, the minimizer $p_{L_k}(x^k)$ is unique.*

Proof. The optimization problem given in (3.5) can be written as

$$\begin{aligned}
&\min_{x\in\mathbf{R}^n} \max_{i=1,\dots,m} \left(\nabla f_i(x^k)^\top (x - x^k) + g_i(x) - g_i(x^k) + L_k B_\omega(x, x^k)\right) \\
&= \min_{x\in\mathbf{R}^n} \max_{i=1,\dots,m} \left(\left\langle \nabla f_i(x^k) - L_k \nabla\omega(x^k), x - x^k \right\rangle + g_i(x) - g_i(x^k)\right. \\
&\qquad\qquad \left. + L_k\omega(x) - L_k\omega(x^k)\right),
\end{aligned}$$

where the equality follows from the definition of $B_\omega(x, x^k)$. Dividing the objective function of the above problem by L_k, we obtain the equivalent problem:

$$\min_{x\in\mathbf{R}^n} \max_{i=1,\dots,m} \left(\left\langle \frac{1}{L_k}\nabla f_i(x^k) - \nabla\omega(x^k), x \right\rangle + \frac{1}{L_k} g_i(x) - \frac{1}{L_k} g_i(x^k) + \omega(x) \right).$$

For some k, defining

$$\varphi(x) := \max_{i=1,\dots,m} \left(\left\langle \frac{1}{L_k}\nabla f_i(x^k) - \nabla\omega(x^k), x \right\rangle + \frac{1}{L_k} g_i(x) - \frac{1}{L_k} g_i(x^k) \right),$$

the problem can be written as

$$\min_{x\in\mathbf{R}^n} \Psi(x) := \big(\varphi(x) + \omega(x)\big).$$

The function Ψ is closed since g, ω, and the maximum of closed functions are closed. It is also proper because $\operatorname{dom} g \cap \operatorname{dom}\omega \neq \emptyset$. Moreover, Ψ is convex because it takes the maximum of convex functions (and g_i is convex). From Assumption 3.1, $\{x \in \operatorname{dom} g \mid \omega(x) \leq \alpha\}$ is compact for all $\alpha \in \mathbf{R}$, so this problem is solvable. This means that $p_{L_k}(x^k)$ is well defined. Since $\operatorname{dom}\Psi$ is nonempty and convex, it follows that there exists x_0 in the relative interior of $\operatorname{dom}\Psi$ [5, Theorem 3.17], and consequently, by [5, Theorem 3.18], $\partial\Psi(x_0) \neq \emptyset$. Furthermore, if ω is strongly convex, we conclude that φ is a proper closed and strongly convex function, and hence, from [5, Lemma 9.7], the subproblem has a unique optimal solution in $\operatorname{dom} g \cap \operatorname{dom}\partial\omega$. □

From Assumption 3.1 (in particular, the relative smoothness of f_i) and the definition of L in (3.1), if $L_k \geq L$, for all i we have

$$F_i(x^{k+1}) - F_i(x^k) \leq \nabla f_i(x^k)^\top (x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k) + L_k B_\omega(x^{k+1}, x^k). \quad (3.6)$$

Since x^{k+1} is the optimal solution of (3.4), the maximum in i of the right-hand side of (3.6) is less than or equal to zero. Thus, for all i

$$F_i(x^{k+1}) \leq F_i(x^k) \quad \text{for all } k, \quad (3.7)$$

that is, the objective functions decrease monotonically. In the following subsections, we consider two stepsize rules for our method.

3.1 Constant stepsize

Now consider the constant stepsize, which we set as $L_k = \bar{L}$ for all k, with $\bar{L} > L$ and L is given in (3.1). Then the proximal gradient method with Bregman distance is given below.

Algorithm 1 Multi-objective proximal gradient method with Bregman distance and constant stepsize

Step 1 Choose $L_k := \bar{L}$ with $\bar{L} > L$, $\varepsilon > 0$, $x^0 \in \operatorname{dom} g \cap \operatorname{dom} \partial\omega$, and set $k := 0$.
Step 2 Compute $p_{L_k}(x^k)$ by solving subproblem (3.4).
Step 3 If $\left\|p_{L_k}(x^k) - x^k\right\| < \varepsilon$, then stop.
Step 4 Set $x^{k+1} := p_{L_k}(x^k)$, $k := k + 1$, and go to Step 2 .

3.2 Backtracking procedure

Now we consider the backtracking procedure. In the beginning, let $L_{-1} = s$ with $s > 0$. At iteration $k \geq 0$, let $L_k = L_{k-1}$. Then, while existing i such that

$$f_i\big(p_{L_k}(x^k)\big) > f_i(x^k) + \left\langle \nabla f_i(x^k), p_{L_k}(x^k) - x^k \right\rangle + L_k B_\omega(p_{L_k}(x^k), x^k),$$

we set $L_k := \eta L_k$ where $\eta > 1$. In other words, $L_k = L_{k-1}\eta^{j_k}$, where j_k is the smallest nonnegative integer given as follows:

$$j_k := \underset{j \in \{0,1,2,\dots\}}{\operatorname{argmin}} \Bigg(F\big(p_{L_{k-1}\eta^{j_k}}(x^k)\big) \preceq F(x^k) + JF(x^k)^\top \big(p_{L_{k-1}\eta^j}(x^k) - x^k\big) + L_{k-1}\eta^j (p_{L_{k-1}\eta^j}(x^k), x^k) \Bigg). \tag{3.8}$$

