Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 22–42

CONVERGENCE RATES OF THE STOCHASTIC ALTERNATING ALGORITHM FOR BI-OBJECTIVE OPTIMIZATION

S. LIU AND L. N. VICENTE

Abstract: Stochastic alternating algorithms for bi-objective optimization are considered when optimizing two conflicting functions for which optimization steps have to be applied separately for each function. Such algorithms consist of applying a certain number of steps of gradient or subgradient descent on each single objective at each iteration. In this paper, we show that stochastic alternating algorithms achieve a sublinear convergence rate of $\mathcal{O}(1/T)$, under strong convexity, for the determination of a minimizer of a weighted-sum of the two functions, parameterized by the number of steps applied on each of them. An extension to the convex case is presented for which the rate weakens to $\mathcal{O}(1/\sqrt{T})$. These rates are valid also in the non-smooth case. Importantly, by varying the proportion of steps applied to each function, one can determine an approximation to the Pareto front.

Keywords: Multi-Objective Optimization, Pareto Front, Stochastic Optimization, Alternating Optimization.
Math. Subject Classification (2010): 90C15, 90C25, 90C29.

1. Introduction

Numerous real-world application scenarios involve several potentially conflicting objectives, which need to be considered simultaneously. Such type of optimization problems is referred to as multi-objective optimization (MOO). One can find various applications spanning across applied engineering, operations management, finance, economics and social sciences, agriculture, green logistics, and health systems [26]. As these objectives are usually competing among each other, it is not possible to find a single solution which is optimal with respect to all objectives. The goal of MOO is to find a set of equally good solutions known as Pareto optimal solutions or efficient points. Roughly speaking for each Pareto optimal solution, there is no other point in the feasible region leading to a simultaneous improvement in all objectives. The determination of the set of Pareto optimal solutions helps decision makers to define the best trade-offs among the several competing criteria.

Received December 6, 2022.
Support for LNV was partially provided by the Centre for Mathematics of the University of Coimbra under grant FCT/MCTES UIDB/MAT/00324/2020.

In this paper, we consider a continuous bi-objective optimization problem

$$\min \ F(x) = (f^a(x), f^b(x))^\top, \quad x \in \mathcal{X}, \tag{1}$$

where $\mathcal{X}$ is a nonempty bounded closed convex region. We say that the multi-objective problem is smooth if all objective functions f^i are continuously differentiable. In the convergence theory of our algorithms, we will assume that both functions are convex. Furthermore, both objectives involve randomness in its parameters, in which case we expect that a true gradient or subgradient is not available or too expensive to compute. Instead, one can access a unbiased estimator $g^i(x, \xi), i \in \{a, b\}$, of a true gradient or subgradient of f^i, where ξ denotes some random variable. The optimality in MOO is defined via the concept of Pareto dominance. More precisely, for the general bi-objective problem (1), we say that x (weakly) dominates y if $f^i(x) \leq f^i(y), \forall i \in \{a, b\}$, and $F(x) \neq F(y)$. A point $x \in \mathcal{X}$ is called a (strictly) non-dominated solution or Pareto optimal solution if it is not (weakly) dominated by any other point in $\mathcal{X}$. The set of non-dominated solutions denoted by $\mathcal{P}$ forms the so-called Pareto front $F(\mathcal{P}) = \{F(x) : x \in \mathcal{P}\}$.

1.1. A quick overview of multi-objective optimization methods. Depending on whether the decision maker's preference is explicitly involved during the optimization process, two categories of approaches, namely *a priori* and *a posteriori* methods, are mainly considered in the existing multi-objective optimization literature.

The *a priori* methods incorporate the preference to reduce the multi-objective problem into a single-objective one, which can then be tackled by classical single-objective optimization algorithms. Such preferences could be expressed via weights put on different objectives, desired upper bounds of some objectives, or utility functions, corresponding to the weight-sum method [25], the ϵ-constrained method [27], and the utility function method [10, 30]. In the weighted-sum method, one assigns each objective a non-negative weight λ_i and then optimizes the weighted-sum function $S(x, \lambda) = \sum_i \lambda_i f^i(x)$. By varying the weights in a convex linear combination one is guaranteed to find all Pareto solutions when the functions are convex. The ϵ-constrained method consists of optimizing one objective using given upper bounds, i.e., $\min \ f^i(x)$ subject to $f^j(x) \leq \epsilon_j, \forall j \neq i$. The utility function method is a more general approach in this category. A utility function U is a real function quantifying the overall preference among the

competing objectives [10]. Given two candidate solutions x and y, then the decision maker prefers x to y if $U(F(x)) > U(F(y))$.

As for the *a posteriori* methods, the solution process aims at producing all the Pareto optimal solutions or a representative subset of the Pareto optimal solutions, out of which a single "best" solution is then selected according to the decision maker's preferences. A technique often used is to maintain and update a list of candidates iteratively based on a *heuristic* or *rigorous* mechanism. The class of (*a posteriori*) evolutionary multi-objective optimization algorithms [3, 13] using *metaheuristics* has been a very popular research topic since 1990's. Moreover, in the past two decades, many *rigorous* descent methods [24] have been developed, namely multi-objective versions of the zero-order method [15], the steepest descent [16, 17, 18, 20, 22, 23], the subgradient methods [5, 14], the proximal gradient [6, 8, 38, 39], the conjugate gradient methods [32], the trust-region methods [35, 40], and the Newton methods [19, 21, 33, 34]. These methods attempt to decrease the individual objective values simultaneously in some sense, and the convergence of the generated iterates to a Pareto first-order stationary point is rigorously established under reasonable assumptions. Such a common descent direction for all the objectives is usually computed at each iteration by solving a certain sub-problem that takes into account the gradient and/or Hessian information of all the objectives.

Stochastic multi-objective optimization (SMOO) deals with the case where one or multiple objective functions involve uncertainty or noisy data. One usually formulates the objective functions in the form of expectation with respect to some random variables and approximate them by the sample average approximation (SAA) technique. Two main approaches [1, 2, 11] for solving the resulting SMOO problems are the *multi-objective* methods and the *stochastic* methods. While the former reduces the SMOO problem into a deterministic MOO problem, the later converts the original problem into a single-objective stochastic problem by scalarization.

The alternating optimization algorithms considered in this paper belong to the *a priori* category. Instead of decreasing individual objectives simultaneously at each iteration, such algorithms separately optimize them, one after the other, and we will see in this paper that these algorithms eventually optimize a certain convex linear combination of the two functions. In terms

of how the stochasticity is handled, our paper uses the stochastic approximation (SA) technique (stochastic gradient-based methods) which has been recently considered for SMOO by [29, 36].

An example of stochastic alternating algorithms for bi-objective optimization arises in fair clustering, where one tries to capture a trade-off between clustering cost and balance among different demographic groups defined by sensitive attributes (see [12, 7, 37, 41]). A perfectly balanced solution requires that each cluster has exactly the same proportions of each group, but such solution is not necessarily the one that minimizes the clustering cost. Fair clustering can be formulated as a bi-objective problem (see [28, Problem (2)]), involving two non-convex and non-smooth functions. The problem is originally discrete but a continuous relaxation of the binary assignment variables can be considered. In [28], we introduced a stochastic alternating balance fair k-means (SAfairKM) approach for fair clustering. In this application we could not find a reasonable way to optimize a convex linear combination of the two functions. Rather, we were faced with a situation where we knew how to optimize them separately.

1.2. Contribution of this paper. In this paper, we establish rates of convergence for stochastic alternating bi-objective optimization algorithms, both in the smooth and non-smooth cases, and under both simple and strong convexity. It is shown that the algorithms exhibit sublinear convergence rates of $\mathcal{O}(1/T)$ when determining a Pareto solution, under strong convexity and classical assumptions of gradients or subgradients. Moreover, we show that downgrading strong convexity to convexity results in a degradation of the convergence rate from $\mathcal{O}(1/T)$ to $\mathcal{O}(1/\sqrt{T})$. It is remarkable that we recover the same convergence rates of the corresponding single-objective stochastic gradient/subgradient methods. Our theory evolves around an ingenious application of the Intermediate Value Theorem (IVT) to aggregate the steps applied when separately optimizing each function. The target Pareto solution is defined by selecting a convex span of the number of steps applied to each function. By varying the weight in this convex span, one can determine the whole Pareto front in the convex case. It is the application of the IVT that brings the convex span of the effort into the convergence rates.

