STIMULUS: Achieving Fast Convergence and Low Sample Complexity in Stochastic Multi-Objective Learning

Zhuqing Liu[1], **Chaosheng Dong**[2], **Michinari Momma**[2], **Simone Shao**[2],
Shaoyuan Xu[2], **Yan Gao**[2], **Haibo Yang**[3], **Jia Liu**[4]

[1]Computer Science and Engineering, University of North Texas, Denton, TX, USA
[2]Amazon, Seattle, WA, USA
[3]Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, USA
[4]Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA

Abstract

Recently, multi-objective optimization (MOO) has gained attention for its broad applications in ML, operations research, and engineering. However, MOO algorithm design remains in its infancy and many existing MOO methods suffer from unsatisfactory convergence rate and sample complexity performance. To address this challenge, in this paper, we propose an algorithm called STIMULUS (stochastic path-integrated multi-gradient recursive estimator), a new and robust approach for solving MOO problems. Different from the traditional methods, STIMULUS introduces a simple yet powerful recursive framework for updating stochastic gradient estimates to improve convergence performance with low sample complexity. In addition, we introduce an enhanced version of STIMULUS, termed STIMULUS-M, which incorporates a momentum term to further expedite convergence. We establish $\mathcal{O}(1/T)$ convergence rates of the proposed methods for non-convex settings and $\mathcal{O}(\exp -\mu T)$ for strongly convex settings, where T is the total number of iteration rounds. Additionally, we achieve the state-of-the-art $O\left(n+\sqrt{n}\epsilon^{-1}\right)$ sample complexities for non-convex settings and $\mathcal{O}\left(n+\sqrt{n}\ln(\mu/\epsilon)\right)$ for strongly convex settings, where $\epsilon > 0$ is a desired stationarity error. Moreover, to alleviate the periodic full gradient evaluation requirement in STIMULUS and STIMULUS-M, we further propose enhanced versions with adaptive batching called STIMULUS$^+$ / STIMULUS-M$^+$ and provide their theoretical analysis.

1 INTRODUCTION

1) Background of multi-objective learning: Machine learning (ML) has always heavily relied on optimization formulations and algorithms. While traditional ML problems generally focus on minimizing a single loss function, many emergent complex-structured multi-task ML problems require balancing *multiple* objectives that are often conflicting (e.g., multi-agent reinforcement learning [Parisi et al., 2014], multi-task fashion representation learning [Jiao et al., 2022, 2023], multi-task recommendation system [Chen et al., 2019, Zhou et al., 2023], multi-model learning in video captioning [Pasunuru and Bansal, 2017], and multi-label learning-to-rank [Mahapatra et al., 2023a,b]). Such ML applications necessitate solving *multi-objective* optimization (MOO) problems, which can be expressed as:

$$\min_{\mathbf{x}\in\mathcal{D}} \mathbf{F}(\mathbf{x}) := [f_1(\mathbf{x}), \cdots, f_S(\mathbf{x})], \qquad (1)$$

where $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^d$ is the model parameters. Here, each f_s denotes the objective function of task $s \in [S]$, $f_s(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} f_{sj}(\mathbf{x};\xi_{sj})$, where n denotes the total number of samples, ξ_{sj} denotes the j-th sample for task s. However, unlike traditional single-objective optimization, there may not exist a common $\mathbf{x}$-solution in MOO that can simultaneously minimize all objective functions. Instead, a more relevant optimality criterion in MOO is the notion of *Pareto-optimal solutions*, where no objective can be further improved without sacrificing other objectives. Moreover, in settings where the set of objective functions are non-convex, searching for Pareto-optimal solutions is intractable in general. In such scenarios, the goal of MOO is usually weakened to finding a *Pareto-stationary solution*, where no improving direction exists for any objective without sacrificing other objectives.

2) Motivating application: Multi-label learning to rank (MLLTR) problem. Problem (1) can be applied to a number of interesting real-world problems. Here, we provide one concrete example to further motivate its practical relevance:

The learning to Rank (LTR) method is a common technique used to rank information based on relevance, but it often struggles with ambiguity because of the noisy nature of human-generated data, like product ratings. To tackle this,

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:2711–2747.

Table 1: Convergence comparisons between MOO algorithms, where n is the size of dataset; ϵ is the convergence error. Our proposed algorithms are marked in a shaded background.

Algorithm	Multi-gradient	Non-convex case		Strongly-Convex case	
		Rate	Sample Complexity	Rate	Sample Complexity
MGD [Fliege et al., 2019]	Deterministic	$\mathcal{O}\left(T^{-1}\right)$	$\mathcal{O}\left(n\epsilon^{-1}\right)$	$\mathcal{O}(\exp(-\mu T))$	$\mathcal{O}\left(n\ln(\mu/\epsilon)\right)$
SMGD [Yang et al., 2022]	Stochastic	$\mathcal{O}\left(T^{-1/2}\right)$	$\mathcal{O}\left(\epsilon^{-2}\right)$	$\mathcal{O}\left(T^{-1}\right)$	$\mathcal{O}\left(\epsilon^{-1}\right)$
MoCo [Fernando et al., 2022]	Stochastic	$\mathcal{O}\left(T^{-1/2}\right)$	$\mathcal{O}\left(\epsilon^{-2}\right)$	$\mathcal{O}\left(T^{-1}\right)$	$\mathcal{O}\left(\epsilon^{-1}\right)$
MoCo+ [Fernando et al., 2024]	Stochastic	$\mathcal{O}\left(T^{-2/3}\right)$	$\mathcal{O}\left(\epsilon^{-1.5}\right)$	-	-
CR-MOGM [Zhou et al., 2022]	Stochastic	$\mathcal{O}\left(T^{-1/2}\right)$	$\mathcal{O}\left(\epsilon^{-2}\right)$	$\mathcal{O}\left(T^{-1}\right)$	$\mathcal{O}\left(\epsilon^{-1}\right)$
STIMULUS/ STIMULUS-M	Stochastic	$\mathcal{O}\left(T^{-1}\right)$	$\mathcal{O}\left(n+\sqrt{n}\epsilon^{-1}\right)$	$\mathcal{O}(\exp(-\mu T))$	$\mathcal{O}\left(n+\sqrt{n}\ln(\mu/\epsilon)\right)$
STIMULUS^{+} / STIMULUS-M^{+}	Stochastic	$\mathcal{O}\left(T^{-1}\right)$	$\mathcal{O}\left(n+\sqrt{n}\epsilon^{-1}\right)$	$\mathcal{O}(\exp(-\mu T))$	$\mathcal{O}\left(n+\sqrt{n}\ln(\mu/\epsilon)\right)$

Multi-Label Learning to Rank (MLLTR) offers a more refined approach. MLLTR addresses the inherent challenges of traditional LTR methods by integrating multiple relevance criteria into the ranking model. This allows for a more comprehensive representation of diverse crucial objectives.

- *Learning to Rank:* Let A be the training set, consisting of pairs $(\mathbf{a}_i, b_i)$ where $\mathbf{a}_i \in \mathbb{R}^d$ representing features, and $\mathbf{b}$ is the corresponding list of relevance labels b_i, and $i = 1, \ldots, n$. We note that the lists $\mathbf{a}$ within the training set may not all be of the same length. $\mathbf{x}$ is the model parameter.

  The goal of the learning-to-rank problem is to find a scoring function f that optimizes a chosen Information Retrieval (IR) metric, such as Normalized Discounted Cumulative Gain (NDCG), on the test set. The scoring function f is trained to minimize the mean of a surrogate loss l across the training data: $f_{single}(\mathbf{x}) = \frac{1}{|A|}\sum_{(\mathbf{a},\mathbf{b})\in A} l(f(\mathbf{x};\mathbf{a}),\mathbf{b})$.
- *Multi-label Learning to Rank:* Learning to Rank from multiple relevance labels. In the problem of Multi-label learning to rank (MLLTR), different relevance criteria are measured, providing multiple labels for each feature vector $\mathbf{a}_i \in \mathbb{R}^d$. The goal of MLLTR is still the same as that of LTR, which is to learn a scoring function $f(\mathbf{x};\mathbf{a})$ that assigns a scalar value to each feature vector $\mathbf{a}_i \in \mathbb{R}^d$. Here, we consider a set of training examples denoted by $\mathbf{a}_i \in \mathbb{R}^d$, where $i = 1, \ldots, n$. Associated with each training example $\mathbf{a}_i$ is a vector of class labels: $\mathbf{b}_i = \left(b_i^1, \ldots, b_i^K\right)$, indicating the labels of $\mathbf{a}_i$. Here, K is the total count of possible labels. In the multi-label learning to rank problem, the objective is to construct K distinct classification functions: $f_k(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$, for $k = 1, \ldots, K$, each tailored to a specific label.

  In MLLTR, the cost is a vector-valued function: $f(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), f_K(\mathbf{x})]$, naturally making it an MOO problem.

In the search ranking domain, the objective is to rank search results based on their relevance to user queries and other factors such as popularity, user feedback, and conversion rates. The loss function in search ranking not only considers relevance but also takes into account various performance metrics, such as click-through rates (CTR), dwell time, or conversion ratesLyu et al. [2020], Yang et al. [2020], Xiao et al. [2020]. The goal is to optimize the ranking of search results to maximize user satisfaction and engagement. Common loss functions used in search ranking include pairwise ranking lossKumar et al. [2020], Jing et al. [2019], Wang et al. [2021], listwise lossRevaud et al. [2019], Yu et al. [2019], or evaluation metrics like normalized discounted cumulative gain (NDCG)Bruch et al. [2019] or mean average precision (MAP)Revaud et al. [2019]. These loss functions aim to capture the overall quality of the search ranking by considering both relevance and performance metrics.

The multi-label learning to rank problem typically involves a larger number of labels, which increases the dimensionality of the output space. This higher dimensionality often necessitates a greater number of samples to accurately train models, resulting in increased sample complexity. Therefore, this motivates us to propose a new family of algorithms for low sample complexity and fast convergence rates.

3) Related works and motivation: To date, existing MOO algorithms in the literature can be generally categorized as gradient-free and gradient-based methods. Typical gradient-free methods include evolutionary MOO algorithms and Bayesian MOO algorithms [Zhang and Li, 2007, Deb et al., 2002, Belakaria et al., 2020, Laumanns and Ocenasek, 2002]. These techniques are suitable for small-scale problems but inefficient in solving high-dimensional MOO models (e.g., deep neural networks). Notably, gradient-based methods have attracted increasing attention recently due to their stronger empirical performances. Specifically, following a similar token of (stochastic) gradient descent methods for single-objective optimization, (stochastic) multi-gradient descent (MGD/SMGD) algorithms have been proposed in [Fliege et al., 2019, Fernando et al., 2022, Zhou et al., 2022, Liu and Vicente, 2021]. The basic idea of MGD/SMGD is to iteratively update the $\mathbf{x}$-variable following a common descent direction for all the objectives through a time-varying convex combination of (stochastic) gradients of all objective functions. Although MGD-type algorithms enjoy a fast $\mathcal{O}(1/T)$ convergence rate (T denotes the number of iterations) in finding a Pareto-stationary solution, their $\mathcal{O}(n)$ per-iteration computation complexity in full multi-gradient evaluations becomes prohibitive when the dataset size n is large. As a result, SMGD-type algorithms are often more favored in practice thanks to the lower per-iteration computation complexity in evaluating stochastic multi-gradients. However, due to the noisy stochastic multi-gradient evaluations, SMGD-type algorithms typically exhibit a slow $\mathcal{O}(1/\sqrt{T})$ convergence rate, which also induces a high $\mathcal{O}(\epsilon^{-2})$ sample complexity. Although SMGD is easier to implement in practice thanks to the use of stochastic multi-gradient, it has been shown that the noisy common descent direction in SMGD could potentially cause divergence (cf. the example in Sec. 4 in [Zhou et al., 2022]). There also have been recent works on using momentum-based methods for bias mitigation in MOO, named MoCo [Fernando et al., 2022], MoCo+ [Fernando et al., 2024], CR-MOGM [Zhou et al., 2022]. For easier comparisons, we summarize the state-of-the-art gradient-based MOO algorithms and their convergence rate results under non-convex and strongly convex settings in Table 1. We note that given the limited research on finite-sum multi-objective optimization, we included broader comparisons.

In light of these major limitations of SMGD-type algorithms, a fundamental question naturally emerges:

> **(Q)**: Is it possible to develop fast-convergent stochastic MOO algorithms in the sense of matching the convergence rate of deterministic MGD-type methods, while having a low per-iteration computation complexity as in SMGD-type algorithms, as well as achieving a low overall sample complexity?

To be specific, our algorithms differ from them in the following key aspects: (i) Our algorithms only require a constant level step size, which is easier to tune in practice. (ii) Our STIMULUS family of algorithms has a lower sample complexity compared to all other existing methods.

4) Technical Challenges: As in traditional single-objective optimization, a natural idea to achieve both fast convergence and low sample complexity in MOO is to employ the so-called "variance reduction" (VR) techniques to tame the noise in stochastic multi-gradients in SMGD-type methods. However, due to the complex coupling nature of MOO problems, developing VR-assisted algorithms for SMGD-type algorithms faces the following challenges *unseen* in their single-objective counterparts:

(1) Since SMGD-type methods aim to identify the Pareto front (i.e., the set of all Pareto-optimal/stationary solutions), it is critical to ensure that the use of VR techniques does not introduce new bias into the already-noisy SGMD-type search process, which drives the search process toward certain regions of the Pareto front. (2) MOO problems often involve higher computational complexity compared to single-objective problems due to the need to evaluate multiple objectives simultaneously. Incorporating VR techniques adds another layer of complexity, as it requires additional computations to estimate and reduce variance across multiple objectives. (3) Conducting theoretical analysis to prove the convergence performance of some proposed VR-based SMGD-type techniques also contains multiple challenges, including how to quantify multiple conflicting objectives, navigating trade-offs between them, handling the non-convexity objective functions, and managing the computational cost of evaluations. All of these analytical challenges are quite different from those in single-objective optimization theoretical analysis, which necessitate specialized proofs and analyses are needed to effectively tackle these challenges and facilitate efficient exploration of the Pareto optimality/stationarity.

5) Main Contributions: The major contribution of this paper is that we overcome the aforementioned technical challenges and develop a suite of new VR-assisted SMGD-based MOO algorithms called STIMULUS (stochastic path-integrated multi-gradient recursive estimator) to achieve both fast convergence and low sample complexity in MOO. Our main technical results are summarized as follows:

- Our STIMULUS algorithm not only enhances computational efficiency but also significantly reduces multi-gradient estimation variance, leading to more stable convergence trajectories and overcoming the divergence problem of SMGD. We theoretically establish a convergence rate of $\mathcal{O}(1/T)$ for STIMULUS in non-convex settings (typical in ML), which further implies a low sample complexity of $O\left(n + \sqrt{n}\epsilon^{-1}\right)$. In the special setting where the objectives are strongly convex, we show that STIMULUS has a linear convergence rate of $\mathcal{O}(\exp(-\mu T))$, which implies an even lower sample complexity of $\mathcal{O}\left(n + \sqrt{n}\ln(\mu/\epsilon)\right)$.
- To further improve the performance of STIMULUS, we develop an enhanced version called STIMULUS-M, which incorporates momentum information to expedite convergence speed. Also, to relax the requirement for periodic full multi-gradient evaluations in STIMULUS and STIMULUS-M, we propose two enhanced variants called STIMULUS^{+} and STIMULUS-M^{+} based on adaptive batching, respectively. We provide theoretical convergence and sample complexity analyses for all these enhanced variants. These enhanced variants expand the practical utility of STIMULUS, offering efficient solutions that not only accelerate optimization processes but also alleviate computational burdens in a wide spectrum of multi-objective optimization applications.
- We conduct extensive experiments on a variety of chal-

lenging MOO problems to verify our theoretical results and illustrate the efficacy of the STIMULUS algorithm family. Our experiments demonstrate the efficiency of the STIMULUS algorithm family over existing state-of-the-art MOO methods, which underscore the robustness, scalability, and flexibility of our STIMULUS algorithm family in complex MOO applications.

2 PRELIMINARIES

To facilitate subsequent technical discussions, in this section, we first provide a primer on MOO fundamentals and formally define the notions of Pareto optimality/stationarity, ϵ-stationarity in MOO, and the associated sample complexity. Then, we will give an overview of the most related work in the MOO literature, thus putting our work into comparative perspectives.