The rule above ensures that (3.6) is still satisfied at each iteration. In addition, the L_k that the backtracking procedure produces satisfies the following bounds for all k:

$$s \leq L_k \leq \max\{\eta L, s\}.$$

The inequality $s \leq L_k$ is trivial. To prove the inequality $L_k \leq \max\{\eta L, s\}$, we note that either $L_k = s$ or $L_k > s$. In the latter case there exists an index $0 \leq k' \leq k$ and some i for which the inequality (3.6) is not satisfied with $k = k'$ and replacing L_k with L_k/η. From the relative smoothness of f_i, this implies in particular that $L_k/\eta < L^i \leq L$. Thus, we have shown that $L_k \leq \max\{\eta L, s\}$. Namely, $L_k \leq \alpha L$, where $\alpha = \max\{\eta, s/L\}$. We also note that the bounds on L_k can be rewritten as

$$\beta L \leq L_k \leq \alpha L,$$

where

$$\alpha = \begin{cases} \frac{\bar{L}}{L}, & \text{constant,} \\ \max\left\{\eta, \frac{s}{L}\right\}, & \text{backtracking,} \end{cases} \qquad \beta = \begin{cases} \frac{\bar{L}}{L}, & \text{constant} \\ \frac{s}{L}, & \text{backtracking.} \end{cases}$$

So the algorithm with backtracking stepsize is given below.

Algorithm 2 Multi-objective proximal gradient method with Bregman distance and backtracking procedure

Step 1 Choose $s > 0$, $\eta > 1$, $\varepsilon > 0$, $x^0 \in \operatorname{dom} g \cap \operatorname{dom} \partial\omega$, and set $L_{-1} = s$, $k := 0$.
Step 2 Compute L_k by solving (3.8).
Step 3 Compute $p_{L_k}(x^k)$ by solving subproblem (3.4).
Step 4 If $\left\|p_{L_k}(x^k) - x^k\right\| < \varepsilon$, then stop.
Step 5 Set $x^{k+1} := p_{L_k}(x^k)$, $k := k + 1$, and go to Step 2 .

4 Convergence Analysis

In this section, we prove that the sequences generated by Algorithms 3.1 and 3.2 converge to Pareto stationary points and discuss their rate of convergence. From now on, let us assume that an infinite sequence is generated.

4.1 Convergence to Pareto stationary points

Now we analyze the convergence of the proximal gradient method with Bregman distance.

Lemma 4.1. *Let $\{x^k\}$ be generated by Algorithms* 3.1 *or* 3.2 *and suppose that $\{F_i(x^k)\}$ is bounded from below for all $i = 1, \ldots, m$. Then we have*

$$\lim_{k\to\infty} B_\omega(x^k, x^{k+1}) = 0. \tag{4.1}$$

Proof. At the kth iteration,

$$\begin{aligned}
&f_i(x^{k+1}) + g_i(x^{k+1}) \\
&\quad = f_i(x^k) + g_i(x^k) + f_i(x^{k+1}) - f_i(x^k) + g_i(x^{k+1}) - g_i(x^k) \\
&\quad \le f_i(x^k) + g_i(x^k) + \nabla f_i(x^k)(x^{k+1} - x^k) + L_k B_\omega(x^{k+1}, x^k) + g_i(x^{k+1}) - g_i(x^k) \\
&\quad \le f_i(x^k) + g_i(x^k) + \psi_{x^k}(x^{k+1}) + L_k B_\omega(x^{k+1}, x^k) \\
&\quad \le f_i(x^k) + g_i(x^k) + \psi_{x^k}(x) + L_k(B_\omega(x, x^k) - B_\omega(x, x^{k+1}))
\end{aligned}$$

for all x. Here, the first inequality follows from (3.7). The second inequality follows from the definition of $\psi_{x^k}(x^{k+1})$. And the third inequality follows from Lemma 2.6 with $\theta = \psi_{x^k}/L_k$. Letting $x = x^k$, and recalling (2.2) and (3.3), we obtain

$$\begin{aligned}
f_i(x^{k+1}) + g_i(x^{k+1}) &\le f_i(x^k) + g_i(x^k) - L_k B_\omega(x^k, x^{k+1}) \\
&\le f_i(x^k) + g_i(x^k) - \beta L B_\omega(x^k, x^{k+1}).
\end{aligned}$$

Since $\{F_i(x^k)\}$ is bounded from below from the assumption, there exists $\tilde{F}_i \le F_i(x^k) = f_i(x^k) + g_i(x^k)$ for all i and k. Adding up the above inequality from $k = 0$ to $k = \hat{k}$, we obtain

$$f_i(x^{\hat{k}+1}) + g_i(x^{\hat{k}+1}) \le f_i(x^0) + g_i(x^0) - \beta L \sum_{k=0}^{\hat{k}} B_\omega(x^k, x^{k+1}).$$

Because $\beta L > 0$, we have

$$\sum_{k=0}^{\hat{k}} B_\omega(x^k, x^{k+1}) \le (\beta L)^{-1}\big(f_i(x^0) + g_i(x^0) - f_i(x^{\hat{k}+1}) - g_i(x^{\hat{k}+1})\big),$$

and thus

$$\sum_{k=0}^{\hat{k}} B_\omega(x^k, x^{k+1}) < \infty.$$

It then follows from the nonnegativity of $B_\omega(x^k, x^{k+1})$ for all k that

$$\lim_{k\to\infty} B_\omega(x^k, x^{k+1}) = 0,$$

which completes the proof. □

In order to give the convergence analysis, we consider the following assumption. In Remark 4.3 and Section 4.1.1 we will see that this assumption is actually not so strict.