Methods like weighted-sum or ϵ-constrained would be the reasonable choices in most practical applications whenever it is possible to optimize a convex linear combination of the objective functions or optimize one of them subject

[27] 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.
[28] S. Liu and L. N. Vicente. A stochastic alternating balance k-means algorithm for fair clustering. *Lecture Notes in Computer Science*, 2023, to appear.
[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.*, 2023, to appear.
[30] K. Miettinen. *Nonlinear Multiobjective Optimization*, volume 12. Springer Science & Business Media, New York, 2012.
[31] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19:1574–1609, 2009.
[32] L. R. Lucambio Pérez and L. F. Prudente. Nonlinear conjugate gradient methods for vector optimization. *SIAM J. Optim.*, 28:2690–2720, 2018.
[33] Z. Povalej. Quasi-Newton's method for multiobjective optimization. *Comput. Optim. Appl.*, 255:765–777, 2014.
[34] S. Qu, M. Goh, and F. T. S. Chan. Quasi-Newton methods for solving multiobjective optimization. *Oper. Res. Lett.*, 39:397–399, 2011.
[35] S. Qu, M. Goh, and B. Liang. Trust region methods for solving multiobjective optimisation. *Optim. Methods Softw.*, 28:796–811, 2013.
[36] M. Quentin, P. Fabrice, and J. A. Désidéri. A stochastic multiple gradient descent algorithm. *European J. Oper. Res.*, 271:808 – 817, 2018.
[37] M. Schmidt, C. Schwiegelshohn, and C. Sohler. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*, pages 232–251. Springer, 2019.
[38] H. Tanabe, E. H. Fukuda, and N. Yamashita. Proximal gradient methods for multiobjective optimization and their applications. *Comput. Optim. Appl.*, 72:339–361, 2019.
[39] H. Tanabe, E. H. Fukuda, and N. Yamashita. Convergence rates analysis of a multiobjective proximal gradient method. *Optim. Lett.*, 2023, to appear.
[40] K. D. Villacorta, P. R. Oliveira, and A. Soubeyran. A trust-region method for unconstrained multiobjective problems with applications in satisficing processes. *J. Optim. Theory Appl.*, 160:865–889, 2014.
[41] I. M. Ziko, E. Granger, J. Yuan, and I. B. Ayed. Variational fair clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11202–11209, 2021.

S. Liu
Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015-1582, USA (sul217@lehigh.edu).

L. N. Vicente
Department of Industrial and Systems Engineering, Lehigh University, 200 West Packer Avenue, Bethlehem, PA 18015-1582, USA and Centre for Mathematics of the University of Coimbra (CMUC), Portugal (lnv@mat.uc.pt).

[4] A. Beck. *First-Order Methods in Optimization*. SIAM, Philadelphia, 2017.
[5] G. C. Bento and J. X. Cruz Neto. A subgradient method for multiobjective optimization on riemannian manifolds. *J. Optim. Theory Appl.*, 159:125–137, 2013.
[6] G. C. Bento, J. X. Cruz Neto, and A. Soubeyran. A proximal point-type method for multicriteria optimization. *Set-Valued and Variational Analysis*, 22:557–573, 2014.
[7] S. Bera, D. Chakrabarty, N. Flores, and M. Negahbani. Fair algorithms for clustering. In *NIPS*, pages 4954–4965, 2019.
[8] H. Bonnel, A. N. Iusem, and B. F. Svaiter. Proximal methods in vector optimization. *SIAM J. Optim.*, 15:953–970, 2005.
[9] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60:223–311, 2018.
[10] E. K. Browning and M. A. Zupan. *Microeconomics: Theory and Applications*. Wiley, Hoboken, NJ, 2020.
[11] R. Caballero, E. Cerdá, M. Munoz, and L. Rey. Stochastic approach versus multiobjective approach for obtaining efficient solutions in stochastic multiobjective programming problems. *European J. Oper. Res.*, 158:633–648, 2004.
[12] F. Chierichetti, R. Kuma, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *NIPS*, pages 5029–5037, 2017.
[13] C. C. Coello. Evolutionary multi-objective optimization: A historical view of the field. *IEEE Computational Intelligence Nagazine*, 1:28–36, 2006.
[14] J. Y. Bello Cruz. A subgradient method for vector optimization problems. *SIAM J. Optim.*, 23:2169–2182, 2013.
[15] 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.
[16] J. A. Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *C. R. Math. Acad. Sci. Paris*, 350:313–318, 2012.
[17] J. A. Désidéri. Multiple-gradient descent algorithm for Pareto-front identification. In *Modeling, Simulation and Optimization for Science and Technology*, pages 41–58. Springer, Dordrecht, 2014.
[18] L. G. Drummond and A. N. Iusem. A projected gradient method for vector optimization problems. *Comput. Optim. Appl.*, 28:5–29, 2004.
[19] L. G. Drummond, F. M. P. Raupp, and B. F. Svaiter. A quadratically convergent Newton method for vector optimization. *Optimization*, 63:661–677, 2014.
[20] L. G. Drummond and B. F. Svaiter. A steepest descent method for vector optimization. *J. Comput. Appl. Math.*, 175:395–414, 2005.
[21] J. Fliege, L. G. Drummond, and B. F. Svaiter. Newton's method for multiobjective optimization. *SIAM J. Optim.*, 20:602–626, 2009.
[22] J. Fliege and B. F. Svaiter. Steepest descent methods for multicriteria optimization. *Math. Methods Oper. Res.*, 51:479–494, 2000.
[23] E. H. Fukuda and L. M. G. Drummond. Inexact projected gradient method for vector optimization. *Comput. Optim. Appl.*, 54:473–493, 2013.
[24] E. H. Fukuda and L. M. G. Drummond. A survey on multiobjective descent methods. *Pesquisa Operacional*, 34:585–620, 2014.
[25] S. Gass and T. Saaty. The computational algorithm for the parametric objective function. *Nav. Res. Logist. Q.*, 2:39–45, 1955.
[26] N. Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1), 2018.

to the others be within a certain range. Our alternating methodology would be chosen when optimizing such a combination is not possible. That is the case when the functions are defined by some form of black-box for which there is a legacy or binary code to carry out the optimization. That is also the case when there is some form of function-dependent descent step [28]. Our methodology would also be chosen when optimizing such combination is possible but not desirable, which may happen for instance when the functions exhibit a scale that is very different and hard to balance.

2. Alternating bi-objective optimization algorithms

The key idea of the alternating bi-objective optimization for solving (1) consists of iteratively taking n_a gradient (or subgradient) descent steps for the first objective and then n_b gradient (or subgradient) descent steps for the second objective. For simplicity, we denote such $n_a + n_b$ steps as one single iteration of the algorithm. The stochastic alternating bi-objective gradient or subgradient algorithm (SA2GD) is formally described in Algorithm 1. At every iteration t, starting at $x_t = y^a_{0,t}$, the algorithm computes two sequences of intermediate iterates, $\{y^a_{r,t}\}_{r=1}^{n_a}$ and $\{y^b_{r,t}\}_{r=1}^{n_b}$. Let $P_{\mathcal{X}}$ denote the orthogonal projection operator that projects the new iterate back to the feasible region. Such an operator is well-defined because we assume a compact and convex feasible region.

Algorithm 1 Stochastic alternating bi-objective gradient (or subgradient) algorithm (SA2GD)

Input: Initial point $x_0 = y^a_{0,0}$ and a step size sequence $\{\alpha_t\}$.
Output: A likely non-dominated or Pareto solution $x_T = y^a_{0,T}$.
for $t = 0, 1, \ldots, T-1$ **do**
  for $r = 0, \ldots, n_a - 1$ **do**
    Generate a stochastic gradient $g^a(y^a_{r,t}, \xi^r_t)$.
    Update $y^a_{r+1,t} = y^a_{r,t} - \alpha_t g^a(y^a_{r,t}, \xi^r_t)$.
  Set $y^b_{0,t} = y^a_{n_a,t}$.
  for $r = 0, \ldots, n_b - 1$ **do**
    Generate a stochastic gradient $g^b(y^b_{r,t}, \xi^{n_a+r}_t)$.
    Update $y^b_{r+1,t} = y^b_{r,t} - \alpha_t g^b(y^b_{r,t}, \xi^{n_a+r}_t)$.
  Set $x_{t+1} = y^a_{0,t+1} = P_{\mathcal{X}}(y^b_{n_b,t})$.

Similarly, in the case where both objectives are non-smooth but sub-differentiable, one can develop a counterpart stochastic alternating bi-objective subgradient method by simply replacing stochastic gradients used in Lines 5 and 9 by stochastic subgradients.

The stochastic gradient algorithm for SMOO (stochastic multi-gradient) computes one point in the Pareto front (following a steepest descent principle) [29, 36]. The alternating one also computes just one point. The difference is that the stochastic multi-gradient method is not parameterized or scalarized, and so the only form to use it to compute the whole Pareto front is to incorporate it in a non-dominated list updating mechanism or some form of hotstarting or randomization. The alternating one works like the weighted-sum and ϵ-constrained methods, and it involves parameters such that when varied allow us to capture the whole Pareto front (the optimization effort quantified by n_a, n_b).

