Journal of Machine Learning Research 25 (2024) 1-52 Submitted 10/23; Revised 3/24; Published 5/24

:

Three-Way Trade-Off in Multi-Objective Learning: Optimization, Generalization and Conflict-Avoidance

Lisha Chen† CHENL21@RPI.EDU
Department of Electrical, Computer & Systems Engineering
Rensselaer Polytechnic Institute, United States

Heshan Fernando† FERNAH@RPI.EDU
Department of Electrical, Computer & Systems Engineering
Rensselaer Polytechnic Institute, United States

Yiming Ying YIMING.YING@SYDNEY.EDU.AU
School of Mathematics and Statistics
University of Sydney, NSW, Australia

Tianyi Chen CHENTIANYI19@GMAIL.COM
Department of Electrical, Computer & Systems Engineering
Rensselaer Polytechnic Institute, United States

Editor: Francesco Orabona

Abstract

Multi-objective learning (MOL) often arises in machine learning problems when there are multiple data modalities or tasks. One critical challenge in MOL is the potential conflict among different objectives during the optimization process. Recent works have developed various *dynamic weighting* algorithms for MOL, where the central idea is to find an update direction that *avoids conflicts* among objectives. Albeit its appealing intuition, empirical studies show that dynamic weighting methods may not outperform static ones. To understand this theory-practice gap, we focus on a stochastic variant of MGDA - the Multi-objective gradient with Double sampling (MoDo), and study the generalization performance and its interplay with optimization through the lens of algorithmic stability in the framework of statistical learning theory. We find that the key rationale behind MGDA – updating along conflict-avoidant direction - may *hinder* dynamic weighting algorithms from achieving the optimal $\mathcal{O}(1/\sqrt{n})$ population risk, where n is the number of training samples. We further demonstrate the impact of dynamic weights on the three-way trade-off among optimization, generalization, and conflict avoidance unique in MOL. We showcase the generality of our theoretical framework by analyzing other algorithms under the framework. Experiments on various multi-task learning benchmarks are performed to demonstrate the practical applicability. Code is available at `https://github.com/heshandevaka/Trade-Off-MOL`.

Symbol † denotes equal contribution. Preliminary results in this paper were presented in part at the 2023 Advances in Neural Information Processing Systems (Chen et al., 2023b). The work of L. Chen, H. Fernando, and T. Chen was supported by the National Science Foundation (NSF) CAREER project 2047177, the Cisco Research Award, and the RPI-IBM Artificial Intelligence Research Collaboration (AIRC). The work of Y. Ying was partially supported by NSF (DMS-2110836, IIS-2103450, and IIS-2110546).

Keywords: Multi-objective optimization, statistical learning theory, algorithm stability, Pareto stationarity, gradient conflict

1 Introduction

Multi-objective learning (MOL) emerges frequently as a new unified learning paradigm from recent machine learning problems such as learning under fairness and safety constraints (Zafar et al., 2017); learning across multiple tasks (Sener and Koltun, 2018); and, learning across multiple agents that may not share a global utility (Moffaert and Nowé, 2014).

This work considers solving the empirical version of MOL defined on the training dataset as $S = \{z_1, \ldots, z_n\}$. The performance of a model $x \in \mathbb{R}^d$ on a datum z for the m-th objective is denoted as $f_{z,m} : \mathbb{R}^d \mapsto \mathbb{R}$, and its performance on the entire training dataset S is measured by the m-th empirical objective $f_{S,m}(x)$ for $m \in [M]$. MOL optimizes the vector-valued objective, given by