Assumption 4.1. Let $\{x^k\}$ be a sequence generated by Algorithms 3.1 or 3.2. If

$$\lim_{k\to\infty} B_\omega(x^k, x^{k+1}) = 0,$$

then

$$z^* := \lim_{k\to\infty} \{\nabla\omega(x^{k+1}) - \nabla\omega(x^k)\} \in N_{\mathrm{cl}(\mathrm{dom}\,\partial\omega)}(x^*),$$

where x^* is an accumulation point of $\{x^k\}$.

Theorem 4.2. *If Assumption 4.1 holds, then every accumulation point of the sequence $\{x^k\}$ generated by Algorithms 3.1 or 3.2, if it exists, is a Pareto stationary point.*

Proof. From the optimality condition of the subproblem (3.4), we have

$$\sum_{i=1}^m \lambda_i^k(\nabla f_i(x^k) + \eta_i(x^{k+1})) + L_k(\nabla\omega(x^{k+1}) - \nabla\omega(x^k)) = 0,$$

where $\eta_i(x^{k+1}) \in \partial g_i(x^{k+1})$, $\sum_{i=1}^m \lambda_i^k = 1$, $\lambda_i^k \geq 0$ for all $i = 1, \ldots, m$, and $\lambda_i^k = 0$ when $i \notin \mathcal{I}_{x^k}(x^{k+1})$ and

$$\begin{aligned}&\mathcal{I}_{x^k}(x^{k+1})\\ &\quad := \left\{i \in \{1,\ldots,m\} \mid \psi_{x^k}(x^{k+1}) = \nabla f_i(x^k)^\top(x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k)\right\}.\end{aligned}$$

Because $\{\lambda_i^k\}$ and $\{L_k\}$ are bounded, we assume without loss of generality that there exist λ_i^* with $\sum_{i=1}^m \lambda_i^* = 1$, $\lambda_i^* \geq 0$ for all $i = 1, \ldots, m$, and $L^* > 0$ such that

$$\begin{aligned}&\lim_{k\to\infty} \sum_{i=1}^m \lambda_i^k(\nabla f_i(x^k) + \eta_i(x^{k+1})) + L_k(\nabla\omega(x^{k+1}) - \nabla\omega(x^k))\\ &= \sum_{i=1}^m \lambda_i^*(\nabla f_i(x^*) + \eta_i(x^*)) + L^* z^*\\ &= 0,\end{aligned}$$

where $\eta_i(x^*) \in \partial g_i(x^*)$ and z^* is given in Assumption 4.1. Then, from this assumption and Lemma 4.1, we have

$$-\sum_{i=1}^m \lambda_i^*(\nabla f_i(x^*) + \eta_i(x^*)) = L^* z^* \in N_{\mathrm{cl}(\mathrm{dom}\,\partial\omega)}(x^*),$$

where we use the fact that $L^* > 0$ and $N_{\mathrm{cl}(\mathrm{dom}\,\partial\omega)}(x^*)$ is a cone. Therefore,

$$\sum_{i=1}^{m} \langle \lambda_i^* (\nabla f_i(x^*) + \eta_i(x^*)), y - x^* \rangle \geq 0 \quad \text{for all} \quad y \in \mathrm{cl}(\mathrm{dom}\,\partial\omega).$$

This implies that for all $y \in \mathrm{cl}(\mathrm{dom}\,\partial\omega)$, there exists at least one i with $\lambda_i^* > 0$ such that

$$\langle \nabla f_i(x^*) + \eta_i(x^*), y - x^* \rangle \geq 0. \tag{4.2}$$

Moreover, since g_i is convex, we have $g_i'(x^*; y - x^*) = \sup_{v \in \partial g_i(x^*)} \langle v, y - x^* \rangle$ and thus

$$g_i'(x^*; y - x^*) \geq \langle \eta_i(x^*), y - x^* \rangle.$$

Therefore, from (4.2), for all $y \in \mathrm{cl}(\mathrm{dom}\,\partial\omega)$ we obtain

$$F_i'(x^*; y - x^*) = \langle \nabla f_i(x^*), y - x^* \rangle + g_i'(x^*; y - x^*) \geq 0$$

for some i, which implies that $\max F_i'(x^*; y - x^*) \geq 0$, and thus x^* is Pareto stationary. □

Remark 4.3. We observe that if $x^* \in \mathrm{int}(\mathrm{dom}\,\partial\omega)$ or if $\nabla\omega$ is Hölder continuous, then $z^* = 0$. This ensures the fulfillment of Assumption 4.1, implying that the accumulation point x^* is Pareto stationary. This aligns with the result presented in [17, Theorem 2(iii)], where $\nabla\omega$ is assumed to be Lipschitz continuous.