Multi-objective Optimization: A primer. As introduced in Section 1, MOO aims to optimize multiple objectives in Eq. (1) simultaneously. However, since in general there may not exist an $\mathbf{x}$-solution that minimizes all objectives at the same time in MOO, the more appropriate notion of optimality in MOO is the so-called *Pareto optimality,* which is formally defined as follows:

Definition 1 ((Weak) Pareto Optimality)**.** *Given two solutions $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x}$ is said to dominate $\mathbf{y}$ only if $f_s(\mathbf{x}) \leq f_s(\mathbf{y}), \forall s \in [S]$ and there exists at least one function, f_s, where $f_s(\mathbf{x}) < f_s(\mathbf{y})$. A solution $\mathbf{x}_*$ is Pareto optimal if no other solution dominates it. A solution $\mathbf{x}$ is defined as weakly Pareto optimal if there is no solution $\mathbf{y}$ for which $f_s(\mathbf{x}) > f_s(\mathbf{y}), \forall s \in [S]$.*

Finding a Pareto-optimal solution in MOO is as complex as solving single-objective non-convex optimization problems and is NP-Hard in general. Consequently, practical efforts in MOO often aim to find a solution that meets the weaker notion called Pareto-stationarity (a necessary condition for Pareto optimality), which is defined as follows Fliege and Svaiter [2000], Miettinen [2012]:

Definition 2 (Pareto Stationarity)**.** *A solution $\mathbf{x}$ is Pareto-stationary if no common descent direction $\mathbf{d} \in \mathbb{R}^d$ exists such that $\nabla f_s(\mathbf{x})^\top \mathbf{d} < 0, \forall s \in [S]$.*

Note also that in the special setting with strongly convex objective functions, Pareto-stationary solutions are Pareto-optimal. Following directly from Pareto-stationarity in Definition 2, gradient-based MOO algorithms strive to find a common descent (i.e., improving) direction $\mathbf{d} \in \mathbb{R}^d$, such that $\nabla f_s(\mathbf{x})^\top \mathbf{d} \leq 0, \forall s \in [S]$. If such a direction does not exist at $\mathbf{x}$, then $\mathbf{x}$ is Pareto-stationary. Toward this end, the MGD method [Désidéri, 2012] identifies an optimal weight $\boldsymbol{\lambda}^*$ for the multi-gradient set $\nabla \mathbf{F}(\mathbf{x}) \triangleq \{\nabla f_s(\mathbf{x}), \forall s \in [S]\}$ by solving $\boldsymbol{\lambda}^*(\mathbf{x}) \in \operatorname{argmin}_{\boldsymbol{\lambda} \in C} \|\boldsymbol{\lambda}^\top \nabla \mathbf{F}(\mathbf{x})\|^2$. Consequently, the common descent direction can be defined as $\mathbf{d} = \boldsymbol{\lambda}^\top \nabla \mathbf{F}(\mathbf{x})$. Then, MGD follows the iterative update rule $\mathbf{x} \leftarrow \mathbf{x} - \eta \mathbf{d}$ in the hope that a Pareto-stationary point can be reached, where η signifies a learning rate. SMGD Liu and Vicente [2021] follows a similar approach, but with full multi-gradients being replaced by stochastic multi-gradients. For both MGD and SMGD, it has been shown that if $\|\boldsymbol{\lambda}^\top \nabla \mathbf{F}(\mathbf{x})\| = 0$ for some $\boldsymbol{\lambda} \in C$, where $C \triangleq \{\mathbf{y} \in [0,1]^S, \sum_{s \in [S]} y_s = 1\}$, then $\mathbf{x}$ is a Pareto stationary solution Fliege et al. [2019], Zhou et al. [2022].

Here, it is insightful to contrast vector-valued MOO with the linear scalarization method with fixed weights for MOO, which is also a relatively straightforward approach commonly seen in the MOO literature. We note that vector-valued MOO offers unique benefits that do not exist in linear scalarization. Specifically, MGD-type methods for vector-valued MOO dynamically calculate the weights for each objective based on the gradient information in each iteration. The dynamic weighting in MGD-type approach adapts much better to the landscapes of different MOO problems, which enables a much more flexible exploration on the Pareto front. In contrast, the linear scalarization method uses fixed or pre-defined weights for each objective. As a result, linear scalarization methods are limited to identifying the convex hull of the Pareto front [Boyd and Vandenberghe, 2004, Ehrgott, 2005], whereas (stochastic) multi-gradient methods, including our proposed VR-based algorithms, have the capability to uncover the Pareto front.

In this paper, we focus on MOO problems in two settings: (i) non-convex MOO and (ii) strongly convex MOO. Clearly, the non-convex setting is applicable to many learning problems in practice (e.g., neural network models). The strongly convex setting is also interesting due to many applications in practice (e.g., linear models with quadratic regularizations).

Next, to introduce the notion of sample complexity in MOO, we first need the following definitions for the non-convex and strongly convex settings, respectively.

Definition 3 (ϵ-Stationarity (Nonconvex Setting))**.** *A solution $\mathbf{x}$ is ϵ-stationary in MOO problem if the common descent direction at $\mathbf{x}$ satisfies the following condition: $\min_{\boldsymbol{\lambda} \in C} \mathbb{E}\|\boldsymbol{\lambda}^\top \nabla \mathbf{F}(\mathbf{x})\|^2 \leq \epsilon$ in non-convex MOO problems, where $C \triangleq \{\mathbf{y} \in [0,1]^S, \sum_{s \in [S]} y_s = 1\}$.*

Definition 4 (ϵ-Optimality (Strongly-Convex Setting))**.** *In the strongly-convex setting, a solution $\mathbf{x}$ is ϵ-optimal if $\mathbb{E}[\|\mathbf{x} - \mathbf{x}^*\|^2] \leq \epsilon$ in MOO problems, where $\mathbf{x}^*$ is a Pareto-optimal solution of Problem (1).*

With the above definitions, we are now in a position to define the concept of sample complexity in MOO as follows:

Definition 5 (Sample Complexity)**.** *The sample complexity in MOO is defined as the total number of incremental first-order oracle (IFO) calls required by a MOO algorithm to converge to an ϵ-stationary (or ϵ-optimal in the strongly* *convex setting) point, where one IFO call evaluates the multi-gradient $\nabla_{\mathbf{x}} f_{sj}(\mathbf{x};\xi_{sj})$ for all tasks s.*

3 THE STIMULUS ALGORITHM FAMILY

In this section, we first present the basic version of the STIMULUS algorithm in Section 3.1, which is followed by its momentum and adaptive-batching variants in Sections 3.2 and 3.3, respectively.

3.1 THE STIMULUS ALGORITHM

Our STIMULUS algorithm is presented in Algorithm 1, where we propose a new variance-reduced (VR) multi-gradient estimator. It can be seen from Algorithm 1 that our proposed VR approach has a double-loop structure, where the inner loop is of length $q > 0$. More specifically, different from MGD where a full multi-gradient direction $\mathbf{u}_t^s = \nabla f_s(\mathbf{x}_t), \forall s \in [S]$ is evaluated in all iterations, our STIMULUS algorithm only evaluates a full multi-gradient every q iterations (i.e., $\mathrm{mod}(t, q) = 0$). For all other iterations t with $\mathrm{mod}(t, q) \neq 0$, our STIMULUS algorithm uses a *stochastic* multi-gradient estimator $\mathbf{u}_t^s$ based on a mini-batch $\mathcal{A}$ with a recursive correction term as follows:

$$\mathbf{u}_t^s = \mathbf{u}_{t-1}^s + \frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}(\nabla f_{sj}(\mathbf{x}_t;\xi_{sj}) - \nabla f_{sj}(\mathbf{x}_{t-1};\xi_{sj})), \text{for all } s \in [S]. \quad (2)$$

Eq. (2) shows that the estimator is constructed iteratively based on information from $\mathbf{x}_{t-1}$ and $\mathbf{u}_{t-1}^s$, both of which are obtained from the previous update. We will show later in Section 4 that, thanks to the q-periodic full multi-gradients and the recursive correction terms, STIMULUS is able to achieve a convergence rate of $\mathcal{O}(1/T)$. Moreover, due to the stochastic subsampling in mini-batch $\mathcal{A}$, STIMULUS has a lower sample complexity than MGD. In STIMULUS, the update rule for parameters in $\mathbf{x}$ is written as: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{d}_t$, where η is the learning rate. Here, the direction $\mathbf{d}_t$ is defined as $\mathbf{d}_t := \sum_{s\in[S]}\lambda_t^s\mathbf{u}_t^s$, where the λ_t^s-values are obtained by solving the following quadratic optimization problem:

$$\min_{\lambda_t^s \geq 0}\Big\|\sum_{s\in[S]}\lambda_t^s\mathbf{u}_t^s\Big\|^2, \text{ s.t. } \sum_{s\in[S]}\lambda_t^s = 1. \quad (3)$$

The iterative update in Eqs. (3) follows the same token as in the MGDA algorithm [Mukai, 1980, Sener and Koltun, 2018, Lin et al., 2019, Fliege et al., 2019].

3.2 THE STIMULUS-M ALGORITHM

Although it can be shown that STIMULUS achieves a theoretical $\mathcal{O}(1/T)$ convergence rate, it could be sensitive to the choice of learning rate and suffer from similar oscillation issues in practice as gradient-descent-type methods do in single-objective optimization when some objectives are ill-conditioned.

Algorithm 1 STIMULUS algorithm and its variants.

```
Require: Initial point $\mathbf{x}_0$, parameters $T$, $q$.
Initialize: Choose $\mathbf{x}_0$.
for $t = 0, 1, \dots, T$ do
    if $\mathrm{mod}(t, q) = 0$ then
        if STIMULUS or STIMULUS-M then
            Compute: $\mathbf{u}_t^s = \frac{1}{n}\sum_{j=1}^n \nabla f_{sj}(\mathbf{x}_t;\xi_{sj}), \forall s \in [S]$.
        end if
        if STIMULUS$^+$ or STIMULUS-M$^+$ then
            Compute: $\mathbf{u}_t^s$ as in Eq. (5).
        end if
    else
        Compute $\mathbf{u}_t^s$ as in Eq. (2).
    end if
    Compute $\boldsymbol{\lambda}_t^* \in [0, 1]^S$ by solving Eq. (3).
    Compute: $\mathbf{d}_t = \sum_{s\in[S]}\lambda_t^{s,*}\mathbf{u}_t^s$.
    if STIMULUS or STIMULUS$^+$ then
        Update: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{d}_t$.
    end if
    if STIMULUS-M or STIMULUS-M$^+$ then
        Update: $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha(\mathbf{x}_t - \mathbf{x}_{t-1}) - \eta\mathbf{d}_t$.
    end if
end for
```

To further improve the empirical performance of STIMULUS, we now propose a momentum-assisted enhancement for STIMULUS called STIMULUS-M. The idea behind STIMULUS-M is to take into account the past trajectories to smooth the update direction. Specifically, in addition to the combined iterative update as in $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{d}_t$ and (3), the update rule in STIMULUS-M incorporates an α-parameterized momentum term as follows:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{d}_t + \underbrace{\alpha(\mathbf{x}_t - \mathbf{x}_{t-1})}_{\text{Momentum}}, \forall s \in [S], \quad (4)$$

where $\alpha \in (0, 1)$ is the momentum coefficient.

3.3 STIMULUS$^+$ /STIMULUS-M$^+$ ALGORITHMS

Note that in both STIMULUS and STIMULUS-M, one still needs to evaluate a full multi-gradient every q iteration, which remains computationally demanding in the large data regime. Moreover, if the objectives are in an expectation or "online" form rather than the finite-sum setting, it is infeasible to compute a full multi-gradient. To address these limitations, we propose two *adaptive-batching* enhanced versions for STIMULUS and STIMULUS-M called STIMULUS$^+$ and STIMULUS-M$^+$, respectively. Specifically, rather than using a q-periodic full multi-gradient $\mathbf{u}_t^s = \nabla f_s(\mathbf{x}_t) = \frac{1}{n}\sum_{j=1}^n \nabla f_{sj}(\mathbf{x}_t;\xi_{sj}), \forall s \in [S]$, in iteration t with $\mathrm{mod}(t, q) = 0$, we utilize an adaptive-batching stochastic multi-gradient as follows:

$$\mathbf{u}_t^s = \frac{1}{|\mathcal{N}_s|} \sum_{j \in \mathcal{N}_s} \nabla f_{sj}(\mathbf{x}_t; \xi_{sj}), \quad \forall s \in [S], \tag{5}$$

where $\mathcal{N}_s$ is an ϵ-adaptive batch sampled from the dataset uniformly at random with size:

$$|\mathcal{N}_s| = \min\left\{c_\gamma \sigma^2 \gamma_t^{-1}, c_\epsilon \sigma^2 \epsilon^{-1}, n\right\}. \tag{6}$$

We choose constants $c_\gamma \geq 8$, $c_\epsilon \geq \eta$ in non-convex case and $c_\gamma \geq \frac{8\mu}{\eta}, c_\epsilon \geq \frac{\mu}{2}$ in strongly-convex case (see detailed discussions in Section 4). The σ^2 represents the variance bound of stochastic gradient norms (cf. Assumption. 2). In STIMULUS$^+$, we choose $\gamma_{t+1} = \sum_{i=(n_k-1)q}^{t} \frac{\|\mathbf{d}_i\|^2}{q}$, while in the momentum based algorithm STIMULUS-M$^+$, we choose $\gamma_{t+1} = \sum_{i=(n_k-1)q}^{t} \|\alpha^{(t-i)}\mathbf{d}_i\|^2/q$. The term γ_{t+1} offers further refinement to improve convergence.

4 PARETO STATIONARITY CONVERGENCE ANALYSIS

In this section, we theoretically analyze the Pareto stationarity convergence of our STIMULUS algorithms in non-convex and strongly convex settings, beginning with two necessary assumptions.

Assumption 1 (L-Lipschitz Smoothness). *There exists a constant $L > 0$ such that $\|\nabla f_s(\mathbf{x}) - \nabla f_s(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\forall s \in [S]$.*

Assumption 2 (Bounded Variance). *There exists a constant $\sigma > 0$ such that for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbb{E}\|\nabla_{\mathbf{x}} f_s(\mathbf{x}; \xi) - \nabla_{\mathbf{x}} f_s(\mathbf{x})\|^2 \leq \sigma^2$, $\forall s \in S$.*

With these assumptions, we are now in a position to discuss the Pareto stationary convergence of the STIMULUS family.

4.1 PARETO-STATIONARITY CONVERGENCE OF STIMULUS

1) STIMULUS**: The Non-convex Setting.** First, we show that the basic STIMULUS algorithm achieves an $\mathcal{O}(1/T)$ convergence rate for non-convex MOO problems in the following theorem. Note that this result matches that of the deterministic MGD method.

Theorem 1 (STIMULUS for Non-convex MOO). *Under Assumption 1, let $\eta \leq \frac{1}{2L}$, if at least one objective function $f_s(\cdot)$, $s \in [S]$ is bounded from below by $f_s^{\min}$, then the sequence $\{\mathbf{x}_t\}$ output by* STIMULUS *satisfies: $\frac{1}{T}\sum_{t=0}^{T-1} \min_{\boldsymbol{\lambda} \in C} \mathbb{E}\|\boldsymbol{\lambda}^\top \nabla \mathbf{F}(\mathbf{x}_t)\|^2 = \mathcal{O}(1/T)$.*

Following from Theorem. 1, we immediately have the following sample complexity for the STIMULUS algorithm by choosing $q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$:

Corollary 1 (Sample Complexity of STIMULUS). *By choosing $\eta \leq \frac{1}{2L}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$, the overall sample complexity of* STIMULUS *for finding an ϵ-stationary point for non-convex MOO problems is $\mathcal{O}\left(\sqrt{n}\epsilon^{-1} + n\right)$.*

Several interesting remarks regarding Theorem 1 and Corollary 1 are in order: **1)** Our proof of STIMULUS's Pareto-stationarity convergence only relies on standard assumptions commonly used in first-order optimization techniques. This is in stark contrast to prior research, where unconventional and hard-to-verify assumptions were required (e.g., an assumption on the convergence of $\mathbf{x}$-sequence is used in Fliege et al. [2019]). **2)** While both MGD and our methods share the same $\mathcal{O}(1/T)$ convergence rate, STIMULUS enjoys a substantially lower sample complexity than MGD. More specifically, the sample complexity of STIMULUS is reduced by a factor of $\sqrt{n}$ when compared to MGD. This becomes particularly advantageous in the "big data" regime where n is large.

2) STIMULUS**: The Strongly Convex Setting.** Now, we consider the strongly convex setting, which is more tractable but still of interest in many learning problems in practice (e.g., multi-objective ridge regression). In the strongly convex setting, we have the following additional assumption:

Assumption 3 (μ-Strongly Convex Function). *Each objective $f_s(\mathbf{x})$, $s \in [S]$ is a μ-strongly convex function, i.e., $f_s(\mathbf{y}) \geq f_s(\mathbf{x}) + \nabla f_s(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$, $\forall \mathbf{x}, \mathbf{y}$, for some $\mu > 0$.*

Assumption 4. *For any objective function f_j, there exists a positive real number c_j such that for any $\mathbf{x}$ in $\mathbb{R}^n$ the following relation holds $f_j(\mathbf{x}) - f_j(\mathbf{x}^*) \geq \frac{c_j}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$ a.s. ;$j \in S$.*

Assumption 4 asserts that the function value increases at least quadratically as you move away from $\mathbf{x}_*$, ensuring consistent progress towards the optimum. It is a reasonable assumption since it is also based on the strong convexity property. The above assumption has also been adopted in Mercier et al. [2018].

For strongly convex MOO problems, the next result says that STIMULUS achieves a much stronger expected linear Pareto-optimality convergence performance:

Theorem 2 (STIMULUS for μ-Strongly Convex MOO). *Under Assumption 1, 3, 4, let $\eta \leq \min\{\frac{1}{2}, \frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}$, $q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$. Under Assumptions 1–4, pick $\mathbf{x}_t$ as the final output of* STIMULUS *with probability $w_t = (1 - \frac{3\mu\eta}{4})^{1-t}$. Then, we have $\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|^2\mu\exp(-\frac{3\eta\mu T}{4})$.*

Further, Theorem 2 immediately implies following with logarithmic sample complexity (in terms of ϵ) STIMULUS with a proper choice of learning rate and $q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$.

Corollary 2 (Sample Complexity of STIMULUS)**.** *By choosing $\eta \le \min\{\frac{1}{2}, \frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil\}$, the overall sample complexity of* STIMULUS *for solving strongly convex MOO is $\mathcal{O}\left(n + \sqrt{n}\ln(\mu/\epsilon)\right)$.*

There are also several interesting insights from Theorem 2 and Corollary 2 regarding STIMULUS's performance for solving strongly convex MOO problems: **1)** STIMULUS achieves an expected linear convergence rate of $\mathcal{O}(\mu\exp(-\mu T))$. Interestingly, this convergence rate matches that of MGD for strongly convex MOO problems as well as gradient descent for strongly convex single-objective optimization. **2)** Another interesting feature of STIMULUS for strongly convex MOO stems from its use of randomly selected outputs $\mathbf{x}_t$ along with associated weights w_t from the trajectory of $\mathbf{x}_t$, which is inspired by the similar idea for stochastic gradient descent (SGD) [Ghadimi and Lan, 2013]. Note that, for implementation in practice, one does not need to store all $\mathbf{x}_t$-values. Instead, the algorithm can be implemented by using a random clock for stopping [Ghadimi and Lan, 2013].

