Stochastic block coordinate and function alternation for multi-objective optimization and learning

Trang H. Tran* L. N. Vicente†

October 24, 2026

Multi-objective optimization is central to many engineering and machine learning applications in which several objectives must be balanced. Standard multi-gradient methods typically require access to every objective gradient at each iteration, which is unsuitable when the objectives arise from separate data sources, models, or computational systems and only one objective can be accessed or optimized at a time. Moreover, when the decision variable is high dimensional or naturally partitioned, computing and updating a full gradient may be prohibitively expensive, whereas partial gradients with respect to individual variable blocks may be substantially cheaper or may be the only information available. We therefore propose Function-Block-Alternate, a stochastic framework that combines these two forms of alternation by performing each elementary step using one objective function and one variable block. The function and block steps can be selected in a flexible order, including deterministic or randomized. The frequency of the function steps determines the target weighted scalarization and enables the exploration of the Pareto front. The framework includes function-alternating, block-coordinate, and stochastic-gradient methods as special cases. We establish convergence guarantees in smooth nonconvex, convex, and Polyak-Łojasiewicz settings, recovering classical single-objective rates without assuming bounded iterates or uniformly bounded stochastic gradients. Numerical experiments evaluate the computed Pareto fronts under equal objective-block computation costs, including settings in which only one objective or one objective block partial gradient can be accessed at a time.

1 Introduction

Multi-objective optimization (MOO) arises in numerous real-world applications from engineering to finance where several conflicting criteria must be evaluated simultaneously [14, 22, 31, 32]. Because these objectives potentially compete, the goal is to find a set of Pareto optimal solutions where no single objective can be improved without degrading another. Identifying this set of efficient points enables decision-makers to navigate complex trade-offs effectively. Furthermore, in many modern large-scale applications, these competing objectives are inherently stochastic, and the underlying decision variables exhibit a natural block structure [33, 38, 40]. In this paper,

*Department of Mathematics, Emory University, Atlanta, GA 30322-1005, USA (htran31@emory.edu). This work was done when the author was a postdoctoral researcher at Lehigh University.

†Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA (lnv@lehigh.edu).

we consider the following problem:

$$\min_{x\in\mathbb{R}^n} F(x) = (f_1(x), \ldots, f_q(x)) = (\mathbb{E}[g_1(x,\xi)], \ldots, \mathbb{E}[g_q(x,\xi)]),$$

where the decision variable $x \in \mathbb{R}^n$ is partitioned into s disjoint blocks $x = (x_1, \ldots, x_s)$, and each block x_i has dimension n_i with $\sum_{i=1}^{s} n_i = n$. We consider the MOO setting where each individual objective function f_k has a Lipschitz continuous gradient. Moreover, each objective function is stochastic and computing precise gradients is typically either impossible or requires prohibitive computational resources. To overcome this limitation, our framework assumes access to unbiased stochastic estimators of the gradients of each objective f_k, or of their block components, depending on the access regime, where ξ represents a random variable.

1.1 Literature review and motivation

Existing MOO literature generally divides methods into two main categories: *a priori* and *a posteriori* methods, depending on when the decision-maker's preferences are incorporated into the optimization process. In MOO, the *a priori* approach integrates decision-making preferences beforehand to transform the multi-objective problem into a simpler single-objective one, which can then be solved using single-objective optimization algorithms. Common strategies in this category include the weighted-sum method [18], which assigns non-negative weights to combine the objectives into a single convex linear combination; the ϵ-constrained method [23], which minimizes one primary objective while imposing upper bounds on the others; and the utility function method [3, 32], which employs a scalar function to mathematically quantify the overall preference between different objective vectors. In contrast, *a posteriori* methods are designed to generate a comprehensive or representative set of Pareto optimal solutions from which the decision-maker later selects the best option. These methods typically update a list of candidate solutions using either metaheuristics (such as evolutionary algorithms [1, 7]) or rigorous descent mechanisms that compute common descent directions for all objectives simultaneously [9, 12, 17]. The alternating optimization algorithms proposed in this paper belong, in a certain sense, to both categories.

In smooth, deterministic MOO, the multi-gradient algorithm applies the steepest descent step by computing the minimal-norm convex combination of gradients from individual functions. Therefore, a natural approach for stochastic MOO is the stochastic multi-gradient algorithm [29, 35], where the gradients in the minimal-norm convex combination are replaced by its stochastic estimators. It is shown that the stochastic multi-gradient is a biased estimator of the true multi-gradient of the weighted function, and one can impose a bound on the amount of bias estimation by dynamic sampling [29]. With that assumption, [29] establishes sublinear convergence rates that are similar to the ones known for single-objective optimization: $\mathcal{O}(1/T)$ in the strongly convex case and $\mathcal{O}(1/T^{1/2})$ in the convex one, where T is the number of iterations. While [29] only considers convex and strongly convex settings, follow-up work [5, 15, 16, 45] investigates multi-objective stochastic gradient methods in non-convex case, using different strategies to address the bias in the stochastic multi-gradient. Without variance reduction, [5] shows the convergence rate of $\mathcal{O}(1/T^{1/2})$ for their proposed stochastic algorithm with double sampling in the non-convex case.

In single-objective optimization, block coordinate descent refers to a family of algorithms that update one block of the decision variable at each iteration [2, 27, 38, 40]. This approach

is attractive when the variable naturally decomposes into blocks of coordinates and when full-gradient steps are expensive. The update of block variables can be done in either a Gauss-Seidel sequential style or a Gauss-Jacobi parallel style [2, 4, 27]. The order of such block variables can be chosen using a cyclic fashion, a random permutation of all blocks, or a randomized selection, where the new block is chosen uniformly at random. The convergence of deterministic BCD has been analyzed extensively in the literature [2, 20, 30, 33, 36, 39], where [2, 33] consider general smooth convex settings for cyclic BCD and its randomized version, and achieve a convergence rate of $\mathcal{O}(1/T)$. The work [41] pioneered the generalization of BCD for stochastic, single-objective optimization with theoretical results for both convex and non-convex regimes. The update order of the variable blocks can be fixed or shuffled, and the algorithm achieves sublinear convergence rates for convex and strongly convex cases (namely $\mathcal{O}(1/T^{1/2})$ and $\mathcal{O}(1/T)$, respectively), which match the rates of stochastic gradient descent. However, [41] only shows asymptotic convergence in the non-convex setting. Later work [4, 8, 13, 24, 42, 43] expands this research direction, where [4] provides non-asymptotic convergence guarantees for cyclic BCD methods in non-convex settings, however, utilizing a variance reduction technique.

1.2 Contributions of this paper

In this paper, we develop Function-Block-Alternate, a stochastic block-coordinate and function-alternation method for stochastic MOO. At each elementary update, it evaluates a stochastic gradient of only one objective with respect to only one variable block. The frequency assigned to each objective determines the target scalarization, and the blocks and objective functions may be selected in a flexible order, deterministic or randomized. We establish convergence rates in smooth non-convex, convex, and Polyak-Łojasiewicz settings, matching the corresponding classical stochastic-gradient rates while retaining the reduced cost of an objective-block update.

Function-Block-Alternate differs from existing stochastic MOO literature, where the most closely related work is the stochastic alternating bi-objective algorithm [28]. While [28] alternates only function minimization and treats two objectives in convex and strongly convex settings, our method simultaneously alternates objectives and variable blocks and accommodates any number of objectives in convex, non-convex, and PL settings. Function-Alternate is obtained when the decision variable consists of a single block, whereas Function-Block-Alternate permits both function and block alternation. More importantly, the analysis in [28] uses the Intermediate Value Theorem to represent a cycle of separate objective updates through an implicit convex combination. That argument relies on having exactly two objectives and on performing all steps for the first objective before all steps for the second; it does not naturally extend to an arbitrary ordering of objective steps. Our analysis expands every objective-block gradient around the point at the beginning of a main iteration. At that common point, the cumulative contribution of the updates is bounded independently of their order and is related to the gradient of the weighted scalarization (determined by how frequently each objective is selected). Gradient smoothness is then used to bound the error caused by evaluating the actual gradients at successive intermediate points. This decomposition applies to any ordering of the specified objective/block updates and any number of objectives and does not rely on convexity.