4.1.1 Special case

We now show that Assumption 4.1 is not so strict, by showing an example with Kullback-Leibler divergence. For all $i = 1, \ldots, m$, define $g_i(x)$ as the indicator function of $\Delta_n := \{x \in \mathbf{R}^n \mid \sum_{i=1}^n x_i = 1, x \geq 0\}$. Let $\omega(x) = \sum_{i=1}^n x_i \ln x_i$ with constraint condition $x \succeq 0$ for the problem (1.1). Note that $\mathrm{cl}(\mathrm{dom}\,\partial\omega) = \{x \mid x \succeq 0\}$. Also, we recall that in this case, $\nabla\omega$ is not Hölder or Lipschitz continuous.

Suppose that the algorithm generates a sequence $\{x^k\}$. It is well-known that ω is 1-strongly convex over Δ_n, and in this case the following inequality holds:

$$B_\omega(x^{k+1}, x^k) \geq \frac{1}{2} \|x^{k+1} - x^k\|_1,$$

where $\|\cdot\|_1$ denotes the 1-norm. Then, combined with Lemma 4.1, we can conclude that $\lim_{k\to\infty} \|x^{k+1} - x^k\|_1 = 0$, which means, for all i, that

$$\lim_{k\to\infty} |x_i^{k+1} - x_i^k| = 0. \tag{4.3}$$

Now, recall that the optimality condition of the subproblem (3.4) is given by

$$\sum_{i=1}^{m} \lambda_i^k \big(\nabla f_i(x^k) + \eta_i(x^{k+1})\big) + L_k(\nabla\omega(x^{k+1}) - \nabla\omega(x^k)) = 0,$$

where $\sum_{i=1}^m \lambda_i^k = 1$, $\lambda_i^k \geq 0$ for all $i = 1, \ldots, m$, and $\eta_i(x^{k+1}) \in \partial g_i(x^{k+1})$. Note also that for any index j,

$$(\nabla\omega(x^{k+1}) - \nabla\omega(x^k))_j = \ln\left(\frac{x_j^{k+1}}{x_j^k}\right).$$

Let us now consider an accumulation point x^* of $\{x^k\}$. For any index j such that $x_j^* = 0$, if $\sum_{i=1}^m \lambda_i^k(\nabla f_i(x^k) + \eta_i(x^{k+1}))_j < 0$, it follows that $\ln(x_j^{k+1}/x_j^k) > 0$ because $L_k > 0$. This shows that $x_j^{k+1} > x_j^k$, which contradicts the fact that $x_j^* = 0$. Consequently, it must be the case that $\lim_{k\to\infty} \sum_{i=1}^m \lambda_i^k(\nabla f_i(x^k) + \eta_i(x^{k+1}))_j \geq 0$, resulting in $\lim_{k\to\infty} \ln(x_j^{k+1}/x_j^k) \leq 0$. On the other hand, for an index j such that $x_j^* > 0$, the condition (4.3) implies $\lim_{k\to\infty} \ln(x_j^{k+1}/x_j^k) = 0$.

Furthermore, assuming that $\{\nabla f_i(x^k)\}$ and $\{\eta_i(x^{k+1})\}$ are bounded for all i, then $\{\nabla\omega(x^{k+1}) - \nabla\omega(x^k)\}$ is also bounded. If z^* is its accumulation point, then $z_j^* \leq 0$ for all $x_j^* = 0$ and $z_j^* = 0$ for all $x_j^* > 0$, which means $z^* \in N_{\{x|x\geq 0\}}(x^*)$, and thus the Assumption 4.1 is satisfied in this case.

Remark 4.4. We observe that the merit function $v_\ell(x)$ as defined in [17] is not well defined on $\mathrm{bd}(\mathrm{dom}\,\partial\omega)$, especially when $\omega(x) = \sum_{i=1}^n x_i \ln x_i$. Similarly, due to this reason, the discourse on the nonconvex case in [18] is questionable.

4.2 Convergence rate analysis

Let us now recall a merit function for the multi-objective optimization problem, and use it to estimate the convergence rate. Here, we will discuss only the convex and the strongly convex cases. The merit function is the simple function $u_0 \colon \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ defined as follows [37]:

$$u_0(x) := \sup_{y\in\mathbf{R}^n} \min_{i\in\{1,\ldots,m\}} \{F_i(x) - F_i(y)\}. \tag{4.4}$$

4.2.1 The convex case

Here we use the function $u_0(\cdot)$ to analyze the convergence rate. First, we give the following lemma. Note that we state it with f_i and g_i having general convexity parameters, which turn out to be zero in this subsection. It will not make a difference here, but it will be important in the discussion of the next subsection.