4.2 PARETO STATIONARITY CONVERGENCE OF STIMULUS-M

Next, we turn our attention to the Pareto stationarity convergence of the STIMULUS-M algorithm. Again, we analyze STIMULUS-M in non-convex and strongly convex settings:

Theorem 3 (STIMULUS-M for Non-convex MOO)**.** *Let $\eta_t = \eta \le \min\{\frac{1}{2L}, \frac{1}{2}\}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$. Under Assumptions 1, if at least one objective function $f_s(\cdot)$, $s \in [S]$, is bounded from below by $f_s^{\min}$, then the sequence $\{\mathbf{x}_t\}$ output by* STIMULUS-M *satisfies $\frac{1}{T}\sum_{t=0}^{T-1}\min_{\boldsymbol{\lambda}\in C}\mathbb{E}\|\boldsymbol{\lambda}^\top\nabla\mathbf{F}(\mathbf{x}_t)\|^2 = \mathcal{O}(\frac{1}{T})$.*

Similar to the basic STIMULUS algorithm, by choosing the appropriate learning rate and inner loop length parameters, we immediately have the following sample complexity result for STIMULUS-M for solving non-convex MOO problems:

Corollary 3 (Sample Complexity of STIMULUS-M)**.** *By choosing $\eta_t = \eta \le \min\{\frac{1}{2L}, \frac{1}{2}\}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$. The overall sample complexity of* STIMULUS-M *under non-convex objective functions is $\mathcal{O}\left(\sqrt{n}\epsilon^{-1} + n\right)$.*

The next two results state the Pareto optimality and sample complexity results for STIMULUS-M:

Theorem 4 (STIMULUS-M for μ-Strongly Convex MOO)**.** *Let $\eta \le \min\{\frac{1}{2}, \frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$. Under Assumption 1, 3, 4, pick $\mathbf{x}_t$ as the final output of* STIMULUS-M *with probability $w_t = (1 - \frac{3\mu\eta}{4})^{1-t}$. Then, we have $\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 \le \|\mathbf{x}_0 - \mathbf{x}_*\|^2\mu\exp(-\frac{3\eta\mu T}{4})$.*

Corollary 4 (Sample Complexity of STIMULUS-M)**.** *By choosing $\eta \le \min\{\frac{1}{2}, \frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$, the overall sample complexity of* STIMULUS-M *for solving strongly convex MOO is $\mathcal{O}\left(n + \sqrt{n}\ln(\mu/\epsilon)\right)$.*

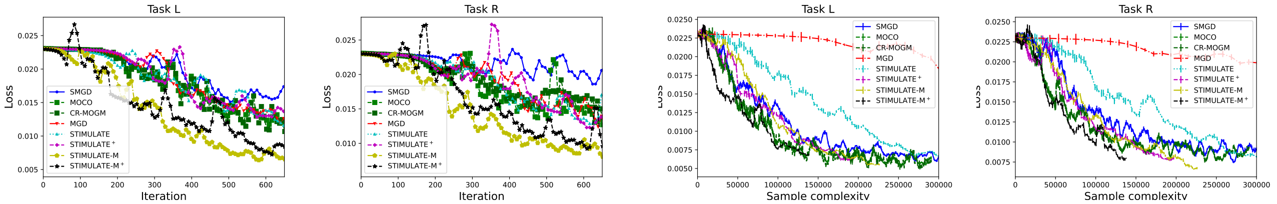
We remark that the convergence rate upper bound of STIMULUS-M is the same as that in Theorem 2, which suggests a potentially loose convergence upper bound in Theorem 4 due to the technicality and intricacies in analyzing momentum-based stochastic multi-gradient algorithms for solving non-convex MOO problems. Yet, we note that even this potentially loose convergence rate upper bound in Theorem 4 already suffices to establish a linear convergence rate for STIMULUS-M in solving strongly convex MOO problems. Moreover, we will show later in Section 5 that this momentum-assisted method significantly accelerates the empirical convergence speed performance. It is also worth noting that there are two key differences in the proofs of Theorem 3 and 4 compared to those of the momentum-based stochastic gradient algorithm for single-objective non-convex optimization: 1) our proof exploits the martingale structure of the $\mathbf{u}_t^s$. This enables us to tightly bound the mean-square error term $\mathbb{E}\left\|\nabla f_s\left(\mathbf{x}_t\right) - \mathbf{u}_t^s\right\|^2$ under the momentum scheme. In contrast, in the traditional analysis of stochastic algorithms with momentum, this error term corresponds to the variance of the stochastic estimator and is typically assumed to be bounded by a universal constant. 2) Our proof requires careful manipulation of the bounding strategy to effectively handle the accumulation of the mean-square error $\mathbb{E}\left\|\nabla f_s\left(\mathbf{x}_k\right) - \mathbf{u}_t^s\right\|^2$ over the entire optimization trajectory in non-convex MOO.

4.3 PARETO STATIONARITY CONVERGENCE RESULTS OF STIMULUS$^+$ AND STIMULUS-M$^+$

Next, we present the Pareto stationarity convergence and the associated sample complexity results of the STIMULUS$^+$ /STIMULUS-M$^+$ algorithms for non-convex MOO as follows:

Theorem 5 (STIMULUS$^+$ /STIMULUS-M$^+$)**.** *Let $\eta \le \min\{\frac{1}{4L}, \frac{1}{2}\}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$. By choosing c_γ and c_ϵ as such that $c_\gamma \ge 8$, and $c_\epsilon \ge \eta$, under Assumptions 1 and 2, if at least one function $f_s(\cdot)$, $s \in [S]$ is bounded from below by $f_s^{\min}$, then the sequence $\{\mathbf{x}_t\}$ output by* STIMULUS$^+$ /STIMULUS-M$^+$ *satisfies: $\frac{1}{T}\sum_{t=0}^{T-1}\min_{\boldsymbol{\lambda}\in C}\mathbb{E}\|\boldsymbol{\lambda}^\top\nabla\mathbf{F}(\mathbf{x}_t)\|^2 = \mathcal{O}(\frac{1}{T})$.*

Corollary 5 (Sample Complexity)**.** *By choosing $\eta \le \min\{\frac{1}{4L}, \frac{1}{2}\}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$, $c_\gamma \ge 8$, and $c_\epsilon \ge \eta$. The overall sample complexity of* STIMULUS$^+$ / STIMULUS-M$^+$ *under non-convex objective functions is $\mathcal{O}\left(\sqrt{n}\epsilon^{-1} + n\right)$.*



(a) Training loss convergence in terms of iterations.

(b) Training loss convergence in terms of samples.

Figure 1: Training loss convergence comparisons between different MOO algorithms.

Theorem 6 (STIMULUS$^+$ /STIMULUS-M$^+$). *Let $\eta \le \min\{\frac{1}{2}, \frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}, c_\gamma \ge \frac{8\mu}{\eta}, c_\epsilon \ge \frac{\mu}{2}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$. Under Assumptions 1- 4, pick $\mathbf{x}_t$ as the final output of the STIMULUS$^+$ /STIMULUS-M$^+$ algorithm with weights $w_t = (1 - \frac{3\mu\eta}{4})^{1-t}$. Then, it holds that $\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 \le \|\mathbf{x}_0 - \mathbf{x}_*\|^2 \mu \exp(-\frac{3\eta\mu T}{4})$.*

Corollary 6 (Sample Complexity). *By choosing $\eta \le \min\{\frac{1}{2}, \frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}, c_\gamma \ge \frac{8\mu}{\eta}, c_\epsilon \ge \frac{\mu}{2}, q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$, the overall sample complexity of STIMULUS$^+$ / STIMULUS-M$^+$ for solving strongly-convex MOO is $\mathcal{O}\left(n + \sqrt{n}\ln(\mu/\epsilon)\right)$.*

We note that, although the theoretical sample complexity bounds of STIMULUS$^+$ / STIMULUS-M$^+$ are the same as those of STIMULUS/ STIMULUS-M, respectively, the fact that STIMULUS$^+$ and STIMULUS-M$^+$ do not need full multi-gradient evaluations implies that STIMULUS/ STIMULUS-M use significantly fewer samples than STIMULUS/ STIMULUS-M in the large dataset regime. Our experimental results in the next section will also empirically confirm this.

5 EXPERIMENTAL RESULTS

In this section, we conduct numerical experiments to validate our STIMULUS algorithm family, focusing on nonconvex MOO problems, while results for strongly convex and 8-objective MOO experiments are in the appendix.

1) Two-Objective Experiments on the MultiMNIST Dataset: First, we test the convergence performance of our STIMULUS using the "MultiMNIST" dataset [Sabour et al., 2017], which is a multi-task learning version of the MNIST dataset [LeCun et al., 1998] from LIBSVM repository. Specifically, MultiMNIST converts the hand-written classification problem in MNIST into a two-task problem, where the two tasks are task "L" (to categorize the top-left digit) and task "R" (to classify the bottom-right digit). The goal is to classify the images of different tasks. We compare our STIMULUS algorithms with MGD, SMGD, CR-MOGM, and MOCO. All algorithms use the same randomly generated initial point. The learning rates are chosen as $\eta = 0.3, \alpha = 0.5$, constant $c = c_\gamma = c_\epsilon = 32$ and solution accuracy $\epsilon = 10^{-3}$. The batch-size for MOCO, CR-MOGM and SMGD is 96. The full batch size for MGD is 1024, and the inner loop batch-size $|\mathcal{N}_s|$ for STIMULUS, STIMULUS-M, STIMULUS$^+$, STIMULUS-M$^+$is 96. As shown in Fig. 1(a), SMGD exhibits the slowest convergence speed, while MOCO has a slightly faster convergence. MGD and our STIMULUS algorithms have comparable performances. The STIMULUS-M /STIMULUS-M$^+$ algorithms converge faster than MGD, STIMULUS , and STIMULUS$^+$, primarily due to the use of momentum. Fig. 1(b) highlights differences in sample complexity. MGD suffers the highest sample complexity, while STIMULUS$^+$ and STIMULUS-M$^+$ demonstrate a more efficient utilization of samples in comparison to STIMULUS and STIMULUS-M. These results are consistent with our theoretical analyses as outlined in Theorems 1, 3, and 5.

2) 40-Objective Experiments with the CelebA Dataset:

Lastly, we conduct large-scale 40-objective experiments with the CelebA dataset [Liu et al., 2015], which contains 200K facial images annotated with 40 attributes. Each attribute corresponds to a binary classification task, resulting in a 40-objective problem.

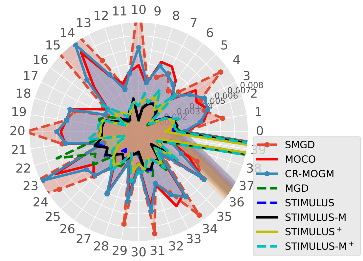


Figure 2: Training loss convergence comparison (40-task).

We use a ResNet-18 He et al. [2016] model without the final layer for each attribute, and we attach a linear layer to each attribute for classification. In this experiment, we set $\eta = 0.0005, \alpha = 0.01$, the full batch size for MGD is 1024, and the batch size for SMGD, CR-MOGM and MOCO and the inner loop batch size $|\mathcal{N}_s|$ for STIMULUS, STIMULUS-M, STIMULUS$^+$, STIMULUS-M$^+$is 32. As shown in Fig. 2, MGD, STIMULUS, STIMULUS-M, STIMULUS$^+$, and STIMULUS-M$^+$significantly outperform SMGD, CR-

MOGM and MOCO in terms of training loss. Also, we would like to note that STIMULUS$^+$ and STIMULUS-M$^+$ consume fewer sample (approximately 11,000) samples compared to STIMULUS and STIMULUS-M , which consume approximately 13,120 samples, and MGD, which consumes roughly 102,400 samples. These results are consistent with our theoretical results in Theorems 1, 3, and 5.

6 CONCLUSION

In this paper, we proposed STIMULUS, a new variance-reduction-based stochastic multi-gradient-based algorithm to achieve fast convergence and low sample complexity multi-objective optimization (MOO). We analyze its Pareto stationarity convergence and sample complexity under non-convex and strongly convex settings. To enhance empirical convergence, we propose STIMULUS-M , which incorporates momentum. To reduce the periodic full multi-gradient evaluation in STIMULUS and STIMULUS-M, we introduce adaptive batching versions, STIMULUS$^+$ /STIMULUS-M$^+$, with theoretical performance analysis. Overall, our STIMULUS algorithm family advances MOO algorithm design and analysis.

7 ACKNOWLEDGEMENT

This work is supported in part by NSF grants CAREER CNS-2110259, CNS-2112471, IIS-2324052, DARPA YFA D24AP00265, DARPA HR0011-25-2-0019, ONR grant N00014-24-1-2729, and AFRL grant PGSC-SC-111374-19s.

References

Syrine Belakaria, Aryan Deshwal, Nitthilan Kannappan Jayakodi, and Janardhan Rao Doppa. Uncertainty-aware search framework for multi-objective bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10044–10052, 2020.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Sebastian Bruch, Xuanhui Wang, Michael Bendersky, and Marc Najork. An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pages 75–78, 2019.

Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. Co-attentive multi-task learning for explainable recommendation. In *IJCAI*, pages 2137–2143, 2019.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

Matthias Ehrgott. *Multicriteria optimization*, volume 491. Springer Science & Business Media, 2005.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Heshan Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent stochastic approach. *arXiv preprint arXiv:2210.12624*, 2022.

Heshan Fernando, Lisha Chen, Songtao Lu, Pin-Yu Chen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, Gaowen Liu, Meng Wang, and Tianyi Chen. Variance reduction can improve trade-off in multi-objective learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6975–6979, 2024. doi: 10.1109/ICASSP48485.2024.10446038.

Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical methods of operations research*, 51(3):479–494, 2000.

Jörg Fliege, A Ismael F Vaz, and Luís Nunes Vicente. Complexity of gradient descent for multiobjective optimization. *Optimization Methods and Software*, 34(5):949–959, 2019.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. cvpr. 2016. *arXiv preprint arXiv:1512.03385*, 2016.

Yang Jiao, Ning Xie, Yan Gao, Chien-Chih Wang, and Yi Sun. Fine-grained fashion representation learning by online deep clustering. In *European Conference on Computer Vision*, pages 19–35. Springer, 2022.

Yang Jiao, Yan Gao, Jingjing Meng, Jin Shang, and Yi Sun. Learning attribute and class-specific representation duet for fine-grained fashion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2023.

Bingzhong Jing, Tao Zhang, Zixian Wang, Ying Jin, Kuiyuan Liu, Wenze Qiu, Liangru Ke, Ying Sun, Caisheng He, Dan Hou, et al. A deep survival analysis method based on ranking. *Artificial intelligence in medicine*, 98:1–9, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 13–18 Jul 2020.

Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. Deep attentive ranking networks for learning to order sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8115–8122, 2020.

Marco Laumanns and Jiri Ocenasek. Bayesian optimization algorithms for multi-objective optimization. In *International Conference on Parallel Problem Solving from Nature*, pages 298–307. Springer, 2002.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *Available: http://yann. lecun. com/exdb/mnist*, 1998.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.

Suyun Liu and Luis Nunes Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. *Annals of Operations Research*, pages 1–30, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Ze Lyu, Yu Dong, Chengfu Huo, and Weijun Ren. Deep match to rank model for personalized click-through rate prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 156–163, 2020.

Debabrata Mahapatra, Chaosheng Dong, Yetian Chen, and Michinari Momma. Multi-label learning to rank through multi-objective optimization. In *Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023a.

Debabrata Mahapatra, Chaosheng Dong, and Michinari Momma. Querywise fair learning to rank through multi-objective optimization. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023b.

Quentin Mercier, Fabrice Poirion, and Jean-Antoine Désidéri. A stochastic multiple gradient descent algorithm. *European Journal of Operational Research*, 271 (3):808–817, 2018.

Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 2012.

Hiroaki Mukai. Algorithms for multicriterion optimization. *IEEE transactions on automatic control*, 25(2):177–186, 1980.

Lin Nie, Keze Wang, Wenxiong Kang, and Yuefang Gao. Image retrieval with attribute-associated auxiliary references. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2017.

Simone Parisi, Matteo Pirotta, Nicola Smacchia, Luca Bascetta, and Marcello Restelli. Policy gradient approaches for multi-objective sequential decision making. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 2323–2330. IEEE, 2014.

Ramakanth Pasunuru and Mohit Bansal. Multi-task video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

Jerome Revaud, Jon Almazan, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Zhitao Wang, Yong Zhou, Litao Hong, Yuanhang Zou, Hanjing Su, and Shouzhi Chen. Pairwise learning for neural link prediction. *arXiv preprint arXiv:2112.02936*, 2021.

Zhibo Xiao, Luwei Yang, Wen Jiang, Yi Wei, Yi Hu, and Hao Wang. Deep multi-interest network for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2265–2268, 2020.

Yi-Che Yang, Ping-Ching Lai, and Hung-Hsuan Chen. Empirically testing deep and shallow ranking models for click-through rate (ctr) prediction. In *2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 147–152. IEEE, 2020.

Yijun Yang, Jing Jiang, Tianyi Zhou, Jie Ma, and Yuhui Shi. Pareto policy pool for model-based offline reinforcement learning. In *International Conference on Learning Representations*, 2022.

Hai-Tao Yu, Adam Jatowt, Hideo Joho, Joemon M Jose, Xiao Yang, and Long Chen. Wassrank: Listwise document ranking using optimal transport theory. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 24–32, 2019.

Qingfu Zhang and Hui Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation*, 11(6):712–731, 2007.

Shiji Zhou, Wenpeng Zhang, Jiyan Jiang, Wenliang Zhong, Jinjie GU, and Wenwu Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Tianchen Zhou, Michinari Momma, Chaosheng Dong, Fan Yang, Chenghuan Guo, Jin Shang, and Jia Kevin Liu. Multi-task learning on heterogeneous graph neural network for substitute recommendation. In *19th International Workshop on Mining and Learning with Graphs*, 2023.

A PROOF OF CONVERGENCE OF STIMULUS

Table 2: List of key notation.