In addition, our paper addresses a critical gap in the block coordinate descent literature for stochastic MOO, as no prior work has analyzed the simultaneous alternation of functions and variable blocks. The closely related BCD work [6] employs scalarization to combine objectives into a single weighted function prior to optimization, rather than treating them separately, but

it lacks a theoretical analysis for its proposed method. Our analysis requires neither bounded iterates nor uniformly bounded stochastic gradients. In the convex case, we prove boundedness of the expected squared distance to a minimizer directly from a recursion coupling this distance with the expected objective-function gap. We impose a relative objective disagreement condition (Assumption 3.3), which is weaker than uniform gradient boundedness: it allows individual objective gradients to grow and controls only their disagreement around the scalarized gradient. It is automatic in the single-objective case and permits substantial conflict among several objectives as long as their dispersion grows at most proportionally to the scalarized gradient.

Finally, our numerical study separates three access regimes. We first compare the four original methods when all objectives and full gradients are available, then restrict access to one objective at a time, and finally consider a real-data problem in which only objective-block gradients are available. We assess the resulting Pareto fronts using purity, hypervolume, and the Γ and Δ spread indicators. When all objectives and full gradients are available, Function-Block-Alternate is competitive with the other methods, especially in purity and spread indicators. When only one objective can be accessed at a time, randomized Function-Block-Alternate generally performs best.

The remainder of this paper is organized as follows. In Section 2, we provide a detailed description of the proposed Function-Block-Alternate algorithm. Section 3 shows the theoretical assumptions and proves the main descent lemma. We then present, in Section 4, the convergence rate analysis of our method for the smooth, non-convex case. In addition, we derive convergence rates for the smooth convex case and the case where the functions satisfy the Polyak-Łojasiewicz (PL) condition. In Section 5, we compare Pareto-front approximations under equal objective-block computation costs across three information-access regimes: access to all objective gradients, access to the full gradient of only one objective at a time, and access to only one objective-block partial gradient at a time. The paper is concluded with some final remarks in Section 6.

2 Stochastic block coordinate and function alternation algorithm

Fix a budget $p \geq q$ and a frequency vector $m = (m_1, \ldots, m_q) \in \mathbb{Z}^q_{\geq 0}$ satisfying $\sum_{k=1}^q m_k = p > 0$. We choose a mapping $\pi_m : \{0, \ldots, p-1\} \to \{1, \ldots, q\}$ that lists the objective k exactly m_k times, i.e., $|\pi_m^{-1}(k)| = m_k$. Define the following weighted-sum function:

$$F_m(x) := \sum_{k=1}^{q} \frac{m_k}{p} f_k(x) = \frac{1}{p} \sum_{j=0}^{p-1} f_{\pi_m(j)}(x). \tag{2.1}$$

As a result, the frequency vector m determines F_m. It also determines a prescribed multiset of block-objective pairs as follows:

$$\mathcal{P}_t = \{(i, \pi_m(j)) : i = 1, \ldots, s,\ j = 0, \ldots, p-1\}.$$

An ordering ρ_t lists every member of $\mathcal{P}_t$ exactly once and is selected before the stochastic samples of main iteration t are generated. For $r = 0, \ldots, sp-1$, we write $(i_{t,r}, k_{t,r}) = \rho_t(r)$. Hence, $\mathcal{P}_t$ can be processed in any implementable order. For example, let $n = 4$ and $q = 3$, and partition the variables into the two blocks B_1 and B_2. Take $m = (1, 2, 3)$, so that $p = 6$, and let $\pi_m = (1, 2, 2, 3, 3, 3)$. Then

$$\rho_t = ((2,1), (1,3), (2,2), (1,2), (2,3), (1,1), (2,3), (1,3), (2,2), (1,3), (2,3), (1,2))$$

is an implementable ordering. Indeed, for each block it contains one occurrence of objective 1, two occurrences of objective 2, and three occurrences of objective 3, while interleaving the blocks and objective functions in an otherwise arbitrary way.

In addition, we let $U_i \in \mathbb{R}^{n \times n_i}$ be the matrix that places a vector from block i in the corresponding coordinates of $\mathbb{R}^n$ and sets all other coordinates to zero. In particular, $\|U_i v\| = \|v\|$ for every $v \in \mathbb{R}^{n_i}$. Then by the construction of the ordering $(i_{t,r}, k_{t,r})$, we have

$$\sum_{r=0}^{sp-1} U_{i_{t,r}} \nabla_{i_{t,r}} f_{k_{t,r}}(x) = \sum_{i=1}^{s} \sum_{j=0}^{p-1} U_i \nabla_i f_{\pi_m(j)}(x) = p \nabla F_m(x). \tag{2.2}$$

Algorithm 2.1: Function-Block-Alternate with a general ordering

1. Input x^0, stepsizes $\{\alpha_t\}$, frequencies m, and budget $p = \sum_k m_k$.
2. For $t = 0, \ldots, T-1$:
   (a) Choose an ordering ρ_t of $\mathcal{P}_t$; set $z^{t,0} = x^t$.
   (b) For $r = 0, \ldots, sp-1$, set $(i_{t,r}, k_{t,r}) = \rho_t(r)$, draw $G_{t,r} = \nabla_{i_{t,r}} g_{k_{t,r}}(z^{t,r}, \xi_{t,r})$, and set $z^{t,r+1} = z^{t,r} - \alpha_t U_{i_{t,r}} G_{t,r}$.
   (c) Set $x^{t+1} = z^{t,sp}$.

It is worth noting that Function-Block-Alternate generalizes several well-known optimization algorithms. First, in the single-objective case, when $q = 1$, our algorithm reduces to the (randomized) stochastic Block Coordinate Descent (BCD) method [41]. Second, when $s = 1$, i.e., the decision variable is treated as a single block, the algorithm simplifies to Function-Alternate for MOO (which generalizes the work [28] to more than two objectives and a flexible ordering). Also, when $s = 1$, one can see our approach as a way to replicate the final effect of the weighted-sum method [18]. Finally, in the most basic case where both $s = 1$ and $q = 1$, Function-Block-Alternate recovers the classical stochastic gradient descent algorithm for single-objective optimization.

3 Assumptions and technical estimates

We first assume that all of the objective functions are Lipschitz smooth, and the weighted-sum function F_m has a lower bound as follows.

Assumption 3.1 (Smoothness) *Each f_k has an L-Lipschitz continuous gradient ($L > 0$). Moreover,*

$$F_* := \inf_x F_m(x) > -\infty.$$

Let $\mathcal{F}_{t,r}$ contain the history through $z^{t,r}$ and the ordering selected at the beginning of a main iteration t. In addition to Assumption 3.1, we impose the following two assumptions.

Assumption 3.2 (Conditional unbiasedness and bounded variance) *For every update,*

$$\mathbb{E}[G_{t,r} \mid \mathcal{F}_{t,r}] = \nabla_{i_{t,r}} f_{k_{t,r}}(z^{t,r}), \qquad \mathbb{E}[\|G_{t,r} - \nabla_{i_{t,r}} f_{k_{t,r}}(z^{t,r})\|^2 \mid \mathcal{F}_{t,r}] \le \nu^2.$$

The samples need not be identically distributed, and no full stochastic gradient is required.

The above assumptions are commonly used for analyzing gradient-type methods, which guarantee that the stochastic gradient is unbiased and its variance is bounded.