Lemma 4.5. *Assume that f_i is μ_i-strongly convex relative to ω and g_i is ν_i-strongly convex relative to ω, and write $\mu := \min_{i\in\{1,\ldots,m\}} \mu_i$ and $\nu := \min_{i\in\{1,\ldots,m\}} \nu_i$. Then, for all $x \in \mathrm{dom}\, g$ it follows that*

$$\sum_{i=1}^m \beta_i^k(F_i(x^{k+1}) - F_i(x)) \leq L_k(B_\omega(x, x^k) - B_\omega(x, x^{k+1})) - \mu B_\omega(x, x^k) - \nu B_\omega(x, x^{k+1}),$$

where β_i^k satisfies the following conditions:

(i) *There exists $\eta_i(x^{k+1}) \in \partial g_i(x^{k+1})$ such that*

$$\sum_{i=1}^m \beta_i^k(\nabla f_i(x^k) + \eta_i(x^{k+1})) + L_k(\nabla\omega(x^{k+1}) - \nabla\omega(x^k)) = 0,$$

(ii) *$\sum_{i=1}^m \beta_i^k = 1$, $\beta_i^k \geq 0$ $(i \in \mathcal{I}_{x^k}(x^{k+1}))$ and $\beta_i^k = 0$ $(i \notin \mathcal{I}_{x^k}(x^{k+1}))$, where $\mathcal{I}_{x^k}(x^{k+1}) := \left\{i \mid \psi_{x^k}(x^{k+1}) = \nabla f_i(x^k)^\top(x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k)\right\}$.*

Proof. For all i, (3.6) holds for both Algorithms 3.1 and 3.2, that is

$$F_i(x^{k+1}) - F_i(x^k) \leq \nabla f_i(x^k)^\top (x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k) + L_k B_\omega(x^{k+1}, x^k). \quad (4.5)$$

The above inequality and relative strong convexity of f_i with modulus μ_i give

$$\begin{aligned}
&F_i(x^{k+1}) - F_i(x) \\
&\quad = (F_i(x^k) - F_i(x)) + (F_i(x^{k+1}) - F_i(x^k)) \\
&\quad \leq (\nabla f_i(x^k)^\top (x^k - x) - \mu_i B_\omega(x, x^k) + g_i(x^k) - g_i(x)) \\
&\qquad + (\nabla f_i(x^k)^\top (x^{k+1} - x^k) + g_i(x^{k+1}) - g_i(x^k) + L_k B_\omega(x^{k+1}, x^k)) \\
&\quad \leq \nabla f_i(x^k)^\top (x^{k+1} - x) + g_i(x^{k+1}) - g_i(x) - \mu B_\omega(x, x^k) + L_k B_\omega(x^{k+1}, x^k) \\
&\quad \leq (\nabla f_i(x^k) + \eta_i(x^{k+1}))^\top (x^{k+1} - x) - \mu B_\omega(x, x^k) - \nu B_\omega(x, x^{k+1}) + L_k B_\omega(x^{k+1}, x^k)
\end{aligned}$$

for all $x \in \operatorname{dom} g$, where the second inequality follows from the definition of μ and the last one comes from the relative strong convexity of g_i. Multiplying the above inequality by β_i^k and summing for all $i \in \{1, \ldots, m\}$, the conditions (i) and (ii) give

$$\begin{aligned}
&\sum_{i=1}^m \beta_i^k (F_i(x^{k+1}) - F_i(x)) \\
&\leq -L_k (\nabla \omega(x^{k+1}) - \nabla \omega(x^k))^\top (x^{k+1} - x) + L_k B_\omega(x^{k+1}, x^k) \\
&\quad - \mu B_\omega(x, x^k) - \nu B_\omega(x, x^{k+1}) \\
&= L_k \Big(\nabla \omega(x^k)^\top x^{k+1} - \nabla \omega(x^k)^\top x + \nabla \omega(x^{k+1})^\top (x - x^{k+1}) + \omega(x^{k+1}) - \omega(x^k) \\
&\quad - \nabla \omega(x^k)^\top (x^{k+1} - x^k) \Big) - \mu B_\omega(x, x^k) - \nu B_\omega(x, x^{k+1}) \\
&= L_k \Big(\omega(x) - \omega(x^k) - \nabla \omega(x^k)^\top (x - x^k) - \omega(x) + \omega(x^{k+1}) + \nabla \omega(x^{k+1})^\top (x - x^{k+1}) \Big) \\
&\quad - \mu B_\omega(x, x^k) - \nu B_\omega(x, x^{k+1}) \\
&= L_k (B_\omega(x - x^k) - B_\omega(x - x^{k+1})) - \mu B_\omega(x, x^k) - \nu B_\omega(x, x^{k+1}),
\end{aligned}$$

where the second and last equalities follow from the definition of Bregman distance. □

We now prove the convergence rate assuming convexity of the objective functions. The following assumption is considered, which is equivalent to the existence of at least one optimal point in the single-objective case. As suggested in [36, Remark 5.2], it is not particularly strong even in the multi-objective case.