3. Convergence analysis

Our convergence analysis starts with an easier case where both objectives are assumed smooth and strongly convex. When we remove smoothness from the assumptions, the same convergence rate is maintained. Furthermore, in both cases, relaxing strong convexity to convexity leads to degradation of convergence rates from $\mathcal{O}(1/T)$ to $\mathcal{O}(1/\sqrt{T})$.

3.1. The smooth and strongly convex case. We first assume that both objective functions are smooth and strongly convex and establish a $\mathcal{O}(1/T)$ sublinear convergence rate for the stochastic alternating bi-objective gradient descent (SA2GD) algorithm. At every iteration t, starting at $x_t = y^a_{0,t}$, SA2GD computes two sequences of intermediate iterates, $\{y^a_{r,t}\}_{r=1}^{n_a}$ and $\{y^b_{r,t}\}_{r=1}^{n_b}$. We first make a formal assumption on the boundedness of the feasible region.

Assumption 3.1. ***(Bounded feasible region)*** *The feasible region $\mathcal{X}$ is compact, and in particular there exists a positive constant Θ such that*

$$\max_{x,\bar{x}\in\mathcal{X}} \|x - \bar{x}\| \leq \Theta < \infty.$$

Then, we formalize the classical smoothness assumption of Lipschitz continuity of the gradients.

By varying the effort put in separately optimizing each objective within a convex linear combination, one can capture a trade-off among the multiple objectives.

In practice, many machine learning models lead to non-convex optimization problems. Non-convexity is present in the original bi-objective fair k-means clustering problem mentioned in the Introduction. A meaningful convergence guarantee for the non-convex case is certainly a topic for future research.

Appendix A. Proposition using Intermediate Value Theorem

Based on the Intermediate Value Theorem, we derive the following proposition for the purpose of convergence rate analysis of the SA2GD algorithm.

Proposition A.1. *Given a continuous real function $\phi(x) : \mathbb{R}^n \to \mathbb{R}$ and a set of points $\{x_j\}_{j=1}^m$, there exists $w \in \mathbb{R}^n$ such that*

$$m\phi(w) = \sum_{j=1}^{m} \phi(x_j),$$

where $w = \sum_{j=1}^m \mu_j x_j$, with $\sum_{j=1}^m \mu_j = 1, \mu_j \geq 0$, $i = 1, \ldots, m$, is a convex linear combination of $\{x_j\}_{j=1}^m$.

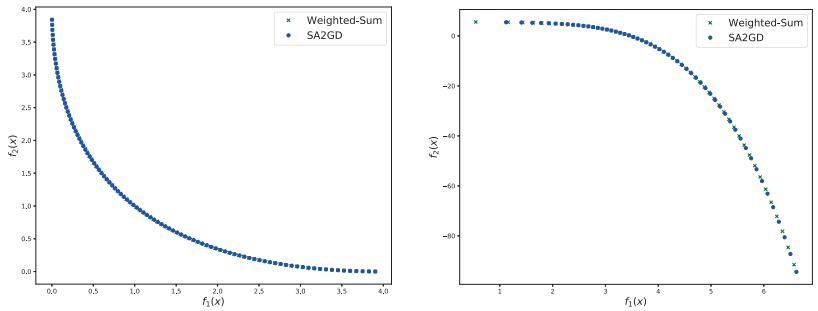
Proof: The proposition is obtained by consecutively applying the Intermediate Value Theorem to $\phi(x)$. First, for the pair of points x_1 and x_2, there exists a point $w_{12} = \mu_{12}x_1 + (1-\mu_{12})x_2, \mu_{12} \in [0,1]$, such that $\phi(w_{12}) = (\phi(x_1) + \phi(x_2))/2$ according to the Intermediate Value Theorem, which implies that $\sum_{j=1}^m \phi(x_j) = 2\phi(w_{12}) + \sum_{j=3}^m \phi(x_j)$. Then, there exists $w_{13} = \mu_{13}w_{12} + (1-\mu_{13})x_3, \mu_{13} \geq 0$, such that $\phi(w_{13}) = (2\phi(w_{12}) + \phi(x_3))/3$ holds given that the average function value $(2\phi(w_{12}) + \phi(x_3))/3$ lies between $\phi(w_{12})$ and $\phi(x_3)$. Notice that w_{13} can also be written as convex linear combination of $\{x_1, x_2, x_3\}$. The proof is concluded by continuing this process until x_m is reached. ■

References

[1] F. B. Abdelaziz. *L'Efficacité en Programmation Multi-Objectifs Stochastique*. PhD thesis, Université de Laval, Québec, 1992.

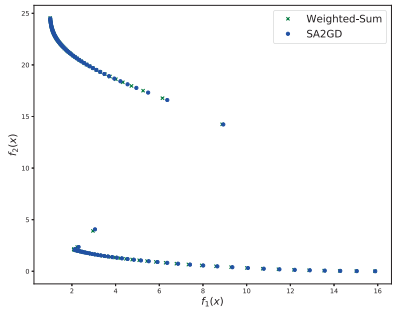
[2] F. B. Abdelaziz. Solution approaches for the multiobjective stochastic programming. *European J. Oper. Res.*, 216:1–16, 2012.

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

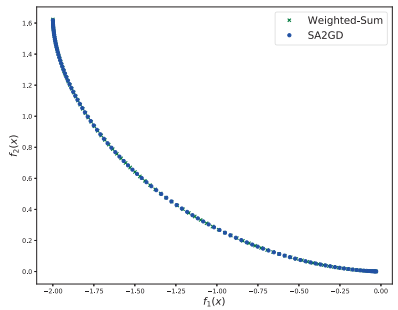


(A) Problem MOP1.

(B) Problem IM1.



(C) Problem MOP3.



(D) Problem FAR1.

FIGURE 1. An illustration of the scalarization by optimization effort of SA2GD when compared to the weighted-sum method.

We ingeniously applied the Intermediate Value Theorem to bundle all steps taken on each of the functions, a process that led to an implicit scalarization $(n_a f^a + n_b f^b)/(n_a + n_b)$ of the two functions. It is however possible, following a similar proof logic, to show that the convergence rates are maintained regardless of the order according to which the gradient (subgradient) steps for the two objectives are executed. In fact, we can even, within a single iteration, randomly select the n_a positions where steps are applied to f^a from the $n_a + n_b$ total number of steps without impacting the rate of convergence.

Furthermore, the proposed stochastic alternating optimization framework and theory could be generalized to more than two conflicting objectives.

Assumption 3.2. ***(Lipschitz continuous gradients)*** *The individual true gradients are Lipschitz continuous with Lipschitz constants $L_i > 0, i \in \{a, b\}$, i.e.,*

$$\|\nabla f^i(x) - \nabla f^i(\bar{x})\| \leq L_i \|x - \bar{x}\|, \quad \forall (x, \bar{x}) \in \mathcal{X} \times \mathcal{X}.$$

Note that any continuous function attains a maximum on a compact set, and thus one can assume the existence of positive constants M^i_∇, $i \in \{a, b\}$, such that

$$\|\nabla f^i(x)\|^2 \leq M^i_\nabla, \quad \forall x \in \mathcal{X}. \tag{2}$$

In addition to Assumption 3.2, we impose strong convexity in both objective functions.

Assumption 3.3. ***(Strong convexity)*** *For all $i = \{a, b\}$, there exists a scalar $c_i > 0$ such that*

$$f^i(\bar{x}) \geq f^i(x) + \nabla f^i(x)^\top (\bar{x} - x) + \frac{c_i}{2} \|\bar{x} - x\|^2, \quad \forall (x, \bar{x}) \in \mathcal{X} \times \mathcal{X}.$$

Note that based on the above assumption, the weighted-sum function $S(x, \lambda) = \lambda f^a(x) + (1 - \lambda) f^b(x), \lambda \in [0, 1]$, is also strongly convex with constant $c = \min(c_a, c_b)$. Given the individual stochastic gradients $g^i(x_t, \xi_t), \forall i \in \{a, b\}$, generated with random variable ξ_t, we use $\mathbb{E}_{\xi_t}[\cdot]$ to denote the conditional expectation taken with respect to ξ_t. We also impose the following two classical assumptions of stochastic gradients.

Assumption 3.4. *For both objective functions $i \in \{a, b\}$, and all iterates $t \in \mathbb{N}$, the individual stochastic gradients $g^i(x_t, \xi_t)$ satisfy the following:*

(a) ***(Unbiased gradient estimation)*** $\mathbb{E}_{\xi_t}[g^i(x_t, \xi_t)] = \nabla f^i(x_t), \ \forall i \in \{a, b\}$.