Assumption 3.3 (Relative objective disagreement) *There are $\zeta \geq 0$ and $c \geq 0$ such that, for every x,*

$$H_m(x) := \frac{1}{p}\sum_{j=0}^{p-1}\|\nabla f_{\pi_m(j)}(x) - \nabla F_m(x)\|^2 \leq \zeta^2 + c\|\nabla F_m(x)\|^2. \tag{3.1}$$

Assumption 3.3 controls the dispersion of each objective function relative to the weighted-sum function F_m. As our algorithm evaluates different objectives at intermediate points, we use thus assumption to bound the corresponding error. Since it does not impose a bound on the gradient, the assumption is relatively mild, and the individual gradients can grow as $\nabla F_m(x)$ grows. The condition is independent of the ordering of the objective occurrences. For one objective, $H_m \equiv 0$. For two objectives with weights $w = m_1/p$ and $1 - w = m_2/p$, direct expansion gives $H_m(x) = w(1-w)\|\nabla f_1(x) - \nabla f_2(x)\|^2$.

The following lemma bounds the error caused by evaluating gradients at the successive points generated during a main iteration.

Lemma 3.1 (Detailed main-iteration estimates) *Suppose Assumptions 3.1–3.3 hold. Let*

$$a := 2sp^2(1+c), \quad b := 2sp^2\zeta^2 + s^2p^2\nu^2, \quad \text{and} \quad R(x) := a\|\nabla F_m(x)\|^2 + b,$$

We also denote $d_{t,r} = z^{t,r} - x^t$. If $\alpha_t \leq \frac{1}{2spL}$, then for every $r \leq sp$,

$$\mathbb{E}_t\|d_{t,r}\|^2 \leq 2\alpha_t^2 R(x^t), \tag{3.2}$$

$$\sum_{r=0}^{sp-1}\mathbb{E}_t\|d_{t,r}\| \leq sp\alpha_t\sqrt{2R(x^t)}, \tag{3.3}$$

where $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t,0}]$. Furthermore, define

$$K := s\sqrt{2a} + \sqrt{b/2}s + a/p \quad \text{and} \quad C_1 := Lsp\sqrt{b/2} + Lb.$$

If additionally $\alpha_t \leq \bar{\alpha} := \frac{1}{2KL} \leq \frac{1}{2spL}$, then

$$\mathbb{E}_t[F_m(x^{t+1})] \leq F_m(x^t) - \frac{p\alpha_t}{2}\|\nabla F_m(x^t)\|^2 + C_1\alpha_t^2. \tag{3.4}$$

Proof. We define $h_{t,r}(x) = U_{i_{t,r}}\nabla_{i_{t,r}} f_{k_{t,r}}(x)$, then fix t and abbreviate $\alpha = \alpha_t$, $x = x^t$, $z^r = z^{t,r}$, $d_r = d_{t,r}$, $i_r = i_{t,r}$, and $k_r = k_{t,r}$. Because $d_0 = 0$ and $z^{\ell+1} - z^\ell = -\alpha U_{i_\ell} G_{t,\ell}$, repeated substitution yields

$$d_r = -\alpha\sum_{\ell=0}^{r-1} U_{i_\ell} G_{t,\ell}. \tag{3.5}$$

Hence, using Cauchy-Schwarz in (3.5), the fact that U_{i_ℓ} is norm preserving, taking $\mathbb{E}_t$, and using $r \leq sp$, we obtain

$$\mathbb{E}_t\|d_r\|^2 \leq \alpha^2 r\sum_{\ell<r}\mathbb{E}_t\|G_{t,\ell}\|^2 \leq sp\alpha^2\sum_{\ell<r}\mathbb{E}_t\|G_{t,\ell}\|^2. \tag{3.6}$$

We write $G_{t,\ell} = \nabla_{i_\ell} f_{k_\ell}(z^\ell) + \varepsilon_\ell$ and from conditional unbiasedness we get $\mathbb{E}[\varepsilon_\ell \mid \mathcal{F}_{t,\ell}] = 0$. Expanding the square and taking the conditional expectation, we obtain

$$\begin{aligned}\mathbb{E}[\|G_{t,\ell}\|^2 \mid \mathcal{F}_{t,\ell}] &= \|\nabla_{i_\ell} f_{k_\ell}(z^\ell)\|^2 + 2\left\langle \nabla_{i_\ell} f_{k_\ell}(z^\ell), \mathbb{E}[\varepsilon_\ell \mid \mathcal{F}_{t,\ell}]\right\rangle + \mathbb{E}[\|\varepsilon_\ell\|^2 \mid \mathcal{F}_{t,\ell}] \\ &\le \|\nabla_{i_\ell} f_{k_\ell}(z^\ell)\|^2 + \nu^2.\end{aligned}$$

Lipschitz continuity and the fact that U_{i_ℓ} is norm preserving imply the bound

$$\begin{aligned}\|\nabla_{i_\ell} f_{k_\ell}(z^\ell)\|^2 &\le 2\|\nabla_{i_\ell} f_{k_\ell}(x)\|^2 + 2\|\nabla_{i_\ell} f_{k_\ell}(z^\ell) - \nabla_{i_\ell} f_{k_\ell}(x)\|^2 \\ &\le 2\|h_{t,\ell}(x)\|^2 + 2L^2\|d_\ell\|^2.\end{aligned}$$

By using the tower property and substituting the terms in (3.6), we have

$$\mathbb{E}_t\|d_r\|^2 \le 2sp\alpha^2 \sum_{\ell<r}\|h_{t,\ell}(x)\|^2 + 2\alpha^2 spL^2 \sum_{\ell<r}\mathbb{E}_t\|d_\ell\|^2 + s^2p^2\alpha^2\nu^2. \tag{3.7}$$

We next bound the first sum. Since the order only permutes the prescribed pairs,

$$\begin{aligned}\sum_{r=0}^{sp-1}\|h_{t,r}(x)\|^2 = \sum_{i=1}^{s}\sum_{j=0}^{p-1}\|\nabla_i f_{\pi_m(j)}(x)\|^2 \;&=\; \sum_{j=0}^{p-1}\|\nabla f_{\pi_m(j)}(x)\|^2 \\ = \sum_{j=0}^{p-1}\|\nabla f_{\pi_m(j)}(x) - \nabla F_m(x)\|^2 + p\|\nabla F_m(x)\|^2 &\le p\big((1+c)\|\nabla F_m(x)\|^2 + \zeta^2\big),\end{aligned}$$

where the last inequality follows from Assumption 3.3. Let $D = \max_{0\le r\le sp}\mathbb{E}_t\|d_r\|^2$. From (3.7) and definition of D we have

$$D \le 2sp^2\alpha^2\big((1+c)\|\nabla F_m(x)\|^2 + \zeta^2\big) + 2s^2p^2\alpha^2L^2D + s^2p^2\alpha^2\nu^2 \le \alpha^2 R(x) + \frac{1}{2}D.$$

The last inequality uses $2sp\alpha L \le 1$ and the definition of $R(x)$. Subtracting $D/2$ from both sides proves $D \le 2\alpha^2 R(x)$, which is (3.2). From, Jensen's inequality and summation over r we obtain (3.3).

It remains to prove descent. From the smoothness of F_m we get

$$F_m(x^{t+1}) \le F_m(x) + \left\langle \nabla F_m(x), x^{t+1} - x\right\rangle + \frac{L}{2}\|x^{t+1} - x\|^2. \tag{3.8}$$