Assumption 4.2. Let X^* be the set of weakly Pareto optimal points for the multi-objective problem, and define the level set of F for $\alpha \in \mathbf{R}^m$ by $\Omega_F(\alpha) := \{x \in \mathbf{R}^n \mid F(x) \preceq \alpha\}$. Then, for all $x \in \Omega_F(F(x^0))$ there exists $x^* \in X^*$ such that $F(x^*) \preceq F(x)$ and

$$R := \sup_{F^* \in F(X^* \cap \Omega_F(F(x^0)))} \inf_{x \in F^{-1}(\{F^*\})} B_\omega(x, x^0) < \infty.$$

Theorem 4.6. *Assume that F_i is convex for all $i \in \{1, \ldots, m\}$. Under Assumption 4.2, Algorithm 3.1 (or Algorithm 3.2) generates a sequence $\{x^k\}$ such that*

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

where $\alpha = \frac{\bar{L}}{L}$ in the constant stepsize setting and $\alpha = \max\{\eta, \frac{s}{L}\}$ if the backtracking rule is employed.

Proof. From Lemma 4.5 and the convexity of f_i and g_i, for all $x \in \mathbf{R}^n$ we have

$$\sum_{i=1}^{m} \beta_i^k (F_i(x^{k+1}) - F_i(x)) \leq L_k (B_\omega(x - x^k) - B_\omega(x - x^{k+1})).$$

Adding up the above inequality from $k = 0$ to $k = \hat{k}$, we obtain

$$\begin{aligned}\sum_{k=0}^{\hat{k}} \sum_{i=1}^{m} \beta_i^k (F_i(x^{k+1}) - F_i(x)) &\leq L_k (B_\omega(x, x^0) - B_\omega(x, x^{\hat{k}+1})) \\ &\leq L_k B_\omega(x, x^0).\end{aligned}$$

The rest of the proof follows similarly to the proof of [36, Theorem 5.2]. □

4.2.2 The strongly convex case

Here, we show that $\{x^k\}$ generated by Algorithms 3.1 and 3.2 converge linearly to a Pareto optimal point if F_i is strongly convex relative to ω.

Theorem 4.7. *Let $\{x^k\}$ be generated by Algorithm 3.1 or 3.2 and suppose that Assumption 4.1 holds. Let f_i and g_i have convexity parameters $\mu_i \in \mathbf{R}$ and $\nu_i \in \mathbf{R}$, respectively, and write $\mu := \min_{i\in\{1,\ldots,m\}} \mu_i$ and $\nu := \min_{i\in\{1,\ldots,m\}} \nu_i$. Then there exists a Pareto optimal point $x^* \in \mathbf{R}^n$ such that for each iteration k,*

$$(L_k + \nu) B_\omega(x^*, x^{k+1}) \leq (L_k - \mu) B_\omega(x^*, x^k). \tag{4.6}$$

Furthermore, assume that ω is σ-strongly convex with $\sigma > 0$, and $\nabla\omega$ is q-Hölder continuous with parameter c and $0 < q \leq 1$. Then, there exists a Pareto optimal point $x^* \in \mathbf{R}^n$ such that for each iteration k,

$$\left\|x^{k+1} - x^*\right\| \leq \sqrt{\frac{c(\alpha L - \mu)}{\sigma(\beta L + \nu)}} \left\|x^k - x^*\right\|^{\frac{q+1}{2}},$$

where

$$\alpha = \begin{cases} \frac{\bar{L}}{L}, & \text{constant,} \\ \max\left\{\eta, \frac{s}{L}\right\}, & \text{backtracking,} \end{cases} \qquad \beta = \begin{cases} \frac{\bar{L}}{L}, & \text{constant} \\ \frac{s}{L}, & \text{backtracking.} \end{cases}$$

Thus, we have

$$\left\|x^k - x^*\right\| \leq \left(\sqrt{\frac{c(\alpha L - \mu)}{\sigma(\beta L + \nu)}}\right)^{\sum_{i=0}^{k-1} (\frac{q+1}{2})^i} \left\|x^0 - x^*\right\|^{(\frac{q+1}{2})^k},$$

and if $0 < c \leq \frac{\sigma(\beta L+\nu)}{\alpha L-\mu}$ then $0 < \sqrt{\frac{c(\alpha L-\mu)}{\sigma(\beta L+\nu)}} \leq 1$. In particular, if $\nabla\omega$ is Lipschitz continuous, we obtain

$$\left\|x^k - x^*\right\| \leq \left(\sqrt{\frac{c(\alpha L - \mu)}{\sigma(\beta L + \nu)}}\right)^k \left\|x^0 - x^*\right\|.$$

Proof. Since each F_i is strongly convex relative to ω, the level set of every F_i is bounded. Thus, $\{x^k\}$ has an accumulation point $x^* \in \mathbf{R}^n$. Note that x^* is Pareto stationary from

Theorem 4.2, and thus Pareto optimal because of the strong convexity assumption. From Lemma 4.5, we have

$$\sum_{i=1}^{m} \beta_i^k (F_i(x^{k+1}) - F_i(x^*)) \leq L_k(B_\omega(x^*, x^k) - B_\omega(x^*, x^{k+1})) \\ - \mu B_\omega(x^*, x^k) - \nu B_\omega(x^*, x^{k+1}),$$

where β_i^k satisfies the conditions (i) and (ii) of Lemma 4.5. Since the left-hand side of the above inequality is nonnegative because of (3.7), we obtain

$$0 \leq L_k(B_\omega(x^*, x^k) - B_\omega(x^*, x^{k+1})) - \mu B_\omega(x^*, x^k) - \nu B_\omega(x^*, x^{k+1}).$$