(b) ***(Bound on the second moment)*** *There exist positive scalars $G_i > 0$ and $\bar{G}_i > 0$ such that*

$$\mathbb{E}_{\xi_t}[g^i(x_t, \xi_t)\|^2] \leq G_i + \bar{G}_i \|\nabla f^i(x_t)\|^2, \quad \forall i \in \{a, b\}.$$

The above assumptions are the commonly used ones in classical stochastic gradient methods [9], basically assuming reasonable bounds on the expectation and variance of the individual stochastic gradients.

Our main result stated below shows that SA2GD drives the expected optimality gap of the weighted-sum function $S(\cdot, \lambda(n_a, n_b))$ to zero at a sublinear rate of $1/T$, where $\lambda(n_a, n_b) = n_a/(n_a + n_b)$, when using a decaying step size

sequence. By varying n_a and n_b in $\{1,\ldots,n_{\text{total}}\}$, such that $n_{\text{total}} = n_a + n_b$, one can capture the entire trade-off between f^a and f^b.

Theorem 3.1. ***(Sublinear convergence rate of SA2GD in the smooth and strongly convex case)*** *Let Assumptions 3.1-3.4 hold and x_* be the unique minimizer of the weighted function $S(\cdot,\lambda_*)$ in $\mathcal{X}$, where $\lambda_* = \lambda(n_a,n_b) = n_a/(n_a+n_b)$. Choosing a diminishing step size sequence $\alpha_t = \frac{2}{c(t+1)(n_a+n_b)}$, the sequence of iterates generated by the SA2GD algorithm satisfies*

$$\min_{t=1,\ldots,T} \mathbb{E}[S(x_t,\lambda_*)] - S(x_*,\lambda_*) \;\leq\; \frac{4}{c(T+1)}\left(\hat{G}^2 + L\Theta\hat{G}\right).$$

where $\hat{G} = \sqrt{G+\bar{G}M_\nabla}$, $L = \max(L_a,L_b)$, $G = \max(G_a,G_b)$, $\bar{G} = \max(\bar{G}_a,\bar{G}_b)$, and $M_\nabla = \max(M_\nabla^a, M_\nabla^b)$.

Proof: The proof is divided in three parts for better organization and understanding. In the first part, one obtains an upper bound on the norm of the iterates, $\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2]$. Strong convexity of the weighted-sum function is applied in the second part. The third part concludes the proof using standard arguments. For simplicity, we let $z_t = y_{0,t}^b = y_{n_a,t}^a$ be the intermediate point at each iteration.

Part I: Bound on the iterates error using the Intermediate Value Theorem. At any iteration t, the sequence of stochastic gradients is computed from drawing the sequence of random variables $\xi_t = \{\xi_t^0,\ldots,\xi_t^{n_a+n_b-1}\}$. We have

$$\begin{aligned} x_{t+1}-x_* &= P_{\mathcal{X}}\left(z_t - \alpha_t \sum_{r=0}^{n_b-1} g^b(y_{r,t}^b,\xi_t^r)\right) - x_*, \\ &= P_{\mathcal{X}}\left(x_t - \alpha_t \sum_{r=0}^{n_a-1} g^a(y_{r,t}^a,\xi_t^r) - \alpha_t \sum_{r=0}^{n_b-1} g^b(y_{r,t}^b,\xi_t^{n_a+r})\right) - x_*. \end{aligned}$$

Since the sequence ξ_t is drawn independently, using Assumption 3.4 (a) one has

$$\begin{aligned} \mathbb{E}_{\xi_t}[g^a(y_{r,t}^a,\xi_t^r)] &= \mathbb{E}_{\xi_t^0,\ldots,\xi_t^{r-1}}[\mathbb{E}_{\xi_t^r}[g^a(y_{r,t}^a,\xi_t^r)]] \;=\; \mathbb{E}_{\xi_t^0,\ldots,\xi_t^{r-1}}[\nabla f^a(y_{r,t}^a)] \\ &= \mathbb{E}_{\xi_t}[\nabla f^a(y_{r,t}^a)], \end{aligned}$$

where the last equality holds due to the independence between $y_{r,t}^a$ and $\{\xi_t^r,\ldots,\xi_t^{n_a+n_b-1}\}$. Similarly, we have $\mathbb{E}_{\xi_t}[g^b(y_{r,t}^b,\xi_t^{n_a+r})] = \mathbb{E}_{\xi_t}[\nabla f^b(y_{r,t}^b)]$.

As a final remark, the given upper bounds in Theorems 3.1-3.2 and Corollaries 3.2-3.3 also hold for the aggregated iterate $\bar{x}_T = \sum_{t=1}^T t x_t / \sum_{t=1}^T t$ in the strongly convex case and $\bar{x}_T = \sum_{t=1}^T x_t/T$ in the convex case. Such claims follow from the following Jensen's inequalities

$$S\left(\tfrac{\sum_{t=1}^T t x_t}{\sum_{t=1}^T t},\lambda_*\right) \;\leq\; \tfrac{\sum_{t=1}^T t S(x_t,\lambda_*)}{\sum_{t=1}^T t} \quad \text{and} \quad S\left(\tfrac{\sum_{t=1}^T x_t}{T},\lambda_*\right) \;\leq\; \tfrac{\sum_{t=1}^T S(x_t,\lambda_*)}{T}.$$

4. A numerical experiment to illustrate the scalarization by optimization effort

We now present an illustration of the behavior of the SA2GD method in what regards its ability to determine the whole Pareto front when applied multiple times over a discretization of the optimization effort scalars n_a and n_b. For simplicity we have applied SA2GD to deterministic problems, thus using full batch gradients. Four deterministic bi-objective problems were selected from [15] involving only simple bound constraints. We will compare the scalarization effort of SA2GD to the one of the weighted-sum method, where each scalarized problem $\min_{x\in\mathcal{X}} \lambda f^a(x) + (1-\lambda) f^b(x)$, $\lambda \in (0,1)$, is solved by the same gradient descent methodology. For both approaches, we use a fixed step size 10^{-3} at each iteration. The starting point for each run is randomly generated within the feasible region.

Figure 1 illustrates the approximated Pareto fronts obtained from both the SA2GD algorithm and the weight-sum algorithm for the selected problems. SA2GD was ran 201 times with $n_a + n_b = 200$ and $n_a \in \{0,\ldots,200\}$. Weighted-sum was also run 201 times with $\lambda \in \{0, 1/200,\ldots,1\}$. In both cases, we stopped each algorithmic run after 300 iterations. Regardless of the shape of Pareto front (convex, concave, or disconnected), one observes that SA2GD performs similarly to the weight-sum approach in terms of capturing well-spread Pareto fronts from scalarization.

5. Conclusions

We have shown that stochastic alternating bi-objective gradient or subgradient descent is convergent, achieving convergence rates aligned with what is known for single-objective gradient (or subgradient) methods. The rates are $\mathcal{O}(1/\sqrt{T})$ for convex functions and $\mathcal{O}(1/T)$ for strongly convex functions, and they hold for both smooth and non-smooth functions.

The analysis considered the case where all gradient (or subgradient) steps are applied first in a block to one function and then to the other also in block.

derive the bound of the iterate error

$$\begin{aligned}\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2] \;\le\; & \|x_t-x_*\|^2+2\alpha_t^2(n_a+n_b)^2\tilde{L}^2\\ & -2\alpha_t(n_af^a(x_t)+n_bf^b(x_t)-(n_af^a(x_*)+n_bf^b(x_*)))\\ & +2\alpha_t\hat{L}\mathbb{E}_{\xi_t}[n_a\|x_t-w_t^a\|+n_b\|x_t-w_t^b\|]\\ & +2\alpha_t\tilde{L}(\sum_{r=0}^{n_a-1}\mathbb{E}_{\xi_t}[\|x_t-y_{r,t}^a\|]+\sum_{r=0}^{n_b-1}\mathbb{E}_{\xi_t}[\|x_t-y_{r,t}^b\|]).\end{aligned} \tag{26}$$

Plugging (25) into (26) yields

$$\begin{aligned}\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2] \;\le\; & \|x_t-x_*\|^2+\alpha_t^2\hat{M}\\ & -2\alpha_t(n_af^a(x_t)+n_bf^b(x_t)-(n_af^a(x_*)+n_bf^b(x_*))),\end{aligned}$$

where $\hat{M}=(4\tilde{L}^2+2\tilde{L}\hat{L})(n_a+n_b)^2$. The proof is completed by applying the same standard arguments as in Theorem 5.3 of [29].

■

The non-smooth and convex case sheds light on the convergence behavior of the SAfairKM algorithm for the original non-convex and non-smooth bi-objective optimization problem [28].