From the updates,

$$x^{t+1} - x = -\alpha\sum_{r=0}^{sp-1} U_{i_r} G_{t,r}.$$

We analyze the linear term by conditioning successively on $\mathcal{F}_{t,r}$ and noting that the gradient is conditionally unbiased:

$$\begin{aligned}\mathbb{E}_t\left\langle \nabla F_m(x), x^{t+1} - x\right\rangle &= -\alpha\sum_r \mathbb{E}_t\left\langle \nabla F_m(x), h_{t,r}(z^r)\right\rangle \\ &= -\alpha\sum_r \left\langle \nabla F_m(x), h_{t,r}(x)\right\rangle - \alpha\sum_r \mathbb{E}_t\left\langle \nabla F_m(x), h_{t,r}(z^r) - h_{t,r}(x)\right\rangle\end{aligned}$$

$$= -p\alpha\|\nabla F_m(x)\|^2 - \alpha \sum_r \mathbb{E}_t \left\langle \nabla F_m(x), h_{t,r}(z^r) - h_{t,r}(x) \right\rangle,$$

where the last equality uses (2.2). We bound the final term using Cauchy-Schwarz inequality, block Lipschitz continuity, and (3.3):

$$\begin{aligned}
&\left| \alpha \sum_r \mathbb{E}_t \left\langle \nabla F_m(x), h_{t,r}(z^r) - h_{t,r}(x) \right\rangle \right| \\
&\leq \alpha \|\nabla F_m(x)\| \sum_{r=0}^{sp-1} \mathbb{E}_t \|U_{i_r}\big(\nabla_{i_r} f_{k_r}(z^r) - \nabla_{i_r} f_{k_r}(x)\big)\| \leq \alpha L \|\nabla F_m(x)\| \sum_r \mathbb{E}_t \|z^r - x\| \\
&\leq Lsp\alpha^2 \|\nabla F_m(x)\| \sqrt{2R(x)} = Lsp\alpha^2 \|\nabla F_m(x)\| \sqrt{2a\|\nabla F_m(x)\|^2 + 2b} \\
&\leq Lsp\alpha^2 \|\nabla F_m(x)\| \left(\sqrt{2a}\|\nabla F_m(x)\| + \sqrt{2b} \right) \\
&\leq Lsp\alpha^2 \|\nabla F_m(x)\|^2 \sqrt{2a} + \sqrt{2b} Lsp\alpha^2 \|\nabla F_m(x)\| \\
&\leq Lsp\alpha^2 \|\nabla F_m(x)\|^2 \sqrt{2a} + \sqrt{b/2} Lsp\alpha^2 \left(\|\nabla F_m(x)\|^2 + 1 \right),
\end{aligned}$$

where we apply the inequality $2g \leq g^2 + 1$. For the quadratic term in (3.8), applying (3.2) with $r = sp$ yields:

$$\frac{L}{2}\mathbb{E}_t \|x^{t+1} - x\|^2 \leq \frac{L}{2}(2\alpha^2 R(x)) = L\alpha^2 \left(a\|\nabla F_m(x)\|^2 + b \right).$$

Combining these terms with (3.8), we obtain

$$\begin{aligned}
\mathbb{E}_t[F_m(x^{t+1})] &\leq F_m(x) - p\alpha\|\nabla F_m(x)\|^2 \\
&\quad + Lsp\alpha^2 \|\nabla F_m(x)\|^2 \sqrt{2a} + \sqrt{b/2} Lsp\alpha^2 \left(\|\nabla F_m(x)\|^2 + 1 \right) + L\alpha^2 \left(a\|\nabla F_m(x)\|^2 + b \right) \\
&\leq F_m(x) - p\alpha\|\nabla F_m(x)\|^2 \\
&\quad + L\alpha^2 \|\nabla F_m(x)\|^2 \left(sp\sqrt{2a} + \sqrt{b/2}sp + a \right) + L\alpha^2(\sqrt{b/2}sp + b).
\end{aligned}$$

We use the condition $\alpha 2L \left(s\sqrt{2a} + \sqrt{b/2}s + a/p \right) \leq 1$ and the definition of C_1 to conclude that

$$\mathbb{E}_t[F_m(x^{t+1})] \leq F_m(x) - \frac{1}{2}p\alpha\|\nabla F_m(x)\|^2 + C_1\alpha^2.$$

This proves the main descent bound. □

4 Convergence analysis

4.1 Smooth nonconvex objectives

We first establish the convergence rate for smooth nonconvex objective functions under the assumptions of Section 3.

Theorem 4.1 *Under Assumptions 3.1–3.3, choose* $\alpha_t = \alpha = \min\{\bar{\alpha}, \gamma T^{-1/2}\}$, *where* $\bar{\alpha}$ *is defined in Lemma 3.1 and* $\gamma > 0$. *Select* $\bar{x}$ *uniformly from* $\{x^0, \ldots, x^{T-1}\}$. *Then*

$$\mathbb{E}\|\nabla F_m(\bar{x})\|^2 \leq \frac{2(F_m(x^0) - F_*)(\gamma + \bar{\alpha}T^{1/2})}{\gamma\bar{\alpha}pT} + \frac{2C_1\gamma}{pT^{1/2}}.$$

Proof. Taking expectation and applying the tower property in the bound (3.4), we have

$$\mathbb{E}F_m(x^{t+1}) \leq \mathbb{E}F_m(x^t) - \frac{p\alpha_t}{2}\mathbb{E}\|\nabla F_m(x^t)\|^2 + C_1\alpha_t^2.$$

Rearranging this inequality and summing from $t = 0$ to $T - 1$ yields

$$\frac{p\alpha_t}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F_m(x^t)\|^2 \leq F_m(x^0) - \mathbb{E}F_m(x^T) + TC_1\alpha_t^2 \leq F_m(x^0) - F_* + TC_1\alpha_t^2.$$

Dividing both sides by $pT\alpha_t/2$ and noting that $\bar{x}$ is sampled uniformly from $\{x^0, \ldots, x^{T-1}\}$ gives us:

$$\begin{aligned}
\mathbb{E}\|\nabla F_m(\bar{x})\|^2 &\leq \frac{2(F_m(x^0) - F_*)}{pT\alpha_t} + \frac{2C_1\alpha_t}{p} \\
&\leq \frac{2(F_m(x^0) - F_*)}{pT\min\{\bar{\alpha}, \gamma T^{-1/2}\}} + \frac{2C_1\min\{\bar{\alpha}, \gamma T^{-1/2}\}}{p} \\
&\leq \frac{2(F_m(x^0) - F_*)(\gamma + \bar{\alpha}T^{1/2})}{\gamma\bar{\alpha}pT} + \frac{2C_1\gamma}{pT^{1/2}},
\end{aligned}$$

□

Our convergence rate for Function-Block-Alternate recovers the classical convergence rate $\mathcal{O}(1/T^{1/2})$ of stochastic gradient descent for single-objective optimization [19]. We note that for MOO, the work [28] does not consider the non-convex setting. Meanwhile, in single-objective optimization, [41] only considers asymptotic theoretical non-convex results for their stochastic BCD algorithm.

4.2 Convex objectives

For the convex case, we impose the existence of a minimizer of the weighted-sum function.

Assumption 4.1 (Convexity and attainability) *F_m is convex and attains its minimum at x^*.*

The following lemma shows a recursion bound for the expected squared distance from the iterate x^t to the solution x^*.

Lemma 4.2 (Distance recursion) *Under Assumptions 3.1–4.1, define*

$$A_0 := Lsp(\sqrt{2a} + \sqrt{2b}), \quad A_1 := Lsp\sqrt{2a} + 2a, \quad \text{and} \quad A_2 := Lsp\sqrt{2b} + 2b.$$

For $\alpha_t \leq \bar{\alpha}$,

$$D_{t+1}^2 \leq (1 + A_0\alpha_t^2)D_t^2 - 2p\alpha_t\mathbb{E}[F_m(x^t) - F_*] + A_1\alpha_t^2\mathbb{E}\|\nabla F_m(x^t)\|^2 + A_2\alpha_t^2, \tag{4.1}$$

where $D_t^2 = \mathbb{E}\|x^t - x^*\|^2$.