Namely,

$$(L_k + \nu) B_\omega(x^*, x^{k+1}) \leq (L_k - \mu) B_\omega(x^*, x^k). \tag{4.7}$$

Similiar to the so-called descent lemma [9, Proposition A.24], using the Hölder continuity of $\nabla\omega$, i.e., $\|\nabla\omega(y) - \nabla\omega(x)\| \leq c\|y - x\|^q$, we obtain for all x, y,

$$\omega(x) \leq \omega(y) + \nabla\omega(y)^\top (x - y) + \frac{c}{2}\|y - x\|^{q+1}.$$

Combined with the σ-strong convexity of ω as well as the definition of Bregman distance, we have

$$\frac{\sigma}{2}\|x - y\|^2 \leq B_\omega(x, y) \leq \frac{c}{2}\|x - y\|^{q+1}.$$

The above inequality and (4.7) give

$$\frac{\sigma(L_k + \nu)}{2}\left\|x^* - x^{k+1}\right\|^2 \leq \frac{c(L_k - \mu)}{2}\left\|x^* - x^k\right\|^{q+1}, \tag{4.8}$$

which is equivalent to

$$\left\|x^{k+1} - x^*\right\| \leq \sqrt{\frac{c(L_k - \mu)}{\sigma(L_k + \nu)}}\left\|x^k - x^*\right\|^{\frac{q+1}{2}}.$$

Using the bounds for L_k, we obtain

$$\left\|x^{k+1} - x^*\right\| \leq \sqrt{\frac{c(\alpha L - \mu)}{\sigma(\beta L + \nu)}}\left\|x^k - x^*\right\|^{\frac{q+1}{2}},$$

where $\alpha = \frac{\bar{L}}{L}$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L}\right\}$ if the backtracking rule is employed. The result then follows by applying this inequality k times. □

Note that under strong convexity assumption, from (4.6) we have

$$B_\omega(x^*, x^k) \leq \left(\frac{\alpha L - \mu}{\beta L + \nu}\right)^k B_\omega(x^*, x^0).$$

Moreover, we also have linear convergence (as usual, defined with Euclidean norm) when Lipschitz continuity of $\nabla\omega$ is satisfied. One future work will be to see if this condition can be removed to establish the same rate of convergence in the strongly convex case.

5 Conclusion

We proposed a proximal gradient method with Bregman distance for multi-objective optimization problems. We also used two-step size strategies: the constant stepsize and the backtracking strategy. We prove that the sequence generated by the algorithms can converge to a Pareto stationary point and further analyze its convergence rate through a merit function. Finally, we proved the convergence rates for convex ($O(1/k)$), and strongly convex ($O(r^k)$) for some $r \in (0,1)$) problems. We point out that our proof is based on Assumption 4.1, which is less strict than assuming Hölder continuity of the gradient of the reference function. For future research, it would be pertinent to investigate under which conditions Assumption 4.1 holds true.

References

[1] M.A.T. Ansary and J. Dutta, A proximal gradient method for multi-objective optimization problems using Bregman functions, *Optimization Online*, (2022).

[2] P.B. Assunção, O.P. Ferreira and L.F. Prudente, Conditional gradient method for multiobjective optimization, *Comput. Optim. Appl.* 78 (2021) 741–768.

[3] P.B. Assunção, O.P. Ferreira and L.F. Prudente, A generalized conditional gradient method for multiobjective composite optimization problems. *Optimization* (2023) 1–31 https://doi.org/10.1080/02331934.2023.2257709.

[4] H.H. Bauschke, J. Bolte, and M. Teboulle, A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications, *Math. Oper. Res.* 42 (2017) 330–348.

[5] A. Beck, *First-Order Methods in Optimization*, SIAM, 2017.

[6] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Oper. Res. Lett.* 31 (2003) 167–175.

[7] Y. Bello-Cruz, J.G. Melo and R.V. Serra, A proximal gradient splitting method for solving convex vector optimization problems, *Optimization* 71 (2022) 33–53.

[8] G.C. Bento, J.X. Cruz Neto, G. López, A. Soubeyran and J.C.O. Souza, The proximal point method for locally Lipschitz functions in multiobjective optimization with application to the compromise problem, *SIAM J. Optim.* 28 (2018) 1104–1120.

[9] D.P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

[10] M. Binder, J. Moosbauer, J. Thomas and B. Bischl, Multi-objective hyperparameter tuning and feature selection using filter ensembles, in: *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, 2020, pp. 471–479.

[11] J. Bolte, S. Sabach, M. Teboulle and Y. Vaisbourd, First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems, *SIAM J. Optim.* 28 (2018) 2131–2151.

[12] H. Bonnel, A.N. Iusem and B.F. Svaiter, Proximal methods in vector optimization, *SIAM J. Optim.* 15 (2005) 953–970.

[13] R. Boţ and S.-M. Grad, Inertial forward–backward methods for solving vector optimization problems, *Optimization* 67 (2018) 1–16.

[14] L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, *USSR Comput. Math. & Math. Phys.* 7 (1967) 200–217.

[15] Y. Censor and A. Lent, An iterative row-action method for interval convex programming, *J. Optim. Theory Appl.* 34 (1981) 321–353.