For completeness, we conclude this section by the following corollary stating the convergence rate for the smooth and convex case without proof. The proof is also derived by getting rid of the quadratic term $-c\alpha_t(n_a+n_b)\|x_t-x_*\|^2$ in (17) and then applying the standard arguments as Theorem 5.3 in [29].

Assumption 3.9. ***(Convexity in the smooth case)*** *For both objective functions $i\in\{a,b\}$, we have*

$$f^i(\bar{x}) \;\ge\; f^i(x)+\nabla f^i(x)^\top(\bar{x}-x),\quad \forall(x,\bar{x})\in\mathcal{X}\times\mathcal{X}.$$

Corollary 3.3. ***(Sublinear convergence rate of SA2GD in the smooth and convex case)*** *Let Assumptions 3.1–3.2, 3.4, and 3.9 hold and x_* be a minimizer of the weighted function $S(\cdot,\lambda_*)$ in $\mathcal{X}$, where $\lambda_*=\lambda(n_a,n_b)=n_a/(n_a+n_b)$. Choosing a diminishing step size sequence $\alpha_t=\frac{\bar{\alpha}}{\sqrt{t}(n_a+n_b)}$, where $\bar{\alpha}$ is any positive constant, the sequence of iterates generated by the SA2GD algorithm satisfies*

$$\min_{t=1,\ldots,T}\mathbb{E}[S(x_t,\lambda_*)]-\mathbb{E}[S(x_*,\lambda_*)] \;\le\; \frac{\frac{\Theta^2}{2\bar{\alpha}}+2\bar{\alpha}\hat{G}^2+2\bar{\alpha}L\Theta\hat{G}}{\sqrt{T}}.$$

Then, taking square norms and expectations over the random variables ξ_t on both sides yields

$$\begin{aligned}\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2] \;\le\; & \|x_t-x_*\|^2\\ & +\alpha_t^2\mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_a-1}g^a(y_{r,t}^a,\xi_t^r)+\sum_{r=0}^{n_b-1}g^b(y_{r,t}^b,\xi_t^{n_a+r})\|^2]\\ & -\mathbb{E}_{\xi_t}[2\alpha_t(x_t-x_*)^\top(\sum_{r=0}^{n_a-1}\nabla f^a(y_{r,t}^a)+\sum_{r=0}^{n_b-1}\nabla f^b(y_{r,t}^b))],\end{aligned} \tag{3}$$

which holds by the non-expansiveness property of the orthogonal projection operator.

We now claim, by applying a version of the Intermediate Value Theorem given in Proposition A.1, that the last term of the right-hand side in (3) can be written as $-2\alpha_t(x_t-x_*)^\top\mathbb{E}_{\xi_t}[n_a\nabla f^a(w_t^a)+n_b\nabla f^b(w_t^b)]$ for some w_t^a and w_t^b. In fact, we apply Proposition A.1 to the real continuous function $\phi^i(y)=-2\alpha_t(x_t-x_*)^\top\nabla f^i(y)$, from which we then know that w_t^i is a convex linear combination of a sequence of points $\{y_{r,t}^i\}_{r=0}^{n_i-1}$ for both $i\in\{a,b\}$.

Using a combination of (2) and Assumption 3.4 (b), the bound for the second moment of the stochastic gradients is given by

$$\mathbb{E}_{\xi_t}[\|g^i(y_{r,t}^i,\xi_t^r)\|^2] \;\le\; G+\bar{G}M_\nabla,\quad \mathbb{E}_{\xi_t}[\|g^i(y_{r,t}^i,\xi_t^r)\|] \;\le\; \sqrt{G+\bar{G}M_\nabla}, \tag{4}$$

where the second bound results from applying Jensen's inequality to the first one. Hence, the second term in the right-hand side of (3) can be bounded by

$$\begin{aligned}& \mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_a-1}g^a(y_{r,t}^a,\xi_t^r)+\sum_{r=0}^{n_b-1}g^b(y_{r,t}^b,\xi_t^{n_a+r})\|^2]\\ & \le\; 2\mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_a-1}g^a(y_{r,t}^a,\xi_t^r)\|^2]+2\mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_b-1}g^b(y_{r,t}^b,\xi_t^{n_a+r})\|^2]\\ & \le\; 2n_a\sum_{r=0}^{n_a-1}\mathbb{E}_{\xi_t}[\|g^a(y_{r,t}^a,\xi_t^r)\|^2]+2n_b\sum_{r=0}^{n_b-1}\mathbb{E}_{\xi_t}[\|g^b(y_{r,t}^b,\xi_t^{n_a+r})\|^2]\\ & \le\; 2(n_a+n_b)^2(G+\bar{G}M_\nabla).\end{aligned}$$

We thus arrive at

$$\begin{aligned}\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2] \;\le\; &\|x_t-x_*\|^2+2\alpha_t^2(n_a+n_b)^2(G+\bar{G}M_\nabla)\\ &-2\alpha_t(x_t-x_*)^\top\mathbb{E}_{\xi_t}[n_a\nabla f^a(w_t^a)+n_b\nabla f^b(w_t^b)].\end{aligned}\tag{5}$$

By adding and subtracting $2\alpha_t(x_t-x_*)^\top(n_a\nabla f^a(x_t)+n_b\nabla f^b(x_t))$ in the right-hand side of (5), we further rewrite it as

$$\begin{aligned}\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2] \;\le\; &\|x_t-x_*\|^2+2\alpha_t^2(n_a+n_b)^2(G+\bar{G}M_\nabla)\\ &-2\alpha_t(x_t-x_*)^\top(n_a\nabla f^a(x_t)+n_b\nabla f^b(x_t))\\ &+2\alpha_t\|x_t-x_*\|\mathbb{E}_{\xi_t}[\|n_a\nabla f^a(w_t^a)-n_a\nabla f^a(x_t)\|]\\ &+2\alpha_t\|x_t-x_*\|\mathbb{E}_{\xi_t}[\|n_b\nabla f^b(w_t^b)-n_b\nabla f^b(x_t)\|].\end{aligned}\tag{6}$$

Note that the last two terms are derived by the Cauchy–Schwarz and Jensen's inequalities.

Part II: Using strong convexity. Selecting $\lambda_*=\lambda(n_a,n_b)=n_a/(n_a+n_b)$, by the strong convexity of the weighted-sum function, one has

$$\nabla_x S(x_t,\lambda_*)^\top(x_t-x_*) \;\ge\; S(x_t,\lambda_*)-S(x_*,\lambda_*)+\frac{c}{2}\|x_t-x_*\|^2,$$

which is equivalent to

$$\begin{aligned}&(x_t-x_*)^\top(n_a\nabla f^a(x_t)+n_b\nabla f^b(x_t))\\ &\ge\; (n_a+n_b)(S(x_t,\lambda_*)-S(x_*,\lambda_*)+\frac{c}{2}\|x_t-x_*\|^2).\end{aligned}\tag{7}$$

From Assumption 3.2, we obtain a bound for the last two terms of (6) in the form

$$\begin{aligned}&2\alpha_t\|x_t-x_*\|\mathbb{E}_{\xi_t}[\|n_i\nabla f^i(w_t^i)-n_i\nabla f^i(x_t)\|]\\ &\le\; 2\alpha_t L n_i\|x_t-x_*\|\mathbb{E}_{\xi_t}[\|x_t-w_t^i\|],\forall i\in\{a,b\}.\end{aligned}\tag{8}$$

According to Proposition A.1, w_t^i is a convex linear combination of a sequence of points $\{y_{r,t}^i\}_{r=0}^{n_i-1}$ for $i\in\{a,b\}$. One can write $w_t^i=\sum_{r=0}^{n_i-1}\beta_r y_{r,t}^i$ with $\beta_r\ge 0$, $r=0,\ldots,n_i-1$, and $\sum_{r=0}^{n_i-1}\beta_r=1$. An explicit upper bound of $\|x_t-w_t^i\|$ can then be derived as follows

$$\|x_t-\sum_{r=0}^{n_i-1}\beta_r y_{r,t}^i\| \;=\; \|\sum_{r=0}^{n_i-1}\beta_r(x_t-y_{r,t}^i)\| \;\le\; \sum_{r=0}^{n_i-1}\beta_r\|x_t-y_{r,t}^i\|.\tag{9}$$

where $M=(4\tilde{L}^2+2\hat{L}\tilde{L}+2\hat{c}\Theta\tilde{L})(n_a+n_b)^2$. Now, we arrive at exactly the same form of upper bound as (17) in the smooth case. Choosing the step size sequence $\alpha_t=\frac{2}{\hat{c}(t+1)(n_a+n_b)}$, the proof is then completed by the same derivation as the Part III of the proof for Theorem 3.1. ∎

In the above proof, since we could not use Lipschitz continuity of subgradients, we applied the Intermediate Value Theorem after (instead of before) using the strong convexity assumption. The resulting quadratic term $\|w_t^i-x_*\|^2$ from strong convexity was then merged into the $\|x_t-x_*\|^2$ term and the other alike $\mathcal{O}(\alpha_t^2)$ terms, converting the bound on the iterate error to what we had seen before in the smooth and strongly convex case.