Proof. We start with the following equality

$$\|x^{t+1} - x^*\|^2 = \|x^t - x^*\|^2 + 2\left\langle x^t - x^*, x^{t+1} - x^t\right\rangle + \|x^{t+1} - x^t\|^2.$$

Using similar argument as in Lemma 3.1, we analyze the linear term:

$$\begin{aligned}
\mathbb{E}_t\left\langle x^t - x^*, x^{t+1} - x^t\right\rangle &= -\alpha_t \sum_r \mathbb{E}_t\left\langle x^t - x^*, h_{t,r}(z^{t,r})\right\rangle \\
&= -\alpha_t \sum_r \mathbb{E}_t\left\langle x^t - x^*, h_{t,r}(x^t)\right\rangle - \alpha_t \sum_r \mathbb{E}_t\left\langle x^t - x^*, h_{t,r}(z^{t,r}) - h_{t,r}(x^t)\right\rangle \\
&= -p\alpha_t\left\langle x^t - x^*, \nabla F_m(x^t)\right\rangle - \alpha_t \sum_r \mathbb{E}_t\left\langle x^t - x^*, h_{t,r}(z^{t,r}) - h_{t,r}(x^t)\right\rangle.
\end{aligned}$$

To analyze the first term, using the convexity of F_m we get

$$\left\langle x^t - x^*, \nabla F_m(x^t)\right\rangle \geq F_m(x^t) - F_m(x^*) = F_m(x^t) - F_*.$$

Similar to Lemma 3.1, to analyze the second term, we use Cauchy-Schwarz inequality, Lipschitz continuity, and (3.3):

$$\begin{aligned}
&2\left|\alpha_t \sum_r \mathbb{E}_t\left\langle x^t - x^*, h_{t,r}(z^{t,r}) - h_{t,r}(x^t)\right\rangle\right| \\
&\leq 2\alpha_t L\|x^t - x^*\| \sum_r \mathbb{E}_t\|d_{t,r}\| \leq 2Lsp\alpha_t^2\|x^t - x^*\|\sqrt{2R(x^t)} \\
&\leq 2Lsp\alpha_t^2\|x^t - x^*\|\left(\sqrt{2a}\|\nabla F_m(x^t)\| + \sqrt{2b}\right) \\
&\leq Lsp\alpha_t^2\sqrt{2a}\left(\|x^t - x^*\|^2 + \|\nabla F_m(x^t)\|^2\right) + Lsp\alpha_t^2\sqrt{2b}\left(\|x^t - x^*\|^2 + 1\right).
\end{aligned}$$

In addition, we note that (3.2) gives the estimate

$$\mathbb{E}_t\|x^{t+1} - x^t\|^2 = \mathbb{E}_t\|d_{t,sp}\|^2 \leq 2\alpha_t^2(a\|\nabla F_m(x^t)\|^2 + b).$$

From these bounds, we finally obtain

$$\begin{aligned}
\mathbb{E}_t\|x^{t+1} - x^*\|^2 \leq\ & \mathbb{E}_t\|x^t - x^*\|^2 - 2p\alpha_t[F_m(x^t) - F_*] \\
&+ Lsp\alpha_t^2\sqrt{2a}\left(\|x^t - x^*\|^2 + \|\nabla F_m(x^t)\|^2\right) + Lsp\alpha_t^2\sqrt{2b}\left(\|x^t - x^*\|^2 + 1\right) \\
&+ 2\alpha_t^2(a\|\nabla F_m(x^t)\|^2 + b).
\end{aligned}$$

Taking total expectation, we arrive at (4.1). □

From this lemma, we are ready to show the convex results as follows.

Theorem 4.3 (Convex rate) *Under Assumptions 3.1–4.1, choose* $\alpha_t = \alpha = \min\{\bar{\alpha}, \gamma T^{-1/2}\}$, *where* $\bar{\alpha}$ *is defined in Lemma 3.1 and* $\gamma > 0$. *Let* $\widehat{x}_T = T^{-1}\sum_{t=0}^{T-1} x^t$. *Additionally, we define*

$$V_t := D_t^2 + \frac{2A_1}{p}\alpha\mathbb{E}[F_m(x^t) - F_*], \qquad C_2 := \frac{2A_1C_1}{p}\bar{\alpha} + A_2,$$

Then

$$\mathbb{E}[F_m(\widehat{x}_T) - F_*] \leq \frac{V_0(\gamma + \bar{\alpha}T^{1/2})}{2pT\gamma\bar{\alpha}} + \frac{[A_0e^{A_0\gamma^2}(V_0 + C_2\gamma^2) + C_2]\gamma}{2pT^{1/2}}.$$

Proof. Throughout this proof, α denotes the common value of α_t for $t = 0, \ldots, T-1$. Rearrange the expected descent inequality (3.4) we get

$$\frac{p\alpha}{2}\mathbb{E}\|\nabla F_m(x^t)\|^2 \leq \mathbb{E}[F_m(x^t) - F_m(x^{t+1})] + C_1\alpha^2.$$

Multiply both sides by $2A_1\alpha/p$ yields

$$A_1\alpha^2\mathbb{E}\|\nabla F_m(x^t)\|^2 \leq \frac{2A_1\alpha}{p}\mathbb{E}[F_m(x^t) - F_m(x^{t+1})] + \frac{2A_1C_1}{p}\alpha^3.$$

Substitute this bound into (4.1) we get

$$D_{t+1}^2 \leq (1 + A_0\alpha^2)D_t^2 - 2p\alpha\mathbb{E}[F_m(x^t) - F_*] + \frac{2A_1\alpha}{p}\mathbb{E}[F_m(x^t) - F_m(x^{t+1})] + \frac{2A_1C_1}{p}\alpha^3 + A_2\alpha^2,$$

which leads to

$$\begin{aligned} D_{t+1}^2 + \frac{2A_1\alpha}{p}\mathbb{E}[F_m(x^{t+1}) - F_*] \leq (1 + A_0\alpha^2)D_t^2 + \frac{2A_1\alpha}{p}\mathbb{E}[F_m(x^t) - F_*] \\ - 2p\alpha\mathbb{E}[F_m(x^t) - F_*] + \left(\frac{2A_1C_1}{p}\bar{\alpha} + A_2\right)\alpha^2. \end{aligned}$$

Hence by definition of V_t and C_2, we have the following recursive bound

$$V_{t+1} \leq (1 + A_0\alpha^2)V_t - 2p\alpha\mathbb{E}[F_m(x^t) - F_*] + C_2\alpha^2. \tag{4.2}$$

Dropping the term $2p\alpha\mathbb{E}[F_m(x^t) - F_*]$ and applying this bound recursively, we obtain

$$V_t \leq (1 + A_0\alpha^2)^t V_0 + C_2\alpha^2\sum_{j=0}^{t-1}(1 + A_0\alpha^2)^j.$$

As $(1 + A_0\alpha^2)^t \leq e^{A_0t\alpha^2} \leq e^{A_0\gamma^2}$ for $t \leq T$, we have

$$V_t \leq e^{A_0\gamma^2}V_0 + C_2t\alpha^2e^{A_0\gamma^2} \leq e^{A_0\gamma^2}(V_0 + C_2\gamma^2).$$

Now we return to (4.2). Summing from $t = 0$ to $T - 1$ we get

$$2p\alpha\sum_{t=0}^{T-1}\mathbb{E}[F_m(x^t) - F_*] \leq V_0 - V_T + A_0\alpha^2\sum_{t=0}^{T-1}V_t + TC_2\alpha^2.$$

Use $0 \leq V_t \leq e^{A_0\gamma^2}(V_0 + C_2\gamma^2)$, and divide by $2pT\alpha$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[F_m(x^t) - F_*] \leq \frac{V_0}{2pT\alpha} + \frac{[A_0e^{A_0\gamma^2}(V_0 + C_2\gamma^2) + C_2]\alpha}{2p}.$$

Substituting $\alpha = \min\{\bar{\alpha}, \gamma T^{-1/2}\}$ we get

$$\mathbb{E}[F_m(\widehat{x}_T) - F_*] \leq \frac{V_0(\gamma + \bar{\alpha}T^{1/2})}{2pT\gamma\bar{\alpha}} + \frac{[A_0e^{A_0\gamma^2}(V_0 + C_2\gamma^2) + C_2]\gamma}{2pT^{1/2}}.$$

□

To compute the whole Pareto front, Algorithm 2.1 can be applied to a collection of frequency vectors whose normalized weights sample the $(q-1)$-dimensional simplex. The convex convergence result applies to each scalarization whenever the corresponding F_m is convex and attains its minimum. Note that F_m is convex for all m if all the individual functions are convex.

Theorem 4.3 recovers the standard convergence rate $\mathcal{O}(1/T^{1/2})$ of BCD single-objective optimization [19, 41] and of stochastic alternating function minimization for bi-objective optimization [28].