Notation	Definition
n	Total number of samples per task
s	Objective/task index
S	Total number of objectives/tasks
t	Iteration number index
T	Total number of iterations
$\mathbf{x} \in \mathbb{R}^d$	Model parameters in Problem (1)
$\mathbf{x}_* \in \mathbb{R}^d$	A pareto optimal solution of Problem (1)
η	The learning rate
α	The momentum constant
ϵ	Stationarity error in Def. 3
μ	Strongly-convex constant in Assumption 3

For clarity of notation, we drop $*$ for λ, that is, we use λ_t^s to represent the solution of quadratic problem for task s in the t-th round.

Lemma 1. *Let Assumption 1 hold. The gradient estimator $\mathbf{u}_t^s$ satisfies for all $(n_t-1)q+1 \le t \le n_t q - 1$:*

$$\mathbb{E}_t\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 \le \frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t} \mathbb{E}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \mathbb{E}_t\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}^s_{(n_t-1)q}\|^2. \tag{7}$$

Proof of Lemma. 1.

Proof. From Lemma 1 in Fang et al. [2018], we have

$$\begin{aligned}\mathbb{E}_t\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 &\overset{(a)}{=} \mathbb{E}_t\|\nabla f_s(\mathbf{x}_{t-1}) - \mathbf{u}_{t-1}^s\|^2 \\ &+\mathbb{E}_t\|\frac{1}{|\mathcal{A}|}\sum_{j\in\mathcal{A}}\left(\nabla f_{sj}(\mathbf{x}_t;\xi_{sj}) - \nabla f_{sj}(\mathbf{x}_{t-1};\xi_{sj}) + \nabla f_s(\mathbf{x}_{t-1}) - \nabla f_s(\mathbf{x}_t)\right)\|^2 \\ &\overset{(b)}{\le}\mathbb{E}_t\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}^s_{(n_t-1)q}\|^2 + L^2\sum_{i=(n_t-1)q}^{t}\frac{1}{|\mathcal{A}|}\mathbb{E}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2.\end{aligned} \tag{8}$$

(a) stems from Proposition 1 in Fang et al. [2018], where the expectation of the gradient difference is broken down. (b) leverages Eq. (2.3) from Fang et al. [2018], applying a bound based on the Lipschitz continuity of the gradient.

Telescoping over from $(n_t-1)\,q+1$ to t, where $t \le n_t q - 1$, we obtain that

$$\mathbb{E}_t\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 \le \mathbb{E}_t\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}^s_{(n_t-1)q}\|^2 + L^2\sum_{i=(n_t-1)q}^{t}\frac{1}{|\mathcal{A}|}\mathbb{E}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 \tag{9}$$

Then, we have

$$\mathbb{E}_t\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 \le \frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t} \mathbb{E}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \mathbb{E}_t\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}^s_{(n_t-1)q}\|^2. \tag{10}$$

□

Lemma 2. *For general L-smooth functions $\{f_s, s \in [S]\}$, choose the learning rate η s.t. $\eta \le \frac{1}{2L}$, the update $\mathbf{d}_t$ of the algorithm satisfies:*

$$f_s(\mathbf{x}_{t+1}) \le f_s(\mathbf{x}_t) + \frac{\eta}{2}\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 - \frac{\eta}{4}\|\mathbf{d}_t\|^2. \tag{11}$$

Proof of Lemma. 2.

Proof.

$$
\begin{aligned}
f_s(\mathbf{x}_{t+1}) &\overset{(a)}{\le} f_s(\mathbf{x}_t) + \langle \nabla f_s(\mathbf{x}_t), -\eta \mathbf{d}_t \rangle + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
&= f_s(\mathbf{x}_t) - \eta \left\langle \nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s, \mathbf{d}_t \right\rangle - \eta \left\langle \mathbf{u}_t^s, \mathbf{d}_t \right\rangle + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
&\overset{(b)}{\le} f_s(\mathbf{x}_t) - \eta \left\langle \nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s, \mathbf{d}_t \right\rangle - \eta \|\mathbf{d}_t\|^2 + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
&\overset{(c)}{\le} f_s(\mathbf{x}_t) + \frac{\eta}{2} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{1}{2}\eta \|\mathbf{d}_t\|^2 - \eta \|\mathbf{d}_t\|^2 + \frac{1}{2} L \eta^2 \|\mathbf{d}_t\|^2 \\
&= f_s(\mathbf{x}_t) + \frac{\eta}{2} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 - \eta \left(\frac{1}{2} - \frac{1}{2} L \eta \right) \|\mathbf{d}_t\|^2. \qquad (12)
\end{aligned}
$$

(a) follows from the objective function f_s is L-smooth. (b) follows from $\langle \mathbf{u}_t^s, \mathbf{d}_t \rangle \ge \|\mathbf{d}_t\|^2$ since $\mathbf{d}_t$ is a general solution in the convex hull of the family of vectors $\{\mathbf{u}_t^s, s \in [S]\}$ (see Lemma 2.1 Désidéri [2012]). (c) follows from the triangle inequality.

By setting $\left(\frac{1}{2} - \frac{L}{2}\eta\right) \ge \frac{1}{4}$, that is, $\eta \le \frac{1}{2L}$, we have

$$
f_s(\mathbf{x}_{t+1}) \le f_s(\mathbf{x}_t) + \frac{\eta}{2} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 - \frac{\eta}{4} \|\mathbf{d}_t\|^2. \qquad (13)
$$

□

Proof of Theorem. 1

Proof. Taking expectation on both sides of the inequality in Lemma. 2, we have

$$
\begin{aligned}
&\mathbb{E}[f_s(\mathbf{x}_{t+1})] \overset{(a)}{\le} \mathbb{E}[f_s(\mathbf{x}_t)] + \frac{\eta}{2} \mathbb{E} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 - \frac{\eta}{4} \mathbb{E} \|\mathbf{d}_t\|^2 \\
&\overset{(b)}{\le} \mathbb{E}[f_s(\mathbf{x}_t)] - \frac{\eta}{4} \mathbb{E} \|\mathbf{d}_t\|^2 + \mathbb{E} \frac{\eta}{2} [\frac{L^2}{|\mathcal{A}|} \sum_{i=(n_t-1)q}^{t} \mathbb{E} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \mathbb{E} \|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}_{(n_t-1)q}^s\|^2] \\
&\overset{(c)}{=} \mathbb{E}[f_s(\mathbf{x}_t)] - \frac{\eta}{4} \mathbb{E} \|\mathbf{d}_t\|^2 + \frac{\eta}{2} [\frac{L^2}{|\mathcal{A}|} \sum_{i=(n_t-1)q}^{t} \eta^2 \mathbb{E} \|\mathbf{d}_i\|^2]. \qquad (14)
\end{aligned}
$$

(a) follows from Lemma. 2. (b) follows from the Lemma. 1. (c) follows from the update rule of $\mathbf{x}$ as shown in Eq. (5) and $\mathbb{E}\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}_{(n_t-1)q}^s\|^2 = 0$ as shown in Line 5 in our Algorithm. 1.

Next, telescoping the above inequality over t from $(n_t - 1)\, q$ to t where $t \le n_t q - 1$ and noting that for $(n_t - 1)\, q \le j \le n_t q - 1, n_j = n_t$, we obtain

$$
\begin{aligned}
&\mathbb{E}[f_s(\mathbf{x}_{t+1})] \\
&\le \mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{4} \sum_{j=(n_t-1)q}^{t} \mathbb{E} \|\mathbf{d}_j\|^2 + \frac{\eta}{2} [\frac{L^2}{|\mathcal{A}|} \sum_{j=(n_t-1)q}^{t} \sum_{i=(n_t-1)q}^{j} \eta^2 \mathbb{E} \|\mathbf{d}_i\|^2] \\
&\overset{(a)}{\le} \mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{4} \sum_{j=(n_t-1)q}^{t} \mathbb{E} \|\mathbf{d}_j\|^2 + \frac{\eta}{2} [\frac{L^2}{|\mathcal{A}|} \sum_{j=(n_t-1)q}^{t} \sum_{i=(n_t-1)q}^{t} \eta^2 \mathbb{E} \|\mathbf{d}_i\|^2] \\
&\overset{(b)}{\le} \mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{4} \sum_{j=(n_t-1)q}^{t} \mathbb{E} \|\mathbf{d}_j\|^2 + \frac{\eta^3 q}{2} [\frac{L^2}{|\mathcal{A}|} \sum_{j=(n_t-1)q}^{t} \mathbb{E} \|\mathbf{d}_j\|^2]
\end{aligned}
$$

$$= \mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - [\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|}] \sum_{j=(n_t-1)q}^{t} \mathbb{E}\|\mathbf{d}_j\|^2. \tag{15}$$

where (a) extends the summation of the third term from j to t, (b) follows from the fact that $t \leq n_t q - 1$.

We continue the proof by further driving

$$\begin{aligned}
&\mathbb{E}[f_s(\mathbf{x}_T)] - \mathbb{E}[f_s(\mathbf{x}_0)] \\
&= (\mathbb{E}[f_s(\mathbf{x}_q)] - \mathbb{E}[f_s(\mathbf{x}_0)]) + (\mathbb{E}[f_s(\mathbf{x}_{2q})] - \mathbb{E}[f_s(\mathbf{x}_q)]) + \cdot + (\mathbb{E}[f_s(\mathbf{x}_T)] - \mathbb{E}[f_s(\mathbf{x}_{(n_T-1)q})]) \\
&\leq -[\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|}] \sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{d}_t\|^2
\end{aligned} \tag{16}$$

Note that $\mathbb{E}[f_s\ (\mathbf{x}_{T+1})] \geq f_s^* \triangleq \inf_{\mathbf{x}\in\mathbb{R}^d} f_s(\mathbf{x})$. Hence, we have

$$[\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|}] \sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{d}_t\|^2 \leq [[f_s(\mathbf{x}_0)] - [f_s(\mathbf{x}_T)]] \leq [[f_s(\mathbf{x}_0)] - f_s^*]. \tag{17}$$

Based on the parameter setting $q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$, we have

$$[\frac{\eta}{4} - \frac{\eta^3 L^2}{2}] \sum_{t=0}^{T-1} \|\mathbf{d}_t\|^2 \leq [[f_s(\mathbf{x}_0)] - f_s^*]. \tag{18}$$

Thus, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{d}_t\|^2 \leq \frac{[[f_s(\mathbf{x}_0)] - f_s^*]}{[\frac{\eta}{4} - \frac{\eta^3 L^2}{2}]T}. \tag{19}$$

Since $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|d_t\|^2$ is just common descent directions. According to Definition. 3 shown in the paper, the quantity to our interest is $\|\sum_{s\in[S]} \lambda_t^s \nabla f(\mathbf{x})\|^2$.

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\sum_{s\in[S]} \lambda_t^s \nabla f_s(\mathbf{x}_t)\|^2 \\
&\overset{(a)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1} 2\mathbb{E}\|\sum_{s\in[S]} \lambda_t^s \nabla f_s(\mathbf{x}_t) - \sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s\|^2 + \frac{1}{T}\sum_{t=0}^{T-1} 2\mathbb{E}\|\sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s\|^2 \\
&\overset{(b)}{=} \frac{1}{T}\sum_{t=0}^{T-1} 2\mathbb{E}\|\sum_{s\in[S]} \lambda_t^s (\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s)\|^2 + \frac{1}{T}\sum_{t=0}^{T-1} 2\mathbb{E}\|\mathbf{d}_t\|^2 \\
&\overset{(c)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1} 2S \sum_{s\in[S]} (\lambda_t^s)^2 \mathbb{E}\|(\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s)\|^2 + \frac{1}{T}\sum_{t=0}^{T-1} 2\mathbb{E}\|\mathbf{d}_t\|^2 \\
&\overset{(d)}{\leq} \frac{1}{T}\sum_{t=0}^{T-1} 2S \sum_{s\in[S]} (\lambda_t^s)^2 [\mathbb{E}_t\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}_{(n_t-1)q}^s\|^2 + L^2 \sum_{i=(n_t-1)q}^{t} \frac{1}{|\mathcal{A}|}\mathbb{E}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2] + \frac{1}{T}\sum_{t=0}^{T-1} 2\mathbb{E}\|\mathbf{d}_t\|^2 \\
&= \frac{1}{T}\sum_{t=0}^{T-1} 2S \sum_{s\in[S]} (\lambda_t^s)^2 [\mathbb{E}_t\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}_{(n_t-1)q}^s\|^2] \\
&\quad + 2SL^2 \frac{1}{T}\sum_{t=0}^{T-1} \sum_{i=(n_t-1)q}^{t} \frac{1}{|\mathcal{A}|}\mathbb{E}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \frac{1}{T}\sum_{t=0}^{T-1} 2\mathbb{E}\|\mathbf{d}_t\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(e)}{\leq}\frac{1}{T}\sum_{t=0}^{T-1}2S\sum_{s\in[S]}(\lambda_t^s)^2[\mathbb{E}_t\|\nabla f_s(\mathbf{x}_{(n_t-1)q})-\mathbf{u}^s_{(n_t-1)q}\|^2]\\
&\quad+2SL^2\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=(n_t-1)q}^{n_tq-1}\frac{1}{|\mathcal{A}|}\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+\frac{1}{T}\sum_{t=0}^{T-1}2\mathbb{E}\|\mathbf{d}_t\|^2\\
&=2SL^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{q}{|\mathcal{A}|}\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+\frac{1}{T}\sum_{t=0}^{T-1}2\mathbb{E}\|\mathbf{d}_t\|^2\\
&\overset{(f)}{=}2SL^2\eta^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2+\frac{1}{T}\sum_{t=0}^{T-1}2\mathbb{E}\|\mathbf{d}_t\|^2\\
&=(2SL^2\eta^2+2)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2
\end{aligned}
\tag{20}
$$

where (a) and (c) hold from the triangle inequality. (b) is because the definition $\mathbf{d}_t=\sum_{s\in[S]}\lambda_t^s\mathbf{u}_t^s$ as shown in Line 14 in Algorithm. 1. (d) follows from the Lemma. 1. (e) is because $t\leq n_tq-1$. (f) is because we have $q=|\mathcal{A}|=\lceil\sqrt{n}\rceil$.

Then, we can conclude that

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\sum_{s\in[S]}\lambda_t^s\nabla f_s(\mathbf{x}_t)\|^2\overset{(a)}{\leq}(2SL^2\eta^2+2)\frac{[[f_s(\mathbf{x}_0)]-f_s^*]}{[\frac{\eta}{4}-\frac{\eta^3L^2}{2}]T},
\tag{21}
$$

where (a) follows from Eqs. (20) and Eqs. (19).

Let $\eta\leq\frac{1}{2L}$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\min_{\boldsymbol{\lambda}\in C}\mathbb{E}\|\boldsymbol{\lambda}^\top\nabla\mathbf{F}(\mathbf{x}_t)\|^2\leq\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\sum_{s\in[S]}\lambda_t^s\nabla f_s(\mathbf{x}_t)\|^2\\
\leq&\frac{(2SL^2\eta^2\frac{1}{T}+2)[[f_s(\mathbf{x}_0)]-f_s^*]}{[\frac{\eta}{8}]T}=\frac{(2SL^2\eta^2+2)\frac{8}{\eta}[[f_s(\mathbf{x}_0)]-f_s^*]}{T}=\mathcal{O}(\frac{1}{T}).
\end{aligned}
\tag{22}
$$

Lastly, to show the sample complexity, the number of samples with $mod(t,q)=0$ can be calculated as: $\lceil\frac{T}{q}\rceil\cdot M$. Also, the number of samples with $mod(t,q)\neq0$ can be calculated as $T\cdot|\mathcal{A}|$. Hence, the total sample complexity can be calculated as: $\lceil\frac{T}{q}\rceil n+T\cdot|\mathcal{A}|\leq\frac{T+q}{q}n+T\sqrt{n}=T\sqrt{n}+n+T\sqrt{n}=O(n+\sqrt{n}\epsilon^{-1})$. Thus, the overall sample complexity is $\mathcal{O}(n+\sqrt{n}\epsilon^{-1})$. This completes the proof.

□

A.1 PROOF OF THEOREM. 2

Proof.

$$
\begin{aligned}
&f_s(\mathbf{x}_{t+1})\\
\leq&f_s(\mathbf{x}_t)+\langle\nabla f_s(\mathbf{x}_t),-\eta\mathbf{d}_t\rangle+\frac{1}{2}L\|\eta\mathbf{d}_t\|^2\\
\overset{(a)}{\leq}&f_s(\mathbf{x}_*)+\langle\nabla f_s(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}_*\rangle-\frac{\mu}{2}\|\mathbf{x}_t-\mathbf{x}_*\|^2+\langle\nabla f_s(\mathbf{x}_t),-\eta\mathbf{d}_t\rangle+\frac{1}{2}L\|\eta\mathbf{d}_t\|^2\\
=&f_s(\mathbf{x}_*)+\langle\nabla f_s(\mathbf{x}_t),\mathbf{x}_t-\mathbf{x}_*-\eta\mathbf{d}_t\rangle-\frac{\mu}{2}\|\mathbf{x}_t-\mathbf{x}_*\|^2+\frac{1}{2}L\|\eta\mathbf{d}_t\|^2\\
\overset{(b)}{\leq}&f_s(\mathbf{x}_*)+\langle\nabla f_s(\mathbf{x}_t)-\mathbf{u}_t^s,\mathbf{x}_t-\mathbf{x}_*-\eta\mathbf{d}_t\rangle+\langle\mathbf{u}_t^s,\mathbf{x}_t-\mathbf{x}_*-\eta\mathbf{d}_t\rangle
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \mathbf{d}_t\|^2 \\
\overset{(c)}{\leq}& f_s(\mathbf{x}_*) + \frac{1}{2\delta}\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{\delta}{2}\|\mathbf{x}_t - \mathbf{x}_* - \eta\mathbf{d}_t\|^2 + \langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta\mathbf{d}_t\rangle \\
&- \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \mathbf{d}_t\|^2 \\
\overset{(d)}{\leq}& f_s(\mathbf{x}_*) + \frac{1}{2\delta}\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta\mathbf{d}_t\|^2 \\
&+ \langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta\mathbf{d}_t\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \mathbf{d}_t\|^2,
\end{aligned} \tag{23}
$$

the first inequality is due to L-smoothness, the second inequality follows from μ-strongly convex. The last two inequality follows from the triangle inequality.