3.3. The convex case. If we replace strong convexity (Assumption 3.7) by (simple) convexity (see below), we will observe a degradation on the convergence rate. Corollary 3.2 can be seen as a counterpart of the single objective cases [4, Theorems 8.30 (b) and 8.35 (b)].

Assumption 3.8. ***(Convexity in the non-smooth case)*** *For both objective functions $i\in\{a,b\}$, we have*

$$f^i(\bar{x}) \;\ge\; f^i(x)+g^i(x)^\top(\bar{x}-x),\quad \forall(x,\bar{x})\in\mathcal{X}\times\mathcal{X},$$

for all subgradients $g^i(x)\in\partial f^i(x)$.

Corollary 3.2. ***(Sublinear convergence rate of SA2GD in the non-smooth and convex case)*** *Let Assumptions 3.1, 3.5–3.6, and 3.8 hold and x_* be a minimizer of the weighted function $S(\cdot,\lambda_*)$ in $\mathcal{X}$, where $\lambda_*=\lambda(n_a,n_b)=n_a/(n_a+n_b)$. Choosing a diminishing step size sequence $\alpha_t=\frac{\bar{\alpha}}{\sqrt{t}(n_a+n_b)}$, where $\bar{\alpha}$ is any positive constant, the sequence of iterates generated by the SA2GD algorithm satisfies*

$$\min_{t=1,\ldots,T}\mathbb{E}[S(x_t,\lambda_*)]-\mathbb{E}[S(x_*,\lambda_*)] \;\le\; \frac{\frac{\Theta^2}{2\bar{\alpha}}+4\bar{\alpha}\tilde{L}^2+2\bar{\alpha}\tilde{L}\hat{L}}{\sqrt{T}}.$$

Proof: The proof differs from the proof of Theorem 3.2 in Parts II and III. When applying convexity to the second inequality of (21), the term $\hat{c}\alpha_t n_a\|w_t^a-x_*\|^2$ does not appear on the right-hand side anymore, and we

Plugging the inequalities (3.2), (22)-(23) back in (19) and adding and subtracting $2\alpha_t(n_a f^a(x_t) + n_b f^b(x_t))$ on the right-hand side yield

$$\begin{aligned}
\mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] \leq\ & (1 - \hat{c}\alpha_t(n_a + n_b))\|x_t - x_*\|^2 + 2\alpha_t^2(n_a + n_b)^2\tilde{L}^2 \\
& - 2\alpha_t(n_a f^a(x_t) + n_b f^b(x_t) - (n_a f^a(x_*) + n_b f^b(x_*))) \\
& - 2\alpha_t\mathbb{E}_{\xi_t}[n_a f^a(w_t^a) - n_a f^a(x_t) + n_b f^b(w_t^b) - n_b f^b(x_t)] \\
& + 2\hat{c}\Theta\alpha_t(n_a\mathbb{E}_{\xi_t}[\|x_t - w_t^a\|] + n_b\mathbb{E}_{\xi_t}[\|x_t - w_t^b\|]) \\
& + 2\alpha_t\tilde{L}(\sum_{r=0}^{n_a-1}\mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^a\|] + \sum_{r=0}^{n_b-1}\mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^b\|]) \\
\leq\ & (1 - \hat{c}\alpha_t(n_a + n_b))\|x_t - x_*\|^2 + 2\alpha_t^2(n_a + n_b)^2\tilde{L}^2 \\
& - 2\alpha_t(n_a f^a(x_t) + n_b f^b(x_t) - (n_a f^a(x_*) + n_b f^b(x_*))) \\
& + 2\hat{L}\alpha_t(n_a\mathbb{E}_{\xi_t}[\|x_t - w_t^a\|] + n_b\mathbb{E}_{\xi_t}[\|x_t - w_t^b\|]) \\
& + 2\hat{c}\Theta\alpha_t(n_a\mathbb{E}_{\xi_t}[\|x_t - w_t^a\|] + n_b\mathbb{E}_{\xi_t}[\|x_t - w_t^b\|]) \\
& + 2\alpha_t\tilde{L}(\sum_{r=0}^{n_a-1}\mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^a\|] + \sum_{r=0}^{n_b-1}\mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^b\|]),
\end{aligned} \tag{24}$$

where, in the second inequality, we applied Assumption 3.5 to the fourth term.

We now recall the bounds (10)–(12) and (14) on the distance from $y_{r,t}^a$, $y_{r,t}^b$, w_t^a, and w_t^b to x_t. Applying expectation to these four bounds, followed by Assumption 3.6 (b), gives us

$$\begin{aligned}
\mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^a\|] &\leq \alpha_t n_a\tilde{L} \\
\mathbb{E}_{\xi_t}[\|x_t - y_{r,t}^b\|] &\leq \alpha_t(n_a + n_b)\tilde{L} \\
\mathbb{E}_{\xi_t}[\|x_t - w_t^a\|] &\leq \alpha_t n_a\tilde{L} \\
\mathbb{E}_{\xi_t}[\|x_t - w_t^b\|] &\leq \alpha_t(n_a + n_b)\tilde{L}.
\end{aligned} \tag{25}$$

Finally, plugging (25) into (24) yields

$$\begin{aligned}
\mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] \leq\ & (1 - \hat{c}\alpha_t(n_a + n_b))\|x_t - x_*\|^2 + \alpha_t^2 M \\
& - 2\alpha_t(n_a f^a(x_t) + n_b f^b(x_t) - (n_a f^a(x_*) + n_b f^b(x_*))),
\end{aligned}$$

Using $y_{r,t}^i = y_{0,t}^i - \sum_{j=0}^{r-1}\alpha_t g^i(y_{j,t}^i, \xi_t^j)$ and applying the triangle inequality, we have

$$\|x_t - y_{r,t}^a\| \leq \alpha_t\sum_{j=0}^{r-1}\|g^a(y_{j,t}^a, \xi_t^j)\|, \tag{10}$$

and

$$\begin{aligned}
\|x_t - y_{r,t}^b\| &= \|x_t - y_{n_a,t}^a + \sum_{j=0}^{r-1}\alpha_t g^b(y_{j,t}^b, \xi_t^{n_a+j})\| \\
&\leq \alpha_t\sum_{j=0}^{n_a-1}\|g^a(y_{j,t}^a, \xi_t^j)\| + \alpha_t\sum_{j=0}^{r-1}\|g^b(y_{j,t}^b, \xi_t^{n_a+j})\|.
\end{aligned} \tag{11}$$

Plugging (10) into (9) with $i = a$ results in (note that $y_{0,t}^a = x_t$)

$$\begin{aligned}
\|x_t - w_t^a\| &\leq \alpha_t\sum_{r=1}^{n_a-1}\beta_r\sum_{j=0}^{r-1}\|g^a(y_{j,t}^a, \xi_t^j)\| \\
&= \alpha_t\sum_{j=0}^{n_a-1}\|g^a(y_{j,t}^a, \xi_t^j)\|\sum_{r=j+1}^{n_a-1}\beta_r \leq \alpha_t\sum_{j=0}^{n_a-1}\|g^a(y_{j,t}^a, \xi_t^j)\|.
\end{aligned} \tag{12}$$

Applying expectations on both sides of (12), followed by the bound (4) on the second moment of the stochastic gradients,

$$\mathbb{E}_{\xi_t}[\|x_t - w_t^a\|] \leq \alpha_t n_a\sqrt{G + \bar{G}M_\nabla}. \tag{13}$$

Similarly, merging (11) into (9) with $i = b$ leads to

$$\|x_t - w_t^b\| \leq \alpha_t\sum_{j=0}^{n_a-1}\|g^a(y_{j,t}^a, \xi_t^j)\| + \alpha_t\sum_{j=0}^{n_b-1}\|g^b(y_{j,t}^b, \xi_t^{n_a+j})\|, \tag{14}$$

and applying expectations

$$\mathbb{E}_{\xi_t}[\|x_t - w_t^b\|] \leq \alpha_t(n_a + n_b)\sqrt{G + \bar{G}M_\nabla}. \tag{15}$$