4.3 Polyak-Lojasiewicz functions

In this section, we consider the Polyak-Łojasiewicz (PL) inequality, a generalization of strong-convexity [25, 34] , which we impose for the weighted-sum function F_m.

Assumption 4.2 (PL condition) *There is $\mu > 0$ such that $\|\nabla F_m(x)\|^2 \geq 2\mu(F_m(x) - F_*)$ for every x.*

It is well known that a function satisfying the PL condition is not necessarily convex [25]. Under this assumption, one can show that single-objective stochastic gradient descent achieves the same theoretical rate as the sublinear rate $\mathcal{O}(1/T)$ in the strongly convex setting [11, 21]. Similarly, we show the same convergence rate for the optimality gap of Function-Block-Alternate as follows.

Theorem 4.4 *Under Assumptions 3.1–3.3 and 4.2, choose*

$$\alpha_t = \min\{\bar{\alpha}, 2/[\mu p(t+2)]\},$$

where $\bar{\alpha}$ is defined in Lemma 3.1. Let

$$T_0 = \left\lfloor \frac{2}{p\mu\bar{\alpha}} \right\rfloor \quad \textit{and} \quad K = \max\left\{ (F_m(x^0) - F_* + T_0 C_1 \bar{\alpha}^2)(T_0+1), \frac{4C_1}{\mu^2 p^2} \right\}.$$

Then we have

$$\mathbb{E}[F_m(x^T) - F_*] \leq \frac{K}{T+1}, \quad \textit{for} \quad T \geq T_0.$$

Proof. Taking expectation in (3.4) and applying the PL inequality gives

$$\begin{aligned}\mathbb{E}[F_m(x^{t+1}) - F_*] &\leq \mathbb{E}[F_m(x^t) - F_*] - \frac{p\alpha_t}{2}(2\mu\mathbb{E}[F_m(x^t) - F_*]) + C_1\alpha_t^2 \\ &= (1 - \mu p\alpha_t)\mathbb{E}[F_m(x^t) - F_*] + C_1\alpha_t^2.\end{aligned}$$

After a finite number of initial iterations $T_0 = \lfloor \frac{2}{p\mu\bar{\alpha}} \rfloor$, we have $\alpha_t = 2/[\mu p(t+2)]$. As the quantity $\mathbb{E}[F_m(x^t) - F_*]$ increases by at most $C_1\bar{\alpha}^2$ after each iteration, we can only consider the algorithm process for $t \geq T_0$. Then

$$\mathbb{E}[F_m(x^{t+1}) - F_*] \leq \frac{t}{t+2}\mathbb{E}[F_m(x^t) - F_*] + \frac{4C_1}{\mu^2 p^2 (t+2)^2}.$$

We prove $\mathbb{E}[F_m(x^t) - F_*] \leq K/(t+1)$ by induction for $t \geq T_0$. It holds at $t = T_0$. Assuming it at t, we obtain

$$\begin{aligned}\mathbb{E}[F_m(x^{t+1}) - F_*] &\leq \frac{tK}{(t+1)(t+2)} + \frac{K}{(t+2)^2} \\ &\leq \frac{tK}{(t+1)(t+2)} + \frac{K}{(t+1)(t+2)} = \frac{K}{t+2},\end{aligned}$$

which shows the bound for $t+1$. □

The Polyak-Łojasiewicz convergence rate analysis of Function-Block-Alternate matches the strongly convex convergence rate $\mathcal{O}(1/T)$ of BCD single-objective optimization [19, 41] and the stochastic alternating function minimization for bi-objective optimization [29]. However, the Polyak-Łojasiewicz condition is weaker than assuming strong convexity, and in this case we do not need to assume bounded iterates as in [28].

5 Experiments

We compare Pareto-front approximations under three gradient-access regimes: all objective gradients are available; the full gradient of only one objective is used at each elementary iteration; and only one objective-block partial gradient is available at each elementary iteration. These regimes allow us to separate the effects of function alternation and block alternation.

5.1 Multi-environment regression formulation

The first two problems have the following form. Environment k provides a design matrix $X_k \in \mathbb{R}^{N_k \times n}$ and a response vector $y_k \in \mathbb{R}^{N_k}$, defining the objective

$$f_k(x) = \frac{1}{N_k}\|y_k - X_k x\|^2, \qquad k = 1, \ldots, q.$$

The same decision vector $x \in \mathbb{R}^n$ is used in all environments. When the data-generating regression parameters differ across environments, the objectives generally have different minimizers and cannot all be minimized simultaneously. Each environment therefore defines one objective function, while groups of features define the variable blocks.

5.2 Problems and gradient-access regimes

We consider three problems, each with three conflicting objectives.

Problem 1: synthetic multi-environment regression. The decision vector contains 48 regression parameters. Each of the three environments has 4096 training observations and 2048 test observations. In environment k, the response is generated from a regression parameter $\beta_k \in \mathbb{R}^{48}$. Each β_k is the sum of a component common to all three environments and an environment-specific component whose direction and sign vary with k. The optimization variable x is a single regression vector shared by the three environments. Consequently, x cannot in general equal all three individual least-squares minimizers. All three objective gradients are available at every elementary iteration. We compare Weighted-Sum, Function-Alternate, Block-Alternate,

and Function-Block-Alternate. The comparison between Block-Alternate and Function-Block-Alternate also covers the regime in which only partial gradients with respect to one variable block are available across all objectives, since both methods operate within this restriction.

Problem 2: Beijing multi-site air-quality regression. We use the Beijing Multi-Site Air Quality data [26, 44]. The twelve monitoring stations are divided into three groups of four stations. For each group, one objective is the mean squared prediction error for $PM_{2.5}$ over the observations from that group. Thus, the problem has three objectives, all predicting $PM_{2.5}$ but using observations from different station groups. A common nine-dimensional regression vector is used for all groups, and station-indicator variables are excluded. The three groups exhibit different pollution patterns and therefore yield different individual least-squares minimizers.

To study function alternation, we impose a separate-objective access rule: at each elementary iteration, an algorithm may use the gradient of only one selected station-group objective. Although all three objective gradients can be computed from the dataset, the algorithm is not permitted to combine them within a single elementary iteration. Under this rule, we compare cyclic and randomized Function-Alternate with cyclic and randomized Function-Block-Alternate. We do not include Weighted-Sum or Block-Alternate because their elementary updates use gradients from all three objectives and therefore do not satisfy the one-objective-per-iteration rule.

Problem 3: human-activity classification. We use the Human Activity Recognition Using Smartphones data [37]. Its 561 accelerometer- and gyroscope-derived features are used to fit a common scalar score for three mutually exclusive activity classes: walking activities, sitting or standing, and lying. Each class defines one binary least-squares objective, giving three objectives with incompatible targets. At each elementary iteration, the oracle returns the partial gradient of one selected objective with respect to one selected feature block. Function-Alternate is not applicable under this oracle because its elementary iteration requires the full gradient of the selected objective, whereas the oracle supplies only one block of that gradient. We therefore compare cyclic and randomized orderings of Function-Block-Alternate.

5.3 Experimental setup

The variables are partitioned into three blocks. In Problems 1 and 2, we consider all 231 nonnegative integer vectors (m_1, m_2, m_3) satisfying $m_1 + m_2 + m_3 = 20$. In Problem 3, because the decision space is substantially larger, we consider all 66 vectors satisfying $m_1 + m_2 + m_3 = 10$. Each vector determines a weighted-sum scalarization and produces one candidate objective vector.

All problem and algorithmic parameters were fixed by the following common protocol. Every method starts from the zero vector, and the coordinates are split, without using response information, into three contiguous blocks whose sizes differ by at most one. Stepsizes have the form c/L, where $L = \max_k 2\lambda_{\max}(X_k^\top X_k/N_k)$. For every method, the value c is chosen from the same grid $\{0.03, 0.06, 0.12, 0.25, 0.50, 0.75\}$ by minimizing a training loss consisting of equal objective weights.