According to Definition. 3 shown in the paper, the quantity to our interest is $\sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{t+1}) - f_s(\mathbf{x}_*)\right]$, then we have

$$
\begin{aligned}
&\sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{t+1}) - f_s(\mathbf{x}_*)\right] \\
\overset{(a)}{\leq}& \frac{1}{2\delta}\sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta\mathbf{d}_t\|^2 \\
&+ \left\langle \sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \left\langle \sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s, -\eta\mathbf{d}_t \right\rangle + \frac{1}{2}L\|\eta \mathbf{d}_t\|^2 \\
=& \frac{1}{2\delta}\sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta\mathbf{d}_t\|^2 \\
&+ \left\langle \sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta\mathbf{d}_t \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \mathbf{d}_t\|^2 \\
\overset{(b)}{\leq}& \frac{1}{2\delta}\sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta\mathbf{d}_t\|^2 \\
&+ \langle \mathbf{d}_t, \mathbf{x}_t - \mathbf{x}_* - \eta\mathbf{d}_t\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \mathbf{d}_t\|^2 \\
=& \langle \mathbf{d}_t, \mathbf{x}_t - \mathbf{x}_*\rangle - \eta\|\mathbf{d}_t\|^2 - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\eta^2\|\mathbf{d}_t\|^2 \\
&+ \frac{1}{2\delta}\sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta\mathbf{d}_t\|^2 \\
\overset{(c)}{\leq}& \frac{1}{2\eta}\left(\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - \frac{1}{2}\eta\|\mathbf{d}_t\|^2 - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\eta^2\|\mathbf{d}_t\|^2 \\
&+ \frac{4}{\mu}\sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{\mu}{8}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{\mu}{8}\|\eta\mathbf{d}_t\|^2 \\
\overset{(d)}{\leq}& \frac{1}{2\eta}\left((1 - \frac{3\mu\eta}{4})\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2)\|\mathbf{d}_t\|^2 \\
&+ \frac{4}{\mu}\sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 \\
\overset{(e)}{\leq}& \frac{1}{2\eta}\left((1 - \frac{3\mu\eta}{4})\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2)\|\mathbf{d}_t\|^2 \\
&+ \frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}_{(n_t-1)q}^s\|^2)
\end{aligned}
$$

$$=\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\|\mathbf{x}_t-\mathbf{x}_*\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2\right)-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\|\mathbf{d}_t\|^2$$
$$+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2). \tag{24}$$

where (a) follows from Eqs. (23). (b) is because the definition $\mathbf{d}_t=\sum_{s\in[S]}\lambda_t^s\mathbf{u}_t^s$ as shown in Line 14 in Algorithm. 1. (c) is because $\|\mathbf{x}_t-\mathbf{x}_*\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2=-\eta^2\|\mathbf{d}_t\|^2+2\left\langle\eta\mathbf{d}_t,\mathbf{x}_t-\mathbf{x}_*\right\rangle$, and we choose $\delta=\frac{\mu}{8}$ in (d). (e) follows from Lemma. 1.

Next, telescoping the above inequality over t from $(n_t-1)\,q$ to t where $t\leq n_tq-1$ and noting that for $(n_t-1)\,q\leq j\leq n_tq-1, n_j=n_t$, we obtain

$$\sum_{i=(n_t-1)q}^{t}\sum_{s\in[S]}\lambda_i^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]$$
$$\overset{(a)}{\leq}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)$$
$$-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\sum_{i=(n_t-1)q}^{t}\|\mathbf{d}_i\|^2+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_j-1)q}^{j}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)$$
$$\overset{(b)}{\leq}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)$$
$$-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\sum_{i=(n_t-1)q}^{t}\|\mathbf{d}_i\|^2+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)$$
$$=\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)$$
$$-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|})\sum_{i=(n_t-1)q}^{t}\|\mathbf{d}_i\|^2), \tag{25}$$

where (a) is from Eqs. (24). (b) relaxes j to t, since $j\leq t$. We continue the proof by further driving

$$\sum_{i=0}^{T}\sum_{s\in[S]}\lambda_i^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]$$
$$=\sum_{i=0}^{q}\sum_{s\in[S]}\lambda_i^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+\sum_{i=q}^{2q}\sum_{s\in[S]}\lambda_i^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+$$
$$\cdots+\sum_{i=(n_T-1)q}^{T}\sum_{s\in[S]}\lambda_i^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]$$
$$\leq\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)$$
$$-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|})\sum_{i=0}^{T}\|\mathbf{d}_i\|^2), \tag{26}$$

where the last inequality is from Eq. (15) and Eq. (25). Next, we have

$$\begin{aligned}
&\sum_{i=0}^{T} \sum_{s\in[S]} \lambda_i^s \left[f_s(\mathbf{x}_i) - f_s(\mathbf{x}_*)\right] \\
=&\sum_{i=0}^{T} \sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{i+1}) - f_s(\mathbf{x}_*) - f_s(\mathbf{x}_{i+1}) + f_s(\mathbf{x}_i)\right] \\
\leq&\sum_{i=0}^{T} \sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{i+1}) - f_s(\mathbf{x}_*)\right] - \sum_{i=0}^{T} \sum_{s\in[S]} \lambda_t^s |f_s(\mathbf{x}_{i+1}) - f_s(\mathbf{x}_i)| \\
\leq&\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i - \mathbf{x}_*\|^2 - \sum_{i=0}^{T}\|\mathbf{x}_{i+1} - \mathbf{x}_*\|^2\right) \\
&- (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2 - \frac{4}{\mu}\frac{L^2 q\eta^2}{|\mathcal{A}|} - [\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|}])\sum_{i=0}^{T}\|\mathbf{d}_i\|^2 \qquad (27)
\end{aligned}$$

Let $|\mathcal{A}| = q = \lceil\sqrt{n}\rceil$ and $\eta \leq \min\{\frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}$, we have $(\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2 - \frac{4}{\mu}\frac{L^2 q\eta^2}{|\mathcal{A}|} - [\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|}]) > \frac{\eta}{16} > 0$
Thus, we have

$$\sum_{i=0}^{T} \sum_{s\in[S]} \lambda_i^s \left[f_s(\mathbf{x}_i) - f_s(\mathbf{x}_*)\right] \leq \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i - \mathbf{x}_*\|^2 - \sum_{i=0}^{T}\|\mathbf{x}_{i+1} - \mathbf{x}_*\|^2\right). \qquad (28)$$

Then, we have

$$\mathbb{E}_t[\sum_{s\in[S]} \lambda_i^s \left[f_s(\mathbf{x}_t) - f_s(\mathbf{x}_*)\right]] \leq \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\mathbb{E}_t\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \mathbb{E}_t\|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right). \qquad (29)$$

Based on Assumption.4 and averaging using weight $w_t = (1-\frac{3\mu\eta}{4})^{1-t}$ and using such weight to pick output $\mathbf{x}$, by using Lemma 1 in Karimireddy et al. [2020] with $\eta \geq \frac{1}{uR}$, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 \left[f_s(\mathbf{x}_t) - f_s(\mathbf{x}_*)\right]] &\leq \|\mathbf{x}_0 - \mathbf{x}_*\|^2 \mu \exp(-\frac{3\eta\mu T}{4}) \qquad (30) \\
&= \mathcal{O}(\mu \exp(-\mu T)). \qquad (31)
\end{aligned}$$

Then we have the convergence rate $\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 = \mathcal{O}(\mu\exp(-\mu T))$.

Lastly, the total sample complexity can be calculated as: $\lceil\frac{T}{q}\rceil n + T\cdot|\mathcal{A}| \leq \frac{T+q}{q}n + T\sqrt{n} = T\sqrt{n} + n + T\sqrt{n} = O(n + \sqrt{n}\ln(\mu/\epsilon)$. Thus, the overall sample complexity is $\mathcal{O}(n + \sqrt{n}\ln(\mu/\epsilon)$. This completes the proof.

□

B PROOF OF CONVERGENCE OF STIMULUS-M

Lemma 3. *For general L-smooth functions $\{f_s, s \in [S]\}$, choose the learning rate η s.t. $\eta \leq \frac{1}{2}$, the update d_t of the VR-MOO-M algorithm satisfies:*

$$\begin{aligned}
f_s(\mathbf{x}_{t+1}) \leq& f_s(\mathbf{x}_t) + \frac{\eta}{2}\sum_{i=(n_t-1)q}^{t}\alpha^{(t-i)}\|\nabla f_s(\mathbf{x}_i) - \mathbf{u}_i^s\|^2 - \frac{1}{2}\eta\sum_{i=(n_t-1)q}^{t}\alpha^{(t-i)}\|\mathbf{d}_i\|^2 \\
&+ \frac{1}{2}L\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2. \qquad (32)
\end{aligned}$$

Proof of Lemma. 3.

Proof.

$$
\begin{aligned}
&f_s(\mathbf{x}_{t+1}) \leq f_s(\mathbf{x}_t) + \langle \nabla f_s(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\overset{(a)}{\leq} f_s(\mathbf{x}_t) + \langle \nabla f_s(\mathbf{x}_t), \alpha(\mathbf{x}_{t+1} - \mathbf{x}_t) \rangle + \langle \nabla f_s(\mathbf{x}_t), -\eta \mathbf{d}_t \rangle + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
&\overset{(b)}{=} f_s(\mathbf{x}_t) + \sum_{i=0}^{t} \alpha^{(t-i)} \langle \nabla f_s(\mathbf{x}_i), -\eta \mathbf{d}_i \rangle + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
&= f_s(\mathbf{x}_t) - \eta \sum_{i=0}^{t} \alpha^{(t-i)} \langle \nabla f_s(\mathbf{x}_i) - \mathbf{u}_i^s, \mathbf{d}_i \rangle - \eta \sum_{i=0}^{t} \alpha^{(t-i)} \langle \mathbf{u}_i^s, \mathbf{d}_i \rangle + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\overset{(c)}{\leq} f_s(\mathbf{x}_t) - \eta \sum_{i=0}^{t} \alpha^{(t-i)} \langle \nabla f_s(\mathbf{x}_i) - \mathbf{u}_i^s, \mathbf{d}_i \rangle - \eta \sum_{i=0}^{t} \alpha^{(t-i)} \|\mathbf{d}_i\|^2 + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\overset{(d)}{\leq} f_s(\mathbf{x}_t) + \frac{\eta}{2} \sum_{i=0}^{t} \alpha^{(t-i)} \|\nabla f_s(\mathbf{x}_i) - \mathbf{u}_i^s\|^2 + \frac{1}{2} \eta \sum_{i=0}^{t} \alpha^{(t-i)} \|\mathbf{d}_i\|^2 \\
&- \eta \sum_{i=0}^{t} \alpha^{(t-i)} \|\mathbf{d}_i\|^2 + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= f_s(\mathbf{x}_t) + \frac{\eta}{2} \sum_{i=0}^{t} \alpha^{(t-i)} \|\nabla f_s(\mathbf{x}_i) - \mathbf{u}_i^s\|^2 - \frac{1}{2} \eta \sum_{i=0}^{t} \alpha^{(t-i)} \|\mathbf{d}_i\|^2 + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2. \qquad (33)
\end{aligned}
$$

(a) follows from the objective function f_s is L-smooth. (b) follows from the update rule of $\mathbf{x}_t$ shown in Line 19 in Algorithm. 1. (c) follows from $\langle \mathbf{u}_t^s, \mathbf{d}_t \rangle \geq \|\mathbf{d}_t\|^2$ since $\mathbf{d}_t$ is a general solution in the convex hull of the family of vectors $\{\mathbf{u}_t^s, s \in [S]\}$ (see Lemma 2.1 Désidéri [2012]). (d) follows from the triangle inequality.

□

Proof of Theorem. 3

Proof. Taking expectation on both sides of the inequality in Lemma. 3, we have

$$
\begin{aligned}
&\mathbb{E}[f_s(\mathbf{x}_{t+1})] \\
&\overset{(a)}{\leq} \mathbb{E}[f_s(\mathbf{x}_t)] + \frac{\eta}{2} \sum_{i=0}^{t} \alpha^{(t-i)} \mathbb{E}\|\nabla f_s(\mathbf{x}_i) - \mathbf{u}_i^s\|^2 - \frac{1}{2} \eta \sum_{i=0}^{t} \alpha^{(t-i)} \mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2} L \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\overset{(b)}{\leq} \mathbb{E}[f_s(\mathbf{x}_t)] - \frac{1}{2} \eta \sum_{i=0}^{t} \alpha^{(t-i)} \mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2} L \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\quad + \frac{\eta}{2} \sum_{j=0}^{t} \alpha^{(t-j)} [\frac{L^2}{|\mathcal{A}|} \sum_{i=(n_t-1)q}^{j} \mathbb{E}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \mathbb{E}\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}_{(n_t-1)q}^s\|^2] \\
&= \mathbb{E}[f_s(\mathbf{x}_t)] - \frac{1}{2} \eta \sum_{i=0}^{t} \alpha^{(t-i)} \mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2} L \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\quad + \frac{\eta}{2} \sum_{j=0}^{t} \alpha^{(t-j)} [\frac{L^2}{|\mathcal{A}|} \sum_{i=(n_t-1)q}^{j} \mathbb{E}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2], \qquad (34)
\end{aligned}
$$

where (a) follows from Eqs. 33. (b) follows from the Lemma. 1. (c) follows from $\mathbb{E}\|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}_{(n_t-1)q}^s\|^2 = 0$ as shown in Line 5 in our Algorithm. 1.

Next, telescoping the above inequality over t from $(n_t-1)q$ to t where $t \le n_t q - 1$ and noting that for $(n_t-1)q \le j \le n_t q - 1, n_j = n_t$ and let $\eta \le \frac{1}{4L}$, we obtain

$$
\begin{aligned}
&\mathbb{E}[f_s(\mathbf{x}_{t+1})] \\
\overset{(a)}{\le} &\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2 \\
&+ \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[\frac{L^2}{|\mathcal{A}|}\sum_{r=(n_t-1)q}^{i}\mathbb{E}\|\mathbf{x}_{r+1}-\mathbf{x}_r\|^2] \\
\overset{(b)}{\le} &\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2 \\
&+ \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[\frac{L^2}{|\mathcal{A}|}\sum_{r=(n_t-1)q}^{n_tq-1}\mathbb{E}\|\mathbf{x}_{r+1}-\mathbf{x}_r\|^2] \\
\le &\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2 \\
&+ \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[\frac{L^2}{|\mathcal{A}|}q\mathbb{E}\|\mathbf{x}_{j+1}-\mathbf{x}_j\|^2] \\
\overset{(c)}{=} &\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2 \\
&+ \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[L^2\mathbb{E}\|\mathbf{x}_{j+1}-\mathbf{x}_j\|^2] \\
\overset{(d)}{=} &\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2 \\
&+ \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[L^2\mathbb{E}\|\eta\sum_{r=0}^{j}\alpha^{(j-r)}\mathbf{d}_r\|^2] \\
\overset{(e)}{\le} &\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2 \\
&+ \frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{2(j-i)}[L^2\eta^2\mathbb{E}\|\mathbf{d}_i\|^2] \\
\overset{(f)}{\le} &\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{4}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_i\|^2 + \frac{1}{2}L\sum_{j=(n_t-1)q}^{t}\mathbb{E}\|\eta\sum_{i=0}^{j}\alpha^{(j-i)}\mathbf{d}_j\|^2 \\
\overset{(g)}{\le} &\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{8}\sum_{j=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{d}_j\|^2,
\end{aligned}
\tag{35}
$$

where (a) holds from Eqs. (34). (b) is extend i to t since $i \le n_t q - 1$. (c) is because $q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$. (d) follows from the update rule of $\mathbf{x}_t$ shown in Line 19 in Algorithm. 1. (e) follows from the triangle inequality. (f) and (g) hold from $\eta \le \frac{1}{2L}$ and $0 < \alpha < 1$. We continue the proof by further driving

$$
\begin{aligned}
&[f_s(\mathbf{x}_T)] - [f_s(\mathbf{x}_0)] \\
&= ([f_s(\mathbf{x}_q)] - [f_s(\mathbf{x}_0)]) + ([f_s(\mathbf{x}_{2q})] - [f_s(\mathbf{x}_q)]) + \cdot + ([f_s(\mathbf{x}_T)] - [f_s(\mathbf{x}_{(n_T-1)q})])
\end{aligned}
$$

$$\leq -[\frac{\eta}{8}] \sum_{t=0}^{T-1} \|\mathbf{d}_t\|^2 \tag{36}$$

Note that $[f_s(\mathbf{x}_{T+1})] \geq f_s^* \triangleq \inf_{\mathbf{x}\in\mathbb{R}^d} f_s(\mathbf{x})$. Hence, we have

$$[\frac{\eta}{8}] \sum_{t=0}^{T-1} \|\mathbf{d}_t\|^2 \leq [[f_s(\mathbf{x}_0)] - [f_s(\mathbf{x}_T)]] \leq [[f_s(\mathbf{x}_0)] - f_s^*]. \tag{37}$$

Based on the parameter setting $q = |\mathcal{A}| = \sqrt{n}$, we have

$$[\frac{\eta}{8}] \sum_{t=0}^{T-1} \|\mathbf{d}_t\|^2 \leq [[f_s(\mathbf{x}_0)] - f_s^*]. \tag{38}$$

Since $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|d_t\|^2$ is just common descent directions. According to Definition. 3 shown in the paper, the quantity to our interest is $\|\sum_{s\in[S]} \lambda_t^s \nabla f(\mathbf{x})\|^2$.