Finally, combining (8), (13), and (15) yields

$$\begin{aligned}
& 2\alpha_t\|x_t - x_*\|\sum_{i\in\{a,b\}}\mathbb{E}_{\xi_t}[\|n_i\nabla f^i(w_t^i) - n_i\nabla f^i(x_t)\|] \\
& \leq 2\alpha_t^2 L\Theta(n_a + n_b)^2\sqrt{G + \bar{G}M_\nabla}.
\end{aligned} \tag{16}$$

Part III: Bound on the optimality gap in terms of weighted-sum function. Applying inequalities (7) and (16) to (6) leads to

$$\begin{aligned}\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2] \leq\ & (1-\alpha_t(n_a+n_b)c)\|x_t-x_*\|^2 \\ & -2\alpha_t(n_a+n_b)(S(x_t,\lambda_*)-S(x_*,\lambda_*))+\alpha_t^2 M,\end{aligned} \tag{17}$$

where we let $M = 2(n_a+n_b)^2(G+\bar{G}M_\nabla + L\Theta\sqrt{G+\bar{G}M_\nabla})$. Plugging in $\alpha_t = \frac{2}{c(n_a+n_b)(t+1)}$ and rearranging the last inequality result in

$$\begin{aligned}& S(x_t,\lambda_*)-S(x_*,\lambda_*) \\ \leq\ & \frac{(1-\alpha_t(n_a+n_b)c)\|x_t-x_*\|^2-\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2]+\alpha_t^2 M}{2\alpha_t(n_a+n_b)} \\ \leq\ & \frac{c(t-1)}{4}\|x_t-x_*\|^2-\frac{c(t+1)}{4}\mathbb{E}_{\xi_t}[\|x_{t+1}-x_*\|^2]+\frac{\tilde{M}}{c(t+1)},\end{aligned}$$

where $\tilde{M}=\frac{M}{(n_a+n_b)^2}$. By taking total expectation over $\{\xi_t\}$, multiplying both sides by t, and summing over $t=1,\dots,T$, one obtains

$$\begin{aligned}& \sum_{t=1}^T t(\mathbb{E}[S(x_t,\lambda_*)]-S(x_*,\lambda_*)) \\ \leq\ & \sum_{t=1}^T\left(\frac{ct(t-1)}{4}\mathbb{E}[\|x_t-x_*\|^2]-\frac{ct(t+1)}{4}\mathbb{E}[\|x_{t+1}-x_*\|^2]\right) \\ & +\sum_{t=1}^T\frac{\tilde{M}t}{c(t+1)} \\ \leq\ & -\frac{cT(T+1)}{4}\mathbb{E}[\|x_{T+1}-x_*\|^2]+\sum_{t=1}^T\frac{\tilde{M}t}{c(t+1)}\ \leq\ \frac{T}{c}\tilde{M}.\end{aligned}$$

Dividing both sides of the last inequality by $\sum_{t=1}^T t$ yields

$$\min_{t=1,\dots,T}\mathbb{E}[S(x_t,\lambda_*)]-S(x_*,\lambda_*)\ \leq\ \frac{2}{c(T+1)}\tilde{M},$$

which concludes the proof. ∎

The key of the proof was to bound the deviation from w_t^i to x_t using the step size α_t. It is in fact the use of the decaying step size that compensates for the error $\mathcal{O}(\alpha_t^2)$ generated when bundling the gradients using the Intermediate Value Theorem. Note that the above theorem is still guaranteed when one

$$\begin{aligned}\|w_t^a-x_*\|^2 &\geq (\|w_t^a-x_t\|-\|x_t-x_*\|)^2 \\ &= \|x_t-w_t^a\|^2+\|x_t-x_*\|^2-2\|x_t-w_t^a\|\|x_t-x_*\| \\ &\geq \|x_t-x_*\|^2-2\|x_t-w_t^a\|\|x_t-x_*\|.\end{aligned}$$

Applying the above inequality to (21) and using Assumption 3.1 result in

$$\begin{aligned}& -2\alpha_t(x_t-x_*)^\top\sum_{r=0}^{n_a-1}(f^a)'(y_{r,t}^a) \\ \leq\ & -2\alpha_t(n_a f^a(w_t^a)-n_a f^a(x_*))-\hat{c}\alpha_t n_a\|x_t-x_*\|^2 \\ & +2\hat{c}\Theta\alpha_t n_a\|x_t-w_t^a\|+2\alpha_t\sum_{r=0}^{n_a-1}\|x_t-y_{r,t}^a\|\tilde{L}_a.\end{aligned} \tag{22}$$

Similarly, we can obtain

$$\begin{aligned}& -2\alpha_t(x_t-x_*)^\top\sum_{r=0}^{n_b-1}(f^b)'(y_{r,t}^b) \\ \leq\ & -2\alpha_t(n_b f^b(w_t^b)-n_b f^b(x_*))-\hat{c}\alpha_t n_b\|x_t-x_*\|^2 \\ & +2\hat{c}\Theta\alpha_t n_b\|x_t-w_t^b\|+2\alpha_t\sum_{r=0}^{n_b-1}\|x_t-y_{r,t}^b\|\tilde{L}_b,\end{aligned} \tag{23}$$

where w_t^b is a point in the convex hull of $\{y_{r,t}^b\}_{r=0}^{n_b-1}$.

$2\alpha_t \mathbb{E}_{\xi_t}[(x_t - x_*)^\top \sum_{r=0}^{n_b-1}(f^b)'(y^b_{r,t})]$, we will apply Assumption 3.7 multiple times. For simplicity, the expectation symbol is temporarily ignored,

$$
\begin{aligned}
& - 2\alpha_t (x_t - x_*)^\top \sum_{r=0}^{n_a-1} (f^a)'(y^a_{r,t}) \\
= \; & - 2\alpha_t (x_t - x_*)^\top (f^a)'(x_t) \\
& - 2\alpha_t (y^a_{1,t} - x_*)^\top (f^a)'(y^a_{1,t}) - 2\alpha_t (x_t - y^a_{1,t})^\top (f^a)'(y^a_{1,t}) \\
& \cdots \\
& - 2\alpha_t (y^a_{n_a-1,t} - x_*)^\top (f^a)'(y^a_{n_a-1,t}) - 2\alpha_t (x_t - y^a_{n_a-1,t})^\top (f^a)'(y^a_{n_a-1,t}) \\
\leq \; & - 2\alpha_t \sum_{r=0}^{n_a-1} (f^a(y^a_{r,t}) - f^a(x_*) + \frac{\hat{c}}{2}\|y^a_{r,t} - x_*\|^2) \\
& - 2\alpha_t \sum_{r=0}^{n_a-1} (x_t - y^a_{r,t})^\top (f^a)'(y^a_{r,t}) \\
\leq \; & - 2\alpha_t (n_a f^a(w^a_t) - n_a f^a(x_*)) - \hat{c}\alpha_t n_a \|w^a_t - x_*\|^2 + 2\alpha_t \sum_{r=0}^{n_a-1} \|x_t - y^a_{r,t}\| \tilde{L}_a,
\end{aligned}
\tag{21}
$$

where we applied Assumption 3.6 (a) and Assumption 3.7 to get the first inequality. The last inequality holds by applying Proposition A.1 to the quadratic function $\psi^a(x) = f^a(x) - f^a(x_*) + \frac{\hat{c}}{2}\|x - x_*\|^2$, the Cauchy–Schwarz inequality, and the combination of Assumption 3.6 (b) and the Jensen's inequality $\|\mathbb{E}_{\xi_t}[g^i(x_t, \xi_t)]\|^2 \leq \mathbb{E}_{\xi_t}[\|g^i(x_t, \xi_t)\|^2]$. Here, w^a_t is a point in the convex hull of $\{y^a_{r,t}\}_{r=0}^{n_a-1}$.

The second term $\|w^a_t - x_*\|^2$ on the right hand side can be handled using the triangle inequality, i.e.,

relaxes Assumption 3.3 to the case where both objectives are convex and one of them is strongly convex, as (7) still holds in such a case. A rate in terms of the iterates can be also derived.

Corollary 3.1. *Let Assumptions 3.1-3.4 hold and x_* be the unique minimizer of the weighted function $S(\cdot, \lambda_*)$ in $\mathcal{X}$, where $\lambda_* = \lambda(n_a, n_b) = n_a/(n_a + n_b)$. Choosing a diminishing step size sequence $\alpha_t = \gamma/t$ where $\gamma > \frac{1}{2(n_a+n_b)c}$, the sequence of iterates generated by the SA2GD algorithm satisfies*

$$\mathbb{E}[\|x_T - x_*\|^2] \; \leq \; \frac{\max\{2\gamma^2 M(2c(n_a+n_b)\gamma - 1)^{-1}, \|x_0 - x_*\|^2\}}{T}.$$

where $M = 2(n_a + n_b)^2(G + \bar{G}M_\nabla + L\Theta\sqrt{G + \bar{G}M_\nabla})$.