[16] G. Chen and M. Teboulle, Convergence analysis of a proximal-like minimization algorithm using Bregman functions, *SIAM J. Optim.* 3 (1993) 538–543.

[17] J. Chen, L. Tang and X. Yang, Convergence rates analysis of Interior Bregman Gradient Method for Vector Optimization Problems, *arXiv preprint arXiv:2206.10070*, (2022).

[18] K. Chen, E.H. Fukuda and N. Yamashita, A proximal gradient method with Bregman distance in multi-objective optimization, Master thesis, Kyoto University, 2022. Available at: `https://www.amp.i.kyoto-u.ac.jp/tecrep/ps_file/2022/2022-002.pdf`.

[19] Z. Chen, X. Huang and X. Yang, Generalized proximal point algorithms for multiobjective optimization problems, *Appl. Anal.* 90 (2011) 935–949.

[20] J.X. Cruz Neto, G.J.P. Silva, O.P. Ferreira and J.O. Lopes, A subgradient method for multiobjective optimization, *Comput. Optim. Appl.* 54 (2013) 461–472.

[21] J. Fliege, L.M. Graña Drummond and B.F. Svaiter, Newton's method for multiobjective optimization, *SIAM J. Optim.* 20 (2009) 602–626.

[22] J. Fliege and B.F. Svaiter, Steepest descent methods for multicriteria optimization, *Math. Methods Oper. Res.* 51 (2000) 479–494.

[23] E.H. Fukuda and L.M. Graña Drummond, On the convergence of the projected gradient method for vector optimization, *Optimization* 60 (2011) 1009–1021.

[24] E.H. Fukuda and L.M. Graña Drummond, Inexact projected gradient method for vector optimization, *Comput. Optim. Appl.* 54 (2013) 473–493.

[25] E.H. Fukuda and L.M. Graña Drummond, A survey on multiobjective descent methods, *Pesq. Oper.* 34 (2014) 585–620.

[26] A.M. Geoffrion, Proper efficiency and the theory of vector maximization, *J. Math. Anal. Appl.* 22 (1968) 618–630.

[27] M. Gong, M. Zhang and Y. Yuan, Unsupervised band selection based on evolutionary multiobjective optimization for hyperspectral images, *IEEE Trans. Geosci. Remote Sens.* 54 (2016) 544–557.

[28] S.-M. Grad, A survey on proximal point type algorithms for solving vector optimization problems, in: *Splitting Algorithms, Modern Operator Theory, and Applications*, H.H. Bauschke, R.S. Burachik, and D.R. Luke (eds.), Springer International Publishing, 2019, pp. 269–308.

[29] L.M. Gra 単 a Drummond and A.N. Iusem, A projected gradient method for vector optimization problems, *Comput. Optim. Appl.* 28 (2004) 5–29.

[30] F. Huang, J. Li, S. Gao and H. Huang, Enhanced bilevel optimization via Bregman distance, in: *Adv. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (eds.), vol.35, 2022, pp. 28928–28939.

[31] H. Lu, R.M. Freund and Y. Nesterov, Relatively smooth convex optimization by first-order methods, and applications, *SIAM J. Optim.* 28 (2018) 333–354.

[32] Y. Nishimura, E.H. Fukuda and N. Yamashita, Monotonicity for multiobjective accelerated proximal gradient methods, *J. Oper. Res. Soc. Jpn.* 67 (2024) 1–17.

[33] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[34] J.C.O. Souza, J.X. Cruz Neto, J.O. Lopes, R.C.M. Silva, and S.D.B. Bitar, The proximal point method with a vectorial Bregman regularization in multiobjective DC programming, *Optimization* 71 (2022) 263–284.

[35] H. Tanabe, E.H. Fukuda and N. Yamashita, Proximal gradient methods for multiobjective optimization and their applications, *Comput. Optim. Appl.* 72 (2019) 339–361.

[36] H. Tanabe, E.H. Fukuda and N. Yamashita, Convergence rates analysis of a multiobjective proximal gradient method, *Optim. Lett.* 17 (2023) 333–350.

[37] H. Tanabe, E.H. Fukuda and N. Yamashita, New merit functions for multiobjective optimization and their properties, *Optimization* 1–38, https://doi.org/10.1080/02331934.2023.2232794.

[38] M. Teboulle, A simplified view of first order methods for optimization, *Math. Program.* 170 (2018) 67–96.

[39] H. Wang and A. Banerjee, Bregman alternating direction method of multipliers, in: *Adv. Neural Inf. Process. Syst.*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), vol. 27, 2014.

Manuscript received 31 October 2023
revised 5 February 2024
accepted for publication 22 February 2024

Kangming Chen
Graduate School of Informatics, Kyoto University
Kyoto, 606–8501, Japan
E-mail address: kangming@amp.i.kyoto-u.ac.jp

Ellen H. Fukuda
Graduate School of Informatics, Kyoto University
Kyoto, 606–8501, Japan
E-mail address: ellen@i.kyoto-u.ac.jp

Nobuo Yamashita
Graduate School of Informatics, Kyoto University
Kyoto, 606–8501, Japan
E-mail address: nobuo@i.kyoto-u.ac.jp