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\sum_{s\in[S]} \lambda_t^s \nabla f_s(\mathbf{x}_t)\|^2 \overset{(a)}{\leq} (2SL^2\eta^2\frac{1}{T}+2)\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{d}_t\|^2 \tag{39}$$

where (a) follows from Eqs. (20).

Then, we can conclude that

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\sum_{s\in[S]} \lambda_t^s \nabla f_s(\mathbf{x}_t)\|^2 \overset{(a)}{\leq} (2SL^2\eta^2+2)\frac{[\mathbb{E}[f_s(\mathbf{x}_0)] - f_s^*]}{\frac{\eta}{8}T}, \tag{40}$$

where (a) follows from Eqs. (20) and Eqs. 19.

Thus, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \min_{\boldsymbol{\lambda}\in C} \mathbb{E}\|\boldsymbol{\lambda}^\top \nabla \mathbf{F}(\mathbf{x}_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\sum_{s\in[S]} \lambda_t^s \nabla f_s(\mathbf{x}_t)\|^2 = \mathcal{O}(\frac{1}{T}). \tag{41}$$

The total sample complexity can be calculated as: $\lceil \frac{T}{q} \rceil n + T\cdot|\mathcal{A}| \leq \frac{T+q}{q}n + T\sqrt{n} = T\sqrt{n} + n + T\sqrt{n} = O(n+\sqrt{n}\epsilon^{-1})$. Thus, the overall sample complexity is $\mathcal{O}(n+\sqrt{n}\epsilon^{-1})$. This completes the proof.

□

B.1 PROOF OD THEOREM. 4

Proof.

$$\begin{aligned}
&f_s(\mathbf{x}_{t+1}) \\
&\overset{(a)}{\leq} f_s(\mathbf{x}_t) + \left\langle \nabla f_s(\mathbf{x}_t), -\eta\sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle + \frac{1}{2}L\|\eta\sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2 \\
&\overset{(b)}{\leq} f_s(\mathbf{x}_*) + \langle \nabla f_s(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* \rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \left\langle \nabla f_s(\mathbf{x}_t), -\eta\sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle \\
&\quad + \frac{1}{2}L\|\eta\sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2
\end{aligned}$$

$$
\begin{aligned}
&= f_s(\mathbf{x}_*) + \left\langle \nabla f_s(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&= f_s(\mathbf{x}_*) + \left\langle \nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle + \left\langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle \\
&\quad - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\overset{(c)}{\le} f_s(\mathbf{x}_*) + \frac{1}{2\delta}\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{\delta}{2}\|\mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\quad + \left\langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\overset{(d)}{\le} f_s(\mathbf{x}_*) + \frac{1}{2\delta}\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\quad + \left\langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2, \qquad (42)
\end{aligned}
$$

where (a) is due to L-smoothness, (b) follows from μ-strongly convex. (c) and (d) follow from the Young's inequality.

Next, we have

$$
\begin{aligned}
&\sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{t+1}) - f_s(\mathbf{x}_*)\right] \\
&\overset{(a)}{\le} \frac{1}{2\delta} \sum_{s\in[S]} \lambda_t^s \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\quad + \left\langle \sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \left\langle \sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s, -\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle \\
&\quad + \frac{1}{2} L \|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&= \frac{1}{2\delta} \sum_{s\in[S]} \lambda_t^s \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\quad + \left\langle \sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\overset{(b)}{=} \frac{1}{2\delta} \sum_{s\in[S]} \lambda_t^s \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\quad + \left\langle \mathbf{d}_t, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&\overset{(c)}{\le} \frac{1}{2\eta}\left(\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - \frac{1}{2}\eta\|\sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 \\
&\quad + \frac{1}{2} L \|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 + \frac{4}{\mu} \sum_{s\in[S]} \lambda_t^s \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{\mu}{8}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{\mu}{8}\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&= \frac{1}{2\eta}\left((1 - \frac{3\mu\eta}{4})\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2)\|\sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{4}{\mu}\sum_{s\in[S]}\lambda_t^s\|\nabla f_s(\mathbf{x}_t)-\mathbf{u}_t^s\|^2\\
&\overset{(e)}{\le}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\|\mathbf{x}_t-\mathbf{x}_*\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2\right)-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2\\
&\quad+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2+\sum_{s\in[S]}\lambda_t^s\|\nabla f_s(\mathbf{x}_{(n_t-1)q})-\mathbf{u}^s_{(n_t-1)q}\|^2)\\
&\overset{(f)}{=}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\|\mathbf{x}_t-\mathbf{x}_*\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2\right)-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2\\
&\quad+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2).
\end{aligned}
\tag{43}
$$

where (a) follows from Eqs. (42). (b) is because the definition $\mathbf{d}_t=\sum_{s\in[S]}\lambda_t^s\mathbf{u}_t^s$ as shown in Line 14 in Algorithm. 1. (c) is because $\|\mathbf{x}_t-\mathbf{x}_*\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2=-\eta^2\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2+2\left\langle\eta\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i,\mathbf{x}_t-\mathbf{x}_*\right\rangle$, and we choose $\delta=\frac{\mu}{8}$. (e) and (f) follow from $\sum_{s\in[S]}\lambda_t^s=1$ and $\|\nabla f_s(\mathbf{x}_{(n_t-1)q})-\mathbf{u}^s_{(n_t-1)q}\|^2=0$.

Next, telescoping the above inequality over t from $(n_t-1)\,q$ to t where $t\le n_tq-1$ and noting that for $(n_t-1)\,q\le j\le n_tq-1, n_j=n_t$, we obtain

$$
\begin{aligned}
&\sum_{i=(n_t-1)q}^{t}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]\\
&\overset{(a)}{=}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&\quad-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\sum_{i=(n_t-1)q}^{t}\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_j-1)q}^{j}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)\\
&\overset{(b)}{\le}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&\quad-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\sum_{i=(n_t-1)q}^{t}\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)\\
&\overset{(c)}{=}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&\quad-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|})\sum_{i=(n_t-1)q}^{t}\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2),
\end{aligned}
\tag{44}
$$

where (a) follows from Eqs. (43), (b) extend j to t. (c) follows from the update rule of $\mathbf{x}_{t+1}$ shown in Eqs. (4).

We continue the proof by further driving

$$
\begin{aligned}
&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]\\
&=\sum_{i=0}^{q}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+\sum_{i=q}^{2q}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+
\end{aligned}
$$

$$\begin{aligned}
&\sum_{i=(n_T-1)q}^{T} \sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{i+1}) - f_s(\mathbf{x}_*)\right] \\
&\overset{(a)}{\leq} \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2 - \sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right) \\
&\quad -(\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2 - \frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|})\sum_{t=0}^{T}\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2),
\end{aligned} \tag{45}$$

where (a) follows from Eqs. (44). Next, we have

$$\begin{aligned}
&\sum_{t=0}^{T} \sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_i) - f_s(\mathbf{x}_*)\right] \\
=&\sum_{t=0}^{T} \sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{i+1}) - f_s(\mathbf{x}_*) - f_s(\mathbf{x}_{i+1}) + f_s(\mathbf{x}_i)\right] \\
=&\sum_{t=0}^{T} \sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{i+1}) - f_s(\mathbf{x}_*)\right] - \sum_{t=0}^{T} \sum_{s\in[S]} \lambda_t^s |f_s(\mathbf{x}_{i+1}) - f_s(\mathbf{x}_i)| \\
\overset{(a)}{leq}& \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2 - \sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right) \\
&-(\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2 - \frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|} - [\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|}])\sum_{t=0}^{T}\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2,
\end{aligned} \tag{46}$$

where (a) follows from Eqs. (45). Let $|\mathcal{A}| = q = \lceil\sqrt{n}\rceil$ and $\eta \leq \min\{\frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}$, we have $(\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2 - \frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|} - [\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|}]) > \frac{\eta}{16} > 0$
Thus, we have

$$\sum_{t=0}^{T} \sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_i) - f_s(\mathbf{x}_*)\right] \leq \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2 - \sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right). \tag{47}$$

Then, we have

$$\mathbb{E}[\sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_t) - f_s(\mathbf{x}_*)\right]] \leq \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\mathbb{E}\|\mathbf{x}_t-\mathbf{x}_*\|^2 - \mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2\right). \tag{48}$$

Based on Asumption. 4 and averaging using weight $w_t = (1-\frac{3\mu\eta}{4})^{1-t}$ and using such weight to pick output $\mathbf{x}$, by using Lemma 1 in Karimireddy et al. [2020] with $\eta \geq \frac{1}{uR}$, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}_0 - \mathbf{x}_*\|^2 \mu \exp(-\frac{3\eta\mu T}{4}) \qquad (49)\\
&= \mathcal{O}(\mu\exp(-\mu T)). \qquad (50)
\end{aligned}$$

Then we have the convergence rate $\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 = \mathcal{O}(\mu\exp(-\mu T))$. the total sample complexity can be calculated as: $\lceil\frac{T}{q}\rceil n + T\cdot|\mathcal{A}| \leq \frac{T+q}{q}n + T\sqrt{n} = T\sqrt{n} + n + T\sqrt{n} = O(n+\sqrt{n}\ln(\mu/\epsilon)$. Thus, the overall sample complexity is $\mathcal{O}(n+\sqrt{n}\ln(\mu/\epsilon)$. This completes the proof.

□

C PROOF OF CONVERGENCE OF STIMULUS^{+}

Proof of Theorem. 5 [Part 1]

Proof. Recall that $\mathcal{N}_s = \min\{c_\gamma\sigma^2(\gamma_t)^{-1}, c_\epsilon\sigma^2\epsilon^{-1}, n\}$. Then we have

$$\begin{aligned}\frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s} &\le \frac{1}{\min\{c_\epsilon\sigma^2(\epsilon)^{-1}, c_\gamma\sigma^2(\gamma_t)^{-1}\}}\\ &= \max\{\frac{\gamma_t}{c_\gamma\sigma^2}, \frac{\epsilon}{c_\epsilon\sigma^2}\} \le \frac{\gamma_t}{c_\gamma\sigma^2} + \frac{\epsilon}{c_\epsilon\sigma^2}.\end{aligned} \tag{51}$$

From Lemma. 2, we have

$$\begin{aligned}[f_s(\mathbf{x}_{t+1})] &\overset{(a)}{\le} [f_s(\mathbf{x}_t)] + \frac{\eta}{2}\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 - \frac{\eta}{4}\|\mathbf{d}_t\|^2\\ &\overset{(b)}{\le} [f_s(\mathbf{x}_t)] - \frac{\eta}{4}\|\mathbf{d}_t\|^2\\ &\quad + \frac{\eta}{2}[\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}^s_{(n_t-1)q}\|^2]\\ &\overset{(c)}{\le} [f_s(\mathbf{x}_t)] - \frac{\eta}{4}\|\mathbf{d}_t\|^2 + \frac{\eta}{2}[\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t}\eta^2\|\mathbf{d}_i\|^2 + \frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s}\sigma^2],\end{aligned} \tag{52}$$

where (a) follows from Lemma. 2. (b) follows from Lemma. 1. (c) follows from the update rule shown in Eqs. (5).

Next, telescoping the above inequality over t from $(n_t-1)\,q$ to t where $t \le n_tq - 1$ and noting that for $(n_t-1)\,q \le j \le n_tq - 1, n_j = n_t$, and aking expectation on both sides of the inequality in Eqs. (52),we obtain

$$\begin{aligned}&\mathbb{E}[f_s(\mathbf{x}_{t+1})]\\ &\overset{(a)}{\le}\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{4}\sum_{j=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{d}_j\|^2\\ &\quad + \frac{\eta}{2}[\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_t-1)q}^{j}\eta^2\mathbb{E}\|\mathbf{d}_i\|^2 + \sum_{i=(n_t-1)q}^{t}\frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s}\sigma^2]\\ &\overset{(b)}{\le}\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{4}\sum_{j=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{d}_j\|^2\\ &\quad + \frac{\eta}{2}\sum_{i=(n_t-1)q}^{t}[\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_t-1)q}^{t}\eta^2\mathbb{E}\|\mathbf{d}_i\|^2] + \frac{\eta}{2}\sum_{i=(n_t-1)q}^{t}\frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s}\sigma^2\\ &=\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{4}\sum_{j=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{d}_j\|^2\\ &\quad + \frac{\eta^3q}{2}[\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{d}_j\|^2] + \frac{\eta}{2}\sum_{i=(n_t-1)q}^{t}\frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s}\sigma^2\\ &\overset{(c)}{=}\mathbb{E}[f_s(\mathbf{x}_{(n_t-1)q})] - [\frac{\eta}{4} - \frac{\eta^3q}{2}\frac{L^2}{|\mathcal{A}|}]\sum_{j=(n_t-1)q}^{t}\mathbb{E}\|\mathbf{d}_j\|^2 + \frac{\eta}{2}\sum_{i=(n_t-1)q}^{t}(\frac{\gamma_i}{c_\gamma} + \frac{\epsilon}{c_\epsilon}),\end{aligned} \tag{53}$$

where (a) follows from Eqs. (52), (b) extends j to t. (c) follows from Eqs. (51)

Recall that $\gamma_t = \frac{1}{q}\sum_{i=(n_t-1)q}^{t} \|\mathbf{d}_t\|^2$. Then, we have We continue the proof by further driving

$$\begin{aligned}
&\mathbb{E}[f_s(\mathbf{x}_T) - f_s(\mathbf{x}_0)] \\
&= \mathbb{E}[([f_s(\mathbf{x}_q)] - [f_s(\mathbf{x}_0)]) + ([f_s(\mathbf{x}_{2q})] - [f_s(\mathbf{x}_q)]) + \cdot + ([f_s(\mathbf{x}_T)] - [f_s(\mathbf{x}_{(n_T-1)q})])] \\
&\overset{(a)}{\leq} -[\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|}]\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2 + \frac{\eta}{2}\sum_{t=0}^{T-1}(\frac{\mathbb{E}[\gamma_i]}{c_\gamma} + \frac{\epsilon}{c_\epsilon}) \\
&\overset{(b)}{\leq} -[\frac{\eta}{4} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|} - \frac{\eta}{2c_\gamma}]\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2 + \frac{\eta}{2}T\frac{\epsilon}{c_\epsilon},
\end{aligned} \tag{54}$$

where (a) is from Eqs. (53). (b) follows from $\gamma_t = \frac{1}{q}\sum_{i=(n_t-1)q}^{t} \|\mathbf{d}_t\|^2$.

Note that $[f_s(\mathbf{x}_{T+1})] \geq f_s^* \triangleq \inf_{\mathbf{x}\in\mathbb{R}^d} f_s(\mathbf{x})$. Let $c_\gamma > 4$. Hence, we have

$$[\frac{\eta}{8} - \frac{\eta^3 q}{2}\frac{L^2}{|\mathcal{A}|} - \frac{\eta}{2c_\gamma}]\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2 \leq \mathbb{E}[[f_s(\mathbf{x}_0)] - [f_s(\mathbf{x}_T)]] \leq \mathbb{E}[[f_s(\mathbf{x}_0)] - f_s^*] + \frac{\eta}{2}T\frac{\epsilon}{c_\epsilon}. \tag{55}$$

Based on the parameter setting $q = |\mathcal{A}| = \lceil\sqrt{n}\rceil$, we have

$$[\frac{\eta}{8} - \frac{\eta^3 L^2}{2} - \frac{\eta}{2c_\gamma}]\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2 \leq \mathbb{E}[[f_s(\mathbf{x}_0)] - f_s^*] + \frac{\eta}{2}T\frac{\epsilon}{c_\epsilon}. \tag{56}$$

Thus, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2 \leq \frac{\mathbb{E}[[f_s(\mathbf{x}_0)] - f_s^*]}{[\frac{\eta}{8} - \frac{\eta^3 L^2}{2} - \frac{\eta}{2c_\gamma}]T} + \frac{\eta}{2}\frac{\epsilon}{c_\epsilon}. \tag{57}$$

Let $\eta \leq \frac{1}{4L}, c_\gamma \geq 8, c_\epsilon \geq \eta$, we have

Since $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|d_t\|^2$ is just common descent directions. According to Definition. 3 shown in the paper, the quantity to our interest is $\|\sum_{s\in[S]}\lambda_t^s\nabla f(\mathbf{x})\|^2$.

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\sum_{s\in[S]}\lambda_t^s\nabla f_s(\mathbf{x}_t)\|^2 \overset{(a)}{\leq} (2SL^2\eta^2 + 2)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2 \tag{58}$$

where (a) follows from Eqs. (20).

Then, we can conclude that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\sum_{s\in[S]}\lambda_t^s\nabla f_s(\mathbf{x}_t)\|^2 \overset{(a)}{\leq} (2SL^2\eta^2 + 2)(\frac{\mathbb{E}[[f_s(\mathbf{x}_0)] - f_s^*]}{[\frac{\eta}{8} - \frac{\eta^3 L^2}{2} - \frac{\eta}{2c_\gamma}]T} + \frac{\eta}{2}\frac{\epsilon}{c_\epsilon}), \tag{59}$$

where (a) follows from Eqs. (20) and Eqs. 19.

Thus, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\min_{\boldsymbol{\lambda}\in C}\mathbb{E}\|\boldsymbol{\lambda}^\top\nabla\mathbf{F}(\mathbf{x}_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\sum_{s\in[S]}\lambda_t^s\nabla f_s(\mathbf{x}_t)\|^2 = \mathcal{O}(\frac{1}{T}). \tag{60}$$

The total sample complexity can be calculated as: $\lceil\frac{T}{q}\rceil n + T\cdot|\mathcal{A}| \leq \frac{T+q}{q}n + T\sqrt{n} = T\sqrt{n} + n + T\sqrt{n} = O(n + \sqrt{n}\epsilon^{-1})$. Thus, the overall sample complexity is $\mathcal{O}(n + \sqrt{n}\epsilon^{-1})$. This completes the proof.

□

C.1 PROOF OF THEOREM. 6 [PART 1]

Proof.

$$
\begin{aligned}
& f_s(\mathbf{x}_{t+1}) \\
\overset{(a)}{\le} & f_s(\mathbf{x}_t) + \langle \nabla f_s(\mathbf{x}_t), -\eta \mathbf{d}_t \rangle + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
\overset{(b)}{\le} & f_s(\mathbf{x}_*) + \langle \nabla f_s(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* \rangle - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \langle \nabla f_s(\mathbf{x}_t), -\eta \mathbf{d}_t \rangle + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
= & f_s(\mathbf{x}_*) + \langle \nabla f_s(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* - \eta \mathbf{d}_t \rangle - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
= & f_s(\mathbf{x}_*) + \langle \nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \mathbf{d}_t \rangle + \langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \mathbf{d}_t \rangle \\
& - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
\overset{(c)}{\le} & f_s(\mathbf{x}_*) + \frac{1}{2\delta} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{\delta}{2} \|\mathbf{x}_t - \mathbf{x}_* - \eta \mathbf{d}_t\|^2 + \langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \mathbf{d}_t \rangle \\
& - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \\
\overset{(d)}{\le} & f_s(\mathbf{x}_*) + \frac{1}{2\delta} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta \|\eta \mathbf{d}_t\|^2 \\
& + \langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \mathbf{d}_t \rangle - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2, \qquad (61)
\end{aligned}
$$

where (a) follows from L-smoothness, (b) follows from μ-strongly convexity. (c) follows from Young's inequality, and (d) follows from triangle inequality.

Then, we have

$$
\begin{aligned}
& \sum_{s \in [S]} \lambda_t^s \left[f_s(\mathbf{x}_{t+1}) - f_s(\mathbf{x}_*) \right] \qquad (62) \\
\overset{(a)}{\le} & \frac{1}{2\delta} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta \|\eta \mathbf{d}_t\|^2 \\
& + \left\langle \sum_{s \in [S]} \lambda_t^s \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* \right\rangle - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \left\langle \sum_{s \in [S]} \lambda_t^s \mathbf{u}_t^s, -\eta \mathbf{d}_t \right\rangle + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \qquad (63) \\
= & \frac{1}{2\delta} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta \|\eta \mathbf{d}_t\|^2 \\
& + \left\langle \sum_{s \in [S]} \lambda_t^s \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \mathbf{d}_t \right\rangle - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \qquad (64) \\
\overset{(b)}{\le} & \frac{1}{2\delta} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta \|\eta \mathbf{d}_t\|^2 \\
& + \langle \mathbf{d}_t, \mathbf{x}_t - \mathbf{x}_* - \eta \mathbf{d}_t \rangle - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \|\eta \mathbf{d}_t\|^2 \qquad (65) \\
= & \langle \mathbf{d}_t, \mathbf{x}_t - \mathbf{x}_* \rangle - \eta \|\mathbf{d}_t\|^2 - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \eta^2 \|\mathbf{d}_t\|^2 \\
& + \frac{1}{2\delta} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta \|\eta \mathbf{d}_t\|^2 \\
\overset{(c)}{=} & \frac{1}{2\eta} \left(\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 \right) - \frac{1}{2} \eta \|\mathbf{d}_t\|^2 - \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2} L \eta^2 \|\mathbf{d}_t\|^2 \\
& + \frac{4}{\mu} \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{\mu}{8} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{\mu}{8} \|\eta \mathbf{d}_t\|^2 \qquad (66) \\
= & \frac{1}{2\eta} \left((1 - \frac{3\mu\eta}{4}) \|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 \right) - (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2} L \eta^2) \|\mathbf{d}_t\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{4}{\mu}\|\nabla f_s(\mathbf{x}_t)-\mathbf{u}_t^s\|^2 && (67)\\
&\overset{(d)}{\leq}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\|\mathbf{x}_t-\mathbf{x}_*\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2\right)-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\|\mathbf{d}_t\|^2\\
&\quad+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2+\|\nabla f_s(\mathbf{x}_{(n_t-1)q})-\mathbf{u}^s_{(n_t-1)q}\|^2) && (68)\\
&\overset{(f)}{\leq}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\|\mathbf{x}_t-\mathbf{x}_*\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2\right)-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\|\mathbf{d}_t\|^2\\
&\quad+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)+\frac{\mu}{4}\frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s}\sigma^2. && (69)
\end{aligned}
$$

where (a) follows from Eqs.(61). (b) follows from the definition $\mathbf{d}_t=\sum_{s\in[S]}\lambda_t^s\mathbf{u}_t^s$ as shown in Line 14 in Algorithm. 1. (c) is because $\|\mathbf{x}_t-\mathbf{x}_*\|^2-\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2=-\eta^2\|\mathbf{d}_t\|^2+2\langle\eta\mathbf{d}_t,\mathbf{x}_t-\mathbf{x}_*\rangle$. (d) is from Lemma. 1 and we choose $\delta=\frac{\mu}{8}$. (e) is from Eqs. (51).

Next, telescoping the above inequality over t from $(n_t-1)\,q$ to t where $t\leq n_tq-1$ and noting that for $(n_t-1)\,q\leq j\leq n_tq-1, n_j=n_t$, we obtain

$$
\begin{aligned}
&\sum_{i=(n_t-1)q}^{t}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]\\
&\overset{(a)}{\leq}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&\quad-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\sum_{i=(n_t-1)q}^{t}\|\mathbf{d}_i\|^2+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_j-1)q}^{j}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)\\
&\quad+\frac{\mu S}{4}\sum_{i=(n_t-1)q}^{t}\frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s}\sigma^2\\
&\overset{(b)}{\leq}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&\quad-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\sum_{i=(n_t-1)q}^{t}\|\mathbf{d}_i\|^2+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)\\
&\quad+\frac{\mu}{4}\sum_{i=(n_t-1)q}^{t}\frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s}\sigma^2\\
&\overset{(c)}{\leq}\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&\quad-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|})\sum_{i=(n_t-1)q}^{t}\|\mathbf{d}_i\|^2)\\
&\quad+\frac{\mu}{4}\sum_{i=(n_t-1)q}^{t}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon}), && (70)
\end{aligned}
$$

where (a) follows from Eqs. (62) and the fact that $\lambda_t^s\leq 1\forall s\in[S]$. (b) extends j to t. (c) is because $t-(n_t-1)q\geq q$. We continue the proof by further driving

$$\begin{aligned}
&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]\\
=&\sum_{i=0}^{q}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+\sum_{i=q}^{2q}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+\\
&\cdot+\sum_{i=(n_T-1)q}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]\\
\overset{(a)}{\leq}&\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|}+\frac{\mu}{4c_\gamma})\sum_{t=0}^{T}\|\mathbf{d}_i\|^2+\frac{\mu}{4}T\frac{\epsilon}{c_\epsilon},
\end{aligned}\tag{71}$$

where (a) follows from Eqs. (70) and $\gamma_t=\frac{1}{q}\sum_{i=(n_t-1)q}^{t}\|\mathbf{d}_t\|^2$.

Next, we have

$$\begin{aligned}
&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i})-f_s(\mathbf{x}_*)\right]\\
=&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)-f_s(\mathbf{x}_{i+1})+f_s(\mathbf{x}_i)\right]\\
=&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s|f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_i)|\\
\overset{(a)}{\leq}&\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|}-[\frac{\eta}{4}-\frac{\eta^3q}{2}\frac{L^2}{|\mathcal{A}|}]-\frac{\mu}{4c_\gamma})\sum_{t=0}^{T}\|\mathbf{d}_i\|^2+\frac{\mu}{4}T\frac{\epsilon}{c_\epsilon},
\end{aligned}\tag{72}$$

where (a) follows from Eqs. (71).

Let $|\mathcal{A}|=q=\lceil\sqrt{n}\rceil$ and $\eta\leq\min\{\frac{1}{2\mu},\frac{1}{8L},\frac{\mu}{64L^2}\}, c_\gamma\geq\frac{8\mu}{\eta}$, we have $(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|}-[\frac{\eta}{4}-\frac{\eta^3q}{2}\frac{L^2}{|\mathcal{A}|}]-\frac{\mu}{4c_\gamma})>\frac{\eta}{32}>0$

Thus, we have

$$\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_i)-f_s(\mathbf{x}_*)\right]\leq\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right).\tag{73}$$

Then, we have

$$\mathbb{E}[\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_t)-f_s(\mathbf{x}_*)\right]]\leq\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\mathbb{E}\|\mathbf{x}_t-\mathbf{x}_*\|^2-\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2\right)+\frac{\mu}{4}T\frac{\epsilon}{c_\epsilon}.\tag{74}$$

Averaging using weight $w_t=(1-\frac{3\mu\eta}{4})^{1-t}$ and using such weight to pick output $\mathbf{x}$. By using Lemma 1 in Karimireddy et al. [2020] with $\eta\geq\frac{1}{uR},c_\epsilon>\frac{\mu}{2}$ and Assumption. 4, we have

$$
\begin{aligned}
\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}_0 - \mathbf{x}_*\|^2 \mu \exp(-\frac{3\eta\mu T}{4}) + \frac{\mu}{4} T \frac{\epsilon}{c_\epsilon} && (75)\\
&= \mathcal{O}(\mu \exp(-\mu T)). && (76)
\end{aligned}
$$

Then we have the convergence rate $\mathbb{E}\|\mathbf{x}_t - \mathbf{x}^*\|^2 = \mathcal{O}(\mu \exp(-\mu T))$.

The total sample complexity can be calculated as: $\lceil \frac{T}{q} \rceil n + T \cdot |\mathcal{A}| \leq \frac{T+q}{q} n + T\sqrt{n} = T\sqrt{n} + n + T\sqrt{n} = O(n + \sqrt{n}\ln(\mu/\epsilon)$. Thus, the overall sample complexity is $\mathcal{O}(n + \sqrt{n}\ln(\mu/\epsilon)$. This completes the proof. □

D PROOF OF CONVERGENCE OF STIMULUS-M$^+$

Proof of Theorem. 5 [Part 2]

Proof. From Lemma. 3, we have

$$
\begin{aligned}
&[f_s(\mathbf{x}_{t+1})]\\
&\overset{(a)}{\leq} [f_s(\mathbf{x}_t)] + \frac{\eta}{2} \sum_{i=0}^{t} \alpha^{(t-i)} \|\nabla f_s(\mathbf{x}_i) - \mathbf{u}_i^s\|^2 - \frac{1}{2}\eta \sum_{i=0}^{t} \alpha^{(t-i)} \|\mathbf{d}_i\|^2 + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\\
&\overset{(b)}{\leq} [f_s(\mathbf{x}_t)] - \frac{1}{2}\eta \sum_{i=0}^{t} \alpha^{(t-i)} \|\mathbf{d}_i\|^2 + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\\
&\quad + \frac{\eta}{2} \sum_{j=0}^{t} \alpha^{(t-j)} [\frac{L^2}{|\mathcal{A}|} \sum_{i=(n_t-1)q}^{j} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}^s_{(n_t-1)q}\|^2]\\
&\overset{(c)}{\leq} [f_s(\mathbf{x}_t)] - \frac{1}{2}\eta \sum_{i=0}^{t} \alpha^{(t-i)} \|\mathbf{d}_i\|^2 + \frac{1}{2} L \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \frac{\eta}{2} \sum_{j=0}^{t} \alpha^{(t-j)} [\frac{L^2}{|\mathcal{A}|} \sum_{i=(n_t-1)q}^{j} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2]\\
&\quad + \frac{\eta}{2} \sum_{i=0}^{t} \alpha^{(t-i)} (\frac{\gamma_i}{c_\gamma} + \frac{\epsilon}{c_\epsilon}), \qquad (77)
\end{aligned}
$$

where (a) follows from Lemma 3. (b) follows from Lemma. 1. (c) follows from Eqs. (51).

Next, telescoping the above inequality over t from $(n_t - 1)\,q$ to t where $t \leq n_t q - 1$ and noting that for $(n_t - 1)\,q \leq j \leq n_t q - 1, n_j = n_t$ and let $\eta \leq \frac{1}{4L}$, we obtain

$$
\begin{aligned}
&[f_s(\mathbf{x}_{t+1})]\\
&\overset{(a)}{\leq} [f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{2} \sum_{j=(n_t-1)q}^{t} \sum_{i=0}^{j} \alpha^{(j-i)} \|\mathbf{d}_i\|^2 + \frac{1}{2} L \sum_{i=(n_t-1)q}^{t} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2\\
&\quad + \frac{\eta}{2} \sum_{j=(n_t-1)q}^{t} \sum_{i=0}^{j} \alpha^{(j-i)} [\frac{L^2}{|\mathcal{A}|} \sum_{r=(n_t-1)q}^{i} \|\mathbf{x}_{r+1} - \mathbf{x}_r\|^2]\\
&\quad + \frac{\eta}{2} \sum_{j=(n_t-1)q}^{t} \sum_{i=0}^{j} \alpha^{(j-i)} (\frac{[\gamma_i]}{c_\gamma} + \frac{\epsilon}{c_\epsilon})\\
&\overset{(b)}{\leq} [f_s(\mathbf{x}_{(n_t-1)q})] - \frac{\eta}{2} \sum_{j=(n_t-1)q}^{t} \sum_{i=0}^{j} \alpha^{(j-i)} \|\mathbf{d}_i\|^2 + \frac{1}{2} L \sum_{i=(n_t-1)q}^{t} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[\frac{L^2}{|\mathcal{A}|}q\|\mathbf{x}_{j+1}-\mathbf{x}_j\|^2]\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon})\\
\overset{(c)}{=}&[f_s(\mathbf{x}_{(n_t-1)q})]-\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\|\mathbf{d}_i\|^2+\frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[L^2\|\mathbf{x}_{j+1}-\mathbf{x}_j\|^2]\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon})\\
\overset{(d)}{\leq}&[f_s(\mathbf{x}_{(n_t-1)q})]-\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\|\mathbf{d}_i\|^2+\frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[L^2\|\eta\sum_{r=0}^{j}\alpha^{(j-r)}\mathbf{d}_r\|^2]\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon})\\
=&[f_s(\mathbf{x}_{(n_t-1)q})]-\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\|\mathbf{d}_i\|^2+\frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{3(j-i)}[L^2\eta^2\|\mathbf{d}_i\|^2]\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon})\\
\overset{(e)}{\leq}&[f_s(\mathbf{x}_{(n_t-1)q})]-\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\|\mathbf{d}_i\|^2+\frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}[L^2\eta^2\|\mathbf{d}_i\|^2]\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon})\\
\overset{(f)}{\leq}&[f_s(\mathbf{x}_{(n_t-1)q})]-\frac{\eta}{4}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\|\mathbf{d}_i\|^2+\frac{1}{2}L\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2\\
&+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon})\\
\overset{(g)}{\leq}&[f_s(\mathbf{x}_{(n_t-1)q})]-\frac{\eta}{4}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\|\mathbf{d}_i\|^2+\frac{1}{2}L\sum_{j=(n_t-1)q}^{t}\|\eta\sum_{i=0}^{j}\alpha^{(j-i)}\mathbf{d}_j\|^2
\end{aligned}
$$

$$+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon})$$

$$\overset{(h)}{\leq}[f_s(\mathbf{x}_{(n_t-1)q})]-\frac{\eta}{8}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}\|\mathbf{d}_i\|^2+\frac{\eta}{2}\sum_{j=(n_t-1)q}^{t}\sum_{i=0}^{j}\alpha^{(j-i)}(\frac{[\gamma_i]}{c_\gamma}+\frac{\epsilon}{c_\epsilon}), \tag{78}$$

where (a) follows from Eqs. (77). (b) follows from $i \leq n_t q$. (c) follows from $q = |\mathcal{A}| = \lceil \sqrt{n} \rceil$. (d) and (g) follow from the update rule of $\mathbf{x}_t$ shown in Line 19 in Algorithm. 1. (e) follows from $0 < \alpha < 1$, then we have $\alpha^2(j-i) < \alpha^{(j-i)}$. (f) and (h) follow from $\eta \leq \frac{1}{4L}$ Recall that $\gamma_t = \frac{1}{q}\sum_{i=(n_t-1)q}^{t}\|\mathbf{d}_t\|^2$. Then, we have

$$\begin{aligned}
&\mathbb{E}[f_s(\mathbf{x}_T)]-[f_s(\mathbf{x}_0)]\\
&=\mathbb{E}([f_s(\mathbf{x}_q)]-[f_s(\mathbf{x}_0)])+([f_s(\mathbf{x}_{2q})]-[f_s(\mathbf{x}_q)])+\cdot+([f_s(\mathbf{x}_T)]-[f_s(\mathbf{x}_{(n_T-1)q})])\\
&\overset{(a)}{\leq}-[\frac{\eta}{8}]\sum_{t=0}^{T-1}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_t\|^2+\frac{\eta}{2c_\gamma}\sum_{t=0}^{T-1}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_t\|^2+\frac{\eta}{2}Tq\frac{\epsilon}{c_\epsilon}\\
&\overset{(b)}{\leq}-[\frac{\eta}{16}]\sum_{t=0}^{T-1}\sum_{i=0}^{j}\alpha^{(j-i)}\mathbb{E}\|\mathbf{d}_t\|^2+\frac{\eta}{2}Tq\frac{\epsilon}{c_\epsilon}\\
&\overset{(c)}{\leq}-[\frac{\eta}{16}]\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2+\frac{\eta}{2}Tq\frac{\epsilon}{c_\epsilon},
\end{aligned} \tag{79}$$

where (a) follows from $c_\gamma \geq 8$, (c) follows from $0 < \alpha < 1$.

Note that $[f_s(\mathbf{x}_{T+1})] \geq f_s^* \triangleq \inf_{\mathbf{x}\in\mathbb{R}^d} f_s(\mathbf{x})$. Hence, we have

$$[\frac{\eta}{16}]\sum_{t=0}^{T-1}\|\mathbf{d}_t\|^2 \leq [[f_s(\mathbf{x}_0)]-[f_s(\mathbf{x}_T)]] \leq [[f_s(\mathbf{x}_0)]-f_s^*]. \tag{80}$$

Based on the parameter setting $q^2 = |\mathcal{A}| = \sqrt{n}$, we have

$$[\frac{\eta}{16}]\sum_{t=0}^{T-1}\|\mathbf{d}_t\|^2 \leq [[f_s(\mathbf{x}_0)]-f_s^*]. \tag{81}$$

Thus, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\mathbf{d}_t\|^2 \leq \frac{[[f_s(\mathbf{x}_0)]-f_s^*]}{[\frac{\eta}{16}]T}. \tag{82}$$

Since $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|d_t\|^2$ is just common descent directions. According to Definition. 3 shown in the paper, the quantity to our interest is $\|\sum_{s\in[S]}\lambda_t^s\nabla f(\mathbf{x})\|^2$.

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\sum_{s\in[S]}\lambda_t^s\nabla f_s(\mathbf{x}_t)\|^2 \overset{(a)}{\leq} (2SL^2\eta^2+2)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{d}_t\|^2 \tag{83}$$

where (a) follows from Eqs. (20).

Then, we can conclude that

$$\frac{1}{T}\sum_{t=0}^{T-1}\min_{\boldsymbol{\lambda}\in C}\mathbb{E}\|\boldsymbol{\lambda}^\top\nabla\mathbf{F}(\mathbf{x}_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\sum_{s\in[S]}\lambda_t^s\nabla f_s(\mathbf{x}_t)\|^2 = \mathcal{O}(\frac{1}{T}). \tag{84}$$

The total sample complexity can be calculated as: $\lceil\frac{T}{q}\rceil n + T\cdot|\mathcal{A}| \leq \frac{T+q}{q}n + T\sqrt{n} = T\sqrt{n} + n + T\sqrt{n} = O(n+\sqrt{n}\epsilon^{-1})$. Thus, the overall sample complexity is $\mathcal{O}(n+\sqrt{n}\epsilon^{-1})$. This completes the proof.

□

D.1 PROOF OF THEOREM. 6 [PART 2]

Proof.

$$\begin{aligned}
&f_s(\mathbf{x}_{t+1})\\
\overset{(a)}{\leq}&f_s(\mathbf{x}_t) + \left\langle \nabla f_s(\mathbf{x}_t), -\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle + \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
\overset{(b)}{\leq}&f_s(\mathbf{x}_*) + \langle \nabla f_s(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* \rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \left\langle \nabla f_s(\mathbf{x}_t), -\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle\\
&+ \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
=&f_s(\mathbf{x}_*) + \left\langle \nabla f_s(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
=&f_s(\mathbf{x}_*) + \left\langle \nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle + \left\langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle\\
&- \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
\overset{(c)}{\leq}&f_s(\mathbf{x}_*) + \frac{1}{2\delta}\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{\delta}{2}\|\mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
&+ \left\langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle\\
&- \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
\overset{(d)}{\leq}&f_s(\mathbf{x}_*) + \frac{1}{2\delta}\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
&+ \left\langle \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2, \qquad (85)
\end{aligned}$$

where (a) follows from L-smoothness assumption, (b) follows from μ-strongly convex. (c) and (d) follow from the triangle inequality.

$$\begin{aligned}
&\sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{t+1}) - f_s(\mathbf{x}_*)\right] \qquad (86)\\
\overset{(a)}{\leq}&\frac{1}{2\delta}\sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
&+ \left\langle \sum_{s\in[S]} \lambda_t^s\mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \left\langle \sum_{s\in[S]} \lambda_t^s\mathbf{u}_t^s, -\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i \right\rangle\\
&+ \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2\\
=&\frac{1}{2\delta}\sum_{s\in[S]} \lambda_t^s\|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta \sum_{t=0}^{T} \alpha^{(t-i)}\mathbf{d}_i\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\left\langle \sum_{s\in[S]} \lambda_t^s \mathbf{u}_t^s, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 \\
&+ \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
=&\frac{1}{2\delta}\sum_{s\in[S]} \lambda_t^s \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \delta\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \delta\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&+\left\langle \mathbf{d}_t, \mathbf{x}_t - \mathbf{x}_* - \eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i \right\rangle - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
\overset{(b)}{\le}& \frac{1}{2\eta}\left(\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - \frac{1}{2}\eta\|\sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}_*\|^2 \\
&+ \frac{1}{2}L\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&+ \frac{4}{\mu}\sum_{s\in[S]} \lambda_t^s \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 + \frac{\mu}{8}\|\mathbf{x}_t - \mathbf{x}_*\|^2 + \frac{\mu}{8}\|\eta \sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
=&\frac{1}{2\eta}\left((1 - \frac{3\mu\eta}{4})\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2)\|\sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&+ \frac{4}{\mu}\sum_{s\in[S]} \lambda_t^s \|\nabla f_s(\mathbf{x}_t) - \mathbf{u}_t^s\|^2 \\
\overset{(c)}{\le}& \frac{1}{2\eta}\left((1 - \frac{3\mu\eta}{4})\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2)\|\sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&+ \frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 + \sum_{s\in[S]} \lambda_t^s \|\nabla f_s(\mathbf{x}_{(n_t-1)q}) - \mathbf{u}_{(n_t-1)q}^s\|^2) \\
\overset{(d)}{\le}& \frac{1}{2\eta}\left((1 - \frac{3\mu\eta}{4})\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2\right) - (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2)\|\sum_{t=0}^{T} \alpha^{(t-i)} \mathbf{d}_i\|^2 \\
&+ \frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{i=(n_t-1)q}^{t} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2) + \frac{\mu S}{4}\frac{I_{(\mathcal{N}_s<n)}}{\mathcal{N}_s}\sigma^2.
\end{aligned}
\tag{87}
$$

where (a) follows from Eqs. (85), (b) follows from $\|\mathbf{x}_t - \mathbf{x}_*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 = -\eta^2\|\mathbf{d}_t\|^2 + 2\langle \eta\mathbf{d}_t, \mathbf{x}_t - \mathbf{x}_*\rangle$ and we choose $\delta = \frac{\mu}{8}$. (c) is from Lemma. 1. (d) is from Eqs. (51). (d) follows from $0 < \lambda_t^s < 1, \forall s \in [S]$

Next, telescoping the above inequality over t from $(n_t - 1)\,q$ to t where $t \le n_t q - 1$ and noting that for $(n_t - 1)\,q \le j \le n_t q - 1, n_j = n_t$, we obtain

$$
\begin{aligned}
&\sum_{i=(n_t-1)q}^{t} \sum_{s\in[S]} \lambda_t^s \left[f_s(\mathbf{x}_{i+1}) - f_s(\mathbf{x}_*)\right] \\
\overset{(a)}{\le}& \frac{1}{2\eta}\left((1 - \frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t} \|\mathbf{x}_i - \mathbf{x}_*\|^2 - \sum_{i=(n_t-1)q}^{t} \|\mathbf{x}_{i+1} - \mathbf{x}_*\|^2\right) \\
&- (\frac{1}{2}\eta - \frac{\mu}{8}\eta^2 - \frac{1}{2}L\eta^2)\sum_{i=(n_t-1)q}^{t} \|\sum_{i=0}^{t} \alpha^{(t-i)} \mathbf{d}_i\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_j-1)q}^{j}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)+\frac{\mu}{4c_\gamma}\sum_{i=(n_t-1)q}^{t}\|\alpha^{(t-i)}\mathbf{d}_i\|^2\\
&+\frac{\mu}{4c_\gamma}\sum_{t=(n_t-1)q}^{t}\|\alpha^{(t-i)}\mathbf{d}_i\|^2++\frac{\mu}{4}\sum_{i=(n_t-1)q}^{t}\frac{\epsilon}{c_\epsilon}\\
\overset{(b)}{\le}&\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)\\
&-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2)\sum_{i=(n_t-1)q}^{t}\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2\\
&+\frac{4}{\mu}(\frac{L^2}{|\mathcal{A}|}\sum_{j=(n_t-1)q}^{t}\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_i\|^2)\\
&+\frac{\mu}{4c_\gamma}\sum_{t=(n_t-1)q}^{t}\|\alpha^{(t-i)}\mathbf{d}_i\|^2++\frac{\mu}{4}\sum_{i=(n_t-1)q}^{t}\frac{\epsilon}{c_\epsilon}\\
\overset{(c)}{\le}&\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{i=(n_t-1)q}^{t}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)+\frac{\mu}{4}\sum_{i=(n_t-1)q}^{t}\frac{\epsilon}{c_\epsilon}\\
&-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|})\sum_{i=(n_t-1)q}^{t}\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2)\\
&+\frac{\mu}{4c_\gamma}\sum_{t=(n_t-1)q}^{t}\|\alpha^{(t-i)}\mathbf{d}_i\|^2+\frac{\mu}{4}\sum_{i=(n_t-1)q}^{t}\frac{\epsilon}{c_\epsilon},
\end{aligned}
\tag{88}
$$

where (a) follows from Eqs. (86), (b) extends j to t. (c) follows from $t \le n_t q - 1$.

We continue the proof by further driving

$$
\begin{aligned}
&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]\\
=&\sum_{i=0}^{q}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+\sum_{i=q}^{2q}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+\cdot+\sum_{i=(n_T-1)q}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]\\
\le&\frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|})\sum_{t=0}^{T}\|\sum_{t=0}^{T}\alpha^{(t-i)}\mathbf{d}_i\|^2)\\
&+\frac{\mu}{4c_\gamma}\sum_{t=0}^{T}\|\alpha^{(t-i)}\mathbf{d}_i\|^2+\frac{\mu}{4}T\frac{\epsilon}{c_\epsilon}.
\end{aligned}
\tag{89}
$$

Next, we have

$$
\begin{aligned}
&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_i)-f_s(\mathbf{x}_*)\right]\\
=&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)-f_s(\mathbf{x}_{i+1})+f_s(\mathbf{x}_i)\right]\\
\le&\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_*)\right]+\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s|f_s(\mathbf{x}_{i+1})-f_s(\mathbf{x}_i)|
\end{aligned}
$$

$$\overset{(a)}{\leq} \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)$$
$$-(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|}-[\frac{\eta}{4}-\frac{\eta^3q}{2}\frac{L^2}{|\mathcal{A}|}]-\frac{\mu}{4c_\gamma})\sum_{t=0}^{T}\|\alpha^{(t-i)}\mathbf{d}_i\|^2+\frac{\mu}{4}T\frac{\epsilon}{c_\epsilon}, \tag{90}$$

where (a) follows from Eqs. (89). Let $|\mathcal{A}| = q = \lceil\sqrt{n}\rceil$ and $\eta \leq \min\{\frac{1}{2\mu}, \frac{1}{8L}, \frac{\mu}{64L^2}\}, c_\gamma \geq \frac{8\mu}{\eta}, c_\epsilon \geq \frac{\mu}{2}$, we have $(\frac{1}{2}\eta-\frac{\mu}{8}\eta^2-\frac{1}{2}L\eta^2-\frac{4}{\mu}\frac{L^2q\eta^2}{|\mathcal{A}|}-[\frac{\eta}{4}-\frac{\eta^3q}{2}\frac{L^2}{|\mathcal{A}|}]-\frac{\mu}{4c_\gamma}) > \frac{\eta}{32} > 0$

Thus, we have

$$\sum_{t=0}^{T}\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_i)-f_s(\mathbf{x}_*)\right]$$
$$\leq \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\sum_{i=0}^{T}\|\mathbf{x}_i-\mathbf{x}_*\|^2-\sum_{t=0}^{T}\|\mathbf{x}_{i+1}-\mathbf{x}_*\|^2\right)+\frac{\epsilon}{2}. \tag{91}$$

Then, we have

$$\mathbb{E}[\sum_{s\in[S]}\lambda_t^s\left[f_s(\mathbf{x}_t)-f_s(\mathbf{x}_*)\right]]$$
$$\leq \frac{1}{2\eta}\left((1-\frac{3\mu\eta}{4})\mathbb{E}\|\mathbf{x}_t-\mathbf{x}_*\|^2-\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_*\|^2\right)+\frac{\epsilon}{2}. \tag{92}$$

Based on Assumption. 4 and averaging using weight $w_t = (1-\frac{3\mu\eta}{4})^{1-t}$ and using such weight to pick output $\mathbf{x}$. By using Lemma 1 in Karimireddy et al. [2020] with $\eta \geq \frac{1}{uR}$, we have

$$\mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*\|^2 \leq \|\mathbf{x}_0-\mathbf{x}_*\|^2\mu\exp(-\frac{3\eta\mu T}{4}) \tag{93}$$
$$= \mathcal{O}(\mu\exp(-\mu T)). \tag{94}$$

Then we have the convergence rate $\mathbb{E}\|\mathbf{x}_t-\mathbf{x}^*\|^2 = \mathcal{O}(\mu\exp(-\mu T))$.

The total sample complexity can be calculated as: $\lceil\frac{T}{q}\rceil n+T\cdot|\mathcal{A}| \leq \frac{T+q}{q}n+T\sqrt{n} = T\sqrt{n}+n+T\sqrt{n} = O(n+\sqrt{n}\ln(\mu/\epsilon)$. Thus, the overall sample complexity is $\mathcal{O}(n+\sqrt{n}\ln(\mu/\epsilon)$. This completes the proof.

□

E ADDITIONAL EXPERIMENT RESULTS

1) Strongly-Convex Optimization:

We conducted experiments to assess the performance of our algorithms on a strongly-convex optimization problem, where $\mathbf{F}(\mathbf{x}) = [f_1(\mathbf{x}) = \mathbf{x}^2, f_2(\mathbf{x}) = e^{-\mathbf{x}}]$. For this experiment, we selected hyperparameters $\eta = 0.005$ and $\alpha = 0.3$, while introducing stochasticity into the gradient by adding Gaussian noise with a range of (-1, 1). As shown in Fig. 3, it is evident that all of the algorithms successfully achieved convergence. Notably, the momentum-based algorithms, namely MOCO, STIMULUS-M, and STIMULUS-M$^+$, exhibited faster convergence compared to MGD, MSGD, STIMULUS, and STIMULUS$^+$. We would also like to note that there isn't a significant difference between the stochastic algorithms (SMGD, MGD) and other algorithms. This is not necessarily because the stochastic algorithms are inferior, but perhaps because the strongly-convex function in question is too simplistic.

2) Eight-Objective Experiments on River Flow Dataset:

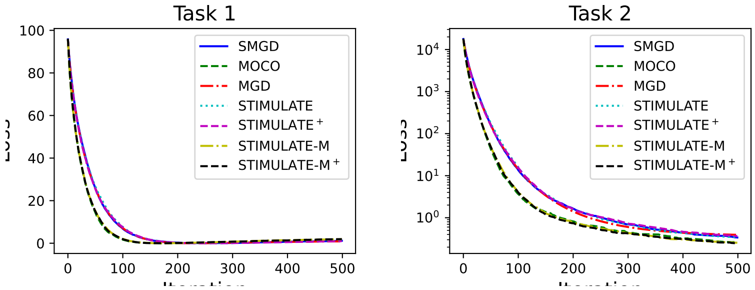


Figure 3: Convergence comparison on strongly-convex optimization problem.

We further test our algorithms on an 8-task problem with the river flow dataset [Nie et al., 2017], which is for flow prediction at eight locations in the Mississippi river network. In this experiment, we set $\eta = 0.001, \alpha = 0.1$, the batch size for MOCO, CR-MOGM and SMGD is 8, the full batch size for MGD is 128, and the inner loop batch size $|\mathcal{N}_s|$ for STIMULUS, STIMULUS-M, STIMULUS$^+$, STIMULUS-M$^+$is eight. To better visualize different tasks, we plot the normalized loss in a radar chart as shown in Fig. 4, where we can see that our STIMULUS algorithms achieve a much smaller footprint, which is desirable. Further, we compare the sample complexity results of all algorithms in Table 3, which reveals a significant reduction in sample utilization by STIMULUS$^+$ /STIMULUS-M$^+$ compared to MGD, while achieving a much better loss compared to SGMD and MOCO (cf. Fig. 4).

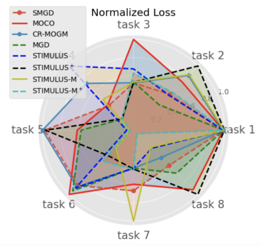


Figure 4: Training loss convergence comparison (8-objective).

Table 3: Results of normalized loss with the river flow dataset and learning tasks.

	# of samples	Tasks							
		0	1	2	3	4	5	6	7
SMGD	8000	0.985	0.558	0.521	0.384	1	0.862	0.667	0.550
MOCO	8000	0.985	0.753	1	0.399	0.632	1	0.595	0.926
MGDA	128000	0.989	0.396	0.532	0.174	0.589	0.945	0.417	0.669
STIMULUS	27200	0.985	0.546	0.675	1	0.077	0.898	0.417	0.281
STIMULUS$^+$	20947	0.996	1	0.528	0.178	0.990	0.395	0.427	1
STIMULUS-M	27200	0.996	0.864	0.530	0.475	0.036	0.271	1	0.264
STIMULUS-M$^+$	21085	1	0.596	0.627	0.1781	0.0376	0.482	0.430	0.055

3) Ablation study on momentum in STIMULUS-M:

Table 4: Loss value vs. Iteration on tasks L and R of STIMULUS-M.

Momentum Term α	Task L				Task R			
	100	200	300	500	100	200	300	500
0.1	0.0228	0.0207	0.0203	0.0153	0.0229	0.0223	0.0205	0.0215
0.3	0.0228	0.0182	0.0179	0.0120	0.0223	0.0191	0.0168	0.0143
0.5	0.0227	0.0174	0.0146	0.0078	0.0215	0.0180	0.0124	0.0091
0.8	0.0225	0.0158	0.0127	0.0065	0.0210	0.0152	0.0113	0.0078

We performed additional experiments to analyze the impact of varying the momentum term in our proposed STIMULUS-M algorithm, as shown in Table 4, on the classification task of the MultiMNIST dataset. The experimental settings are consistent with those in Section 5.1 of the main paper. These results indicate that a larger momentum term leads to faster convergence.