Proof: Since the weighted-sum function $S(\cdot, \lambda_*)$ is strongly convex, one also has

$$(\nabla S(x, \lambda) - \nabla_x S(\bar{x}, \lambda))^\top (x - \bar{x}) \; \geq \; c\|x - \bar{x}\|^2, \quad \forall (x, \bar{x}) \in \mathcal{X} \times \mathcal{X}.$$

Letting $x = x_t$, $\bar{x} = x_*$, and $\lambda = \lambda_k$ in the above inequality leads to

$$\nabla S(x_t, \lambda_*)^\top (x_t - x_*) \; \geq \; c\|x_t - x_*\|^2,$$

where we used the fact that $\nabla S(x_*, \lambda_*)^\top (x_t - x_*) = (P_{\mathcal{X}} \nabla S(x_*, \lambda_*))^\top (x_t - x_*) = 0$. Hence,

$$(x_t - x_*)^\top (n_a \nabla f^a(x_t) + n_b \nabla f^b(x_t)) \; \geq \; c(n_a + n_b)\|x_t - x_*\|^2. \tag{18}$$

Plugging (16) and (18) in (6) results in

$$\mathbb{E}[\|x_{t+1} - x_*\|^2] \; \leq \; (1 - 2c(n_a + n_b)\alpha_t)\mathbb{E}[\|x_t - x_*\|^2] + \alpha_t^2 M.$$

Using $\alpha_t = \gamma/t$ with $\gamma > 1/(2c(n_a + n_b))$ and the induction argument in [31, Eq. (2.9) and (2.10)] lead us to the result. ∎

3.2. The non-smooth and strongly convex case. In this section, we analyze the convergence rates for stochastic alternating bi-objective descent (SA2GD) when subgradients are used in Algorithm 1. Instead of using gradients, the notation $g^i(x, \xi), \forall i \in \{a, b\}$, denotes now the individual stochastic subgradient generated using a random variable ξ. Let $\partial f^i(x)$ be the subdifferential at x. Next, we describe the assumptions under which SA2GD will be analyzed (in addition to Assumption 3.1). We formalize first a classical assumption of Lipschitz continuity of both objective functions, which is often satisfied in practice.

Assumption 3.5. ***(Lipschitz continuous functions)*** *Both objective functions are Lipschitz continuous with Lipschitz constants $\hat{L}_i > 0$, $i \in \{a, b\}$, i.e.,*

$$|f^i(\bar{x}) - f^i(x)| \leq \hat{L}_i \|\bar{x} - x\|, \quad \forall (x, \bar{x}) \in \mathcal{X} \times \mathcal{X}.$$

We also impose the following two classical assumptions about stochastic subgradients.

Assumption 3.6. *For both objective functions $i \in \{a, b\}$, and all iterates $t \in \mathbb{N}$, the stochastic subgradients $g^i(x_t, \xi_t)$ satisfy the following:*

(a) ***(Unbiasedness)*** $\mathbb{E}_{\xi_t}[g^i(x_t, \xi_t)] \in \partial f^i(x_t)$.

(b) ***(Boundness)*** *There exist positive constants $\tilde{L}_i > 0$ such that $\mathbb{E}_{\xi_t}[\|g^i(x_t, \xi_t)\|^2] \leq \tilde{L}_i^2$.*

The above assumptions are commonly used ones in classical stochastic subgradient methods [4, Section 8.3]. For conciseness, we denote $\hat{L} = \max\{\hat{L}_a, \hat{L}_b\}$ and $\tilde{L} = \max\{\tilde{L}_a, \tilde{L}_b\}$.

In addition to the above assumptions, we impose strong convexity in both objective functions.

Assumption 3.7. ***(Strong convexity in the non-smooth case)*** *For both objective functions $i \in \{a, b\}$, there exists a scalar $\hat{c}_i > 0$ such that*

$$f^i(\bar{x}) \geq f^i(x) + g^i(x)^\top(\bar{x} - x) + \frac{\hat{c}_i}{2}\|\bar{x} - x\|^2, \quad \forall (x, \bar{x}) \in \mathcal{X} \times \mathcal{X},$$

for all subgradients $g^i(x) \in \partial f^i(x)$.

Based on the above two assumptions, the weighted-sum function $S(x, \lambda) = \lambda f^a(x) + (1 - \lambda) f^b(x), \forall \lambda \in [0, 1]$, is also Lipschitz continuous with constant $\hat{L}$ and strongly convex with constant $\hat{c} = \min\{\hat{c}_a, \hat{c}_b\}$. Now, we are ready to show a similar convergence rate result as Theorem 3.1. Readers are referred to [4, Theorems 8.31 (a) and 8.37 (a)] for the single-objective counterparts.

Theorem 3.2. ***(Sublinear convergence rate of SA2GD in the non-smooth and strongly convex case)*** *Let Assumptions 3.1 and 3.5–3.7 hold and x_* be the unique minimizer of the weighted function $S(\cdot, \lambda_*)$ in $\mathcal{X}$, where $\lambda_* = \lambda(n_a, n_b) = n_a/(n_a + n_b)$. Choosing a diminishing step size sequence $\alpha_t = \frac{2}{\hat{c}(t+1)(n_a+n_b)}$, the sequence of iterates generated by the SA2GD algorithm satisfies*

$$\min_{t=1,\ldots,T} \mathbb{E}[S(x_t, \lambda_*)] - S(x_*, \lambda_*) \leq \frac{4}{\hat{c}(T+1)}\left(2\tilde{L}^2 + \hat{L}\tilde{L} + \hat{c}\Theta\tilde{L}\right).$$

Proof: Similar to the proof of Theorem 3.1, the proof is divided in three parts. In the first part, one obtains an upper bound on the norm of the iterates, $\mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2]$. Strong convexity is applied in the second part. We skip the third part as it is exactly the same as part III of Theorem 3.1.

Part I: Bound on the iterates error. At any iteration t, the sequence of stochastic subgradients is computed from drawing the sequence of random variables $\xi_t = \{\xi_t^0, \ldots, \xi_t^{n_a+n_b-1}\}$. Replacing the gradients in (3) by the corresponding subgradients, one has

$$\begin{aligned}
&\mathbb{E}_{\xi_t}[\|x_{t+1} - x_*\|^2] \\
&\leq \|x_t - x_*\|^2 + \alpha_t^2 \mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_a-1} g^a(y_{r,t}^a, \xi_t^r) + \sum_{r=0}^{n_b-1} g^b(y_{r,t}^b, \xi_t^{n_a+r})\|^2] \\
&\quad - 2\alpha_t \mathbb{E}_{\xi_t}[(x_t - x_*)^\top (\sum_{r=0}^{n_a-1} (f^a)'(y_{r,t}^a) + \sum_{r=0}^{n_b-1} (f^b)'(y_{r,t}^b))],
\end{aligned} \tag{19}$$

where $(f^a)'(x) \in \partial f^a(x)$ and $(f^b)'(x) \in \partial f^b(x)$ denote certain deterministic subgradients at x for the two objectives respectively.

The upper bound for the second moment of the sequence of stochastic subgradients $\{g^a(y_{r,t}^a, \xi_t^r)\}_{r=0}^{n_a-1}$ and $\{g^b(y_{r,t}^b, \xi_t^{n_a+r})\}_{r=0}^{n_b-1}$ at each iteration t can be derived as follows.

$$\begin{aligned}
&\mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_a-1} g^a(y_{r,t}^a, \xi_t^r) + \sum_{r=0}^{n_b-1} g^b(y_{r,t}^b, \xi_t^{n_a+r})\|^2] \\
&\leq 2\mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_a-1} g^a(y_{r,t}^a, \xi_t^r)\|^2] + 2\mathbb{E}_{\xi_t}[\|\sum_{r=0}^{n_b-1} g^b(y_{r,t}^b, \xi_t^{n_a+r})\|^2] \\
&\leq 2n_a \sum_{r=0}^{n_a-1} \mathbb{E}_{\xi_t}[\|g^a(y_{r,t}^a, \xi_t^r)\|^2] + 2n_b \sum_{r=0}^{n_b-1} \mathbb{E}_{\xi_t}[\|g^b(y_{r,t}^b, \xi_t^{n_a+r})\|^2] \\
&\leq 2(n_a + n_b)^2 \tilde{L}^2,
\end{aligned} \tag{20}$$

where $\tilde{L} = \max(\tilde{L}_a, \tilde{L}_b)$.

Part II: Using strong convexity and Intermediate Value Theorem. In order to get a bound for the term $-2\alpha_t \mathbb{E}_{\xi_t}[(x_t - x_*)^\top \sum_{r=0}^{n_a-1} (f^a)'(y_{r,t}^a)] -$