In Problem 1, for a given frequency vector m, Weighted-Sum takes full-gradient steps on $F_m(x)$, whereas Block-Alternate applies the gradient of the same scalarization to one variable block at a time. In Problems 2 and 3, the cyclic and randomized variants differ as follows. For Function-Alternate, the cyclic variant repeatedly uses a fixed order of the function steps, whereas

the other one randomly permutes the function-step sequence at the beginning of each main iteration. For Function-Block-Alternate, the cyclic variant uses fixed function and block orders and the other one randomly permutes both the function-step sequence and the block sequence at the beginning of each main iteration. The resulting ordering ρ_t is therefore determined by both randomized orders, rather than by shuffling only the functions. Randomized methods are run with ten random seeds while cyclic methods contain no randomness.

All methods compared within a problem use the same computational budget, measured in objective-block gradient evaluations. One partial gradient $\nabla_i f_k$ has unit cost. Because there are three variable blocks, a full gradient ∇f_k costs three units. Computing the gradient of a weighted sum of all three objectives costs nine units. The budget for each scalarization is 5400 units in Problems 1 and 2 and 900 units in Problem 3. Thus, the comparison accounts for the different amounts of first-order information required by full-gradient and block-gradient iterations. In Problem 3, objective-block partial gradients are not assigned an artificial lower cost; they are the only gradients supplied by the oracle.

The algorithms minimize the scalarized training objectives. After optimization, we evaluate each computed solution on the three corresponding test objectives. For each method, the nondominated test-objective vectors form its Pareto-front approximation.

5.4 Pareto-front metrics and numerical results

For each problem, let A_a denote the Pareto-front approximation produced by method a. We form the empirical reference set

$$R = \mathrm{ND}\left(\bigcup_a A_a\right),$$

where ND denotes the nondominated subset.

We use four established Pareto-front metrics. Purity [9] is the fraction of the points in A_a that also belong to R; larger values are better. Before computing hypervolume [7], each objective is scaled to $[0, 1]$ using the smallest and largest values of that objective over $\bigcup_a A_a$. Hypervolume is then computed with the reference point $(1.1, 1.1, 1.1)$, which is componentwise worse than every normalized point; larger values are better. The maximum spread Γ [9] is the largest gap between consecutive points along any objective, including the gaps to the extreme values in R. The generalized spread Δ [10] measures both the nonuniformity of these gaps and failure to attain the extreme values in R. Smaller values are better for both spread metrics.

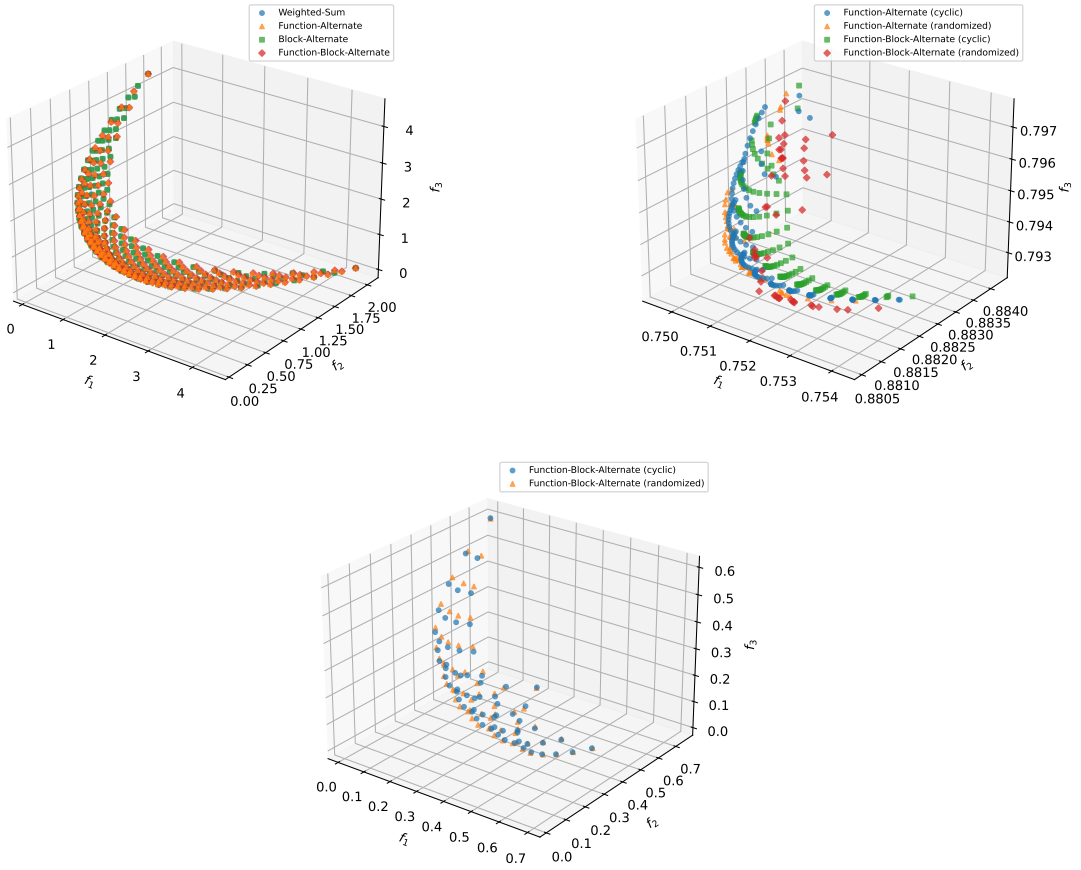


Figure 1: Pareto-front approximations in test-objective space for Problem 1 (top left), Problem 2 (top right), and Problem 3 (bottom). Only methods applicable under each problem's gradient-access regime are shown.

Table 1: Metrics for the Pareto-front approximations in three problems. For the random methods, the numbers reported are the averages over ten comparisons

. Higher is better for purity and hypervolume; lower is better for Γ and Δ.

Problem	Method	Purity	Hypervolume	Γ	Δ
1	Weighted-Sum	**0.991**	**1.06095**	0.0963	0.9797
	Function-Alternate	0.892	1.05653	**0.0641**	**0.9011**
	Block-Alternate	**0.991**	**1.06095**	0.0963	0.9797
	Function-Block-Alternate	0.974	1.05695	**0.0647**	**0.9168**
2	Function-Alternate (cyclic)	0.547	0.89478	**0.2022**	1.1459
	Function-Alternate (randomized)	0.653	0.92455	0.2668	1.1122
	Function-Block-Alternate (cyclic)	0.059	0.85284	0.2499	1.6235
	Function-Block-Alternate (randomized)	**0.709**	**0.95725**	0.2664	**1.1069**
3	Function-Block-Alternate (cyclic)	0.959	0.89538	0.1360	**0.8219**
	Function-Block-Alternate (randomized)	**0.998**	**0.90763**	**0.1233**	0.8399

In Problem 1, Weighted-Sum and Block-Alternate attain the largest purity and hypervolume, whereas Function-Alternate gives the smallest values of Γ and Δ. Function-Block-Alternate

remains close to Function-Alternate on all four metrics. Thus, no method is best according to every metric, and the alternating methods obtain more evenly distributed fronts despite realizing the scalarization indirectly through sequential objective updates.

In Problem 2, randomized Function-Block-Alternate attains the largest mean purity and hypervolume and the smallest mean Δ, whereas cyclic Function-Alternate gives the smallest mean Γ. Randomization improves purity, hypervolume, and Δ for both alternating methods, but increases Γ. Thus, the effect of ordering depends on the chosen Pareto-front metric.

In Problem 3, randomized Function-Block-Alternate improves mean purity, hypervolume, and Γ, while the cyclic variant gives the smaller mean Δ. Thus, under objective–block partial-gradient access, randomization improves convergence and the largest-gap measure, but not the overall uniformity measure.

6 Concluding remarks

The analysis shows that function and block alternation can be combined without imposing a particular order on the objective–block updates. The frequency with which each objective is selected determines the weighted-sum scalarization being optimized, while the ordering of the updates affects only the error generated by evaluations at intermediate points. This separation between frequency and order is the main mechanism that allows the convergence analysis to cover deterministic and randomized orderings and any number of objectives.

The convergence results require neither uniformly bounded stochastic gradients nor bounded iterates. The analysis requires conditional bounded variance on the stochastic gradients and a bounded relative disagreement condition for the objective functions. Under these conditions, Function-Block-Alternate achieves the standard $\mathcal{O}(T^{-1/2})$ rates for smooth nonconvex and convex objectives and the $\mathcal{O}(T^{-1})$ rate under the Polyak–Łojasiewicz condition.

The numerical results clarify when the different forms of alternation are useful. When all objective gradients are available, no method is best according to every Pareto-front metric, and the alternating methods produce more evenly distributed fronts. This is remarkable because Function-Block-Alternate realizes the scalarization indirectly through repeated, sequential steps on individual objective functions, rather than by applying each step directly to the weighted-sum objective. When only one objective gradient can be accessed at a time, the results depend on both the method and the ordering; in Problem 2, randomized Function-Block-Alternate gives the best mean purity, hypervolume, and generalized spread. When only one objective-block partial gradient is available at a time, randomization improves three of the four mean Pareto-front metrics in Problem 3. Thus, experiments show the value of function and block alternation and the randomized flexible ordering in Pareto-front approximation.

Acknowledgments

This work is partially supported by the U.S. Air Force Office of Scientific Research (AFOSR) award FA9550-23-1-0217 and the U.S. Office of Naval Research (ONR) award N000142412656.

References

[1] S. Bechikh, R. Datta, and A. Gupta, editors. *Recent Advances in Evolutionary Multi-Objective Optimization*, volume 20. Springer, New York, 2016.

[2] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM J. Optim.*, 23:2037–2060, 2013.

[3] E. K. Browning and M. A. Zupan. *Microeconomics: Theory and Applications*. Wiley, Hoboken, 2020.

[4] X. Cai, C. Song, S. Wright, and J. Diakonikolas. Cyclic block coordinate descent with variance reduction for composite nonconvex optimization. *International Conference on Machine Learning*, 202:3469–3494, 2023.

[5] L. Chen, H. Fernando, Y. Ying, and T. Chen. Three-Way Trade-Off in Multi-Objective Learning: Optimization, Generalization and Conflict-Avoidance. *Advances in Neural Information Processing Systems*, 36:70045–70093, 2023.

[6] P. Cheng, J. Sulaiman, K. Ghazali, M. K. M. Ali, and M. Xu. Application of Newton-Gauss-Seidel method for solving multi-objective constrained optimization problems. *Transactions on Science and Technology*, 11:43–50, 2024.

[7] C. C. Coello. Evolutionary multi-objective optimization: A historical view of the field. *IEEE Computational Intelligence Magazine*, 1:28–36, 2006.

[8] K. H. Cuevas. *Cyclic stochastic optimization: Generalizations, convergence, and applications in multi-agent systems*. PhD thesis, The Johns Hopkins University, 2017.

[9] A. L. Custódio, J. A. Madeira, A. I. F. Vaz, and L. N. Vicente. Direct multisearch for multiobjective optimization. *SIAM J. Optim.*, 21:1109–1140, 2011.

[10] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6:182–197, 2002.

[11] S. De, A. Yadav, D. Jacobs, and T. Goldstein. Automated Inference with Adaptive Batches. In *Proceedings of Machine Learning Research*, volume 54, pages 1504–1513, 2017.

[12] J. A. Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *C. R. Math. Acad. Sci. Paris*, 350:313–318, 2012.

[13] D. Driggs, J. Tang, J. Liang, M. Davies, and C. Schonlieb. A stochastic proximal alternating minimization for nonsmooth and nonconvex optimization. *SIAM J. Imaging Sci.*, 14:1932–1970, 2021.

[14] M. Ehrgott. *Multicriteria Optimization*, volume 491. Springer Science & Business Media, Berlin, 2005.

[15] H. Fernando, H. Shen, M. Liu, S. Chaudhury, K. Murugesan, and T. Chen. Mitigating gradient bias in multi-objective learning: A provably convergent approach. *International Conference on Learning Representation*, 2023.

[16] H. Fernando, L. Chen, S. Lu, P. Chen, M. Liu, S. Chaudhury, K. Murugesan, G. Liu, M. Wang, and T. Chen. Variance reduction can improve trade-off in multi-objective learning. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6975–6979, 2024.

[17] E. H. Fukuda and L. M. G. Drummond. A survey on multiobjective descent methods. *Pesquisa Operacional*, 34:585–620, 2014.

[18] S. Gass and T. Saaty. The computational algorithm for the parametric objective function. *Nav. Res. Logist. Q.*, 2:39–45, 1955.

[19] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23:2341–2368, 2013.

[20] L. Grippo and M. Sciandrone. On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Oper. Res. Lett.*, 26:127–136, 2000.

[21] R. Gower, O. Sebbouh, and N. Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1315–1323. PMLR, 2021.

[22] N. Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5:1502242, 2018.

[23] Y. V. Haimes. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE Transactions on Systems, Man, and Cybernetics*, 1: 296–297, 1971.

[24] T. Jiang and L. Xiao. Stochastic approximation with block coordinate optimal stepsizes. *arXiv preprint arXiv:2507.08963*, 2025.

[25] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, volume 9851, pages 795–811. Springer International Publishing, Cham, 2016.

[26] M. Kelly, R. Longjohn, and K. Nottingham. The UCI Machine Learning Repository, 2024.

[27] J. Liu, S. Wright, C. Re, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *International Conference on Machine Learning*, 32:469–477, 2014.

[28] S. Liu and L. N. Vicente. Convergence rates of the stochastic alternating algorithm for bi-objective optimization. *J. Optim. Theory Appl.*, 198:165–186, 2023.

[29] S. Liu and L. N. Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. *Ann. Oper. Res.*, 339: 1119–1148, 2024.

[30] Z. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.*, 72:7–35, 1992.

[31] R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. *Struct. Multidiscip. Optim.*, 26:369–395, 2004.

[32] K. Miettinen. *Nonlinear Multiobjective Optimization*, volume 12. Springer Science & Business Media, New York, 2012.

[33] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22:341–362, 2012.

[34] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4:1–17, 1964.

[35] M. Quentin, P. Fabrice, and J. A. Désidéri. A stochastic multiple gradient descent algorithm. *European J. Oper. Res.*, 271:808 – 817, 2018.

[36] M. Razaviyayn, M. Hong, and Z. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM J. Optim.*, 23:1126–1153, 2013.

[37] J. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, and X. Parra. Human Activity Recognition Using Smartphones. UCI Machine Learning Repository, 2013.

[38] H. M. Shi, S. Tu, Y. Xu, and W. Yin. A primer on coordinate descent algorithms. *arXiv preprint arXiv:1610.00040*, 2016.

[39] Z. Wen, D. Goldfarb, and K. Scheinberg. Block coordinate descent methods for semidefinite programming. *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 533–564, 2012.

[40] S. J. Wright. Coordinate descent algorithms. *Math. Program.*, 151:3–34, 2015.

[41] Y. Xu and W. Yin. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM J. Optim.*, 25:1686–1716, 2015.

[42] Y. Yang, M. Pesavento, Z. Luo, and B. Ottersten. Inexact block coordinate descent algorithms for nonsmooth nonconvex optimization. *IEEE Transactions on Signal Processing*, 68:947–961, 2019.

[43] Z. Yu and D. W. Ho. Zeroth-order stochastic block coordinate type methods for nonconvex optimization. *arXiv preprint arXiv:1906.05527*, 2019.

[44] S. Zhang. Beijing Multi-Site Air-Quality Data. UCI Machine Learning Repository, 2019.

[45] S. Zhou, W. Zhang, J. Jiang, W. Zhong, J. Gu, and W. Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond. *Advances in Neural Information Processing Systems*, 35:38103–38115, 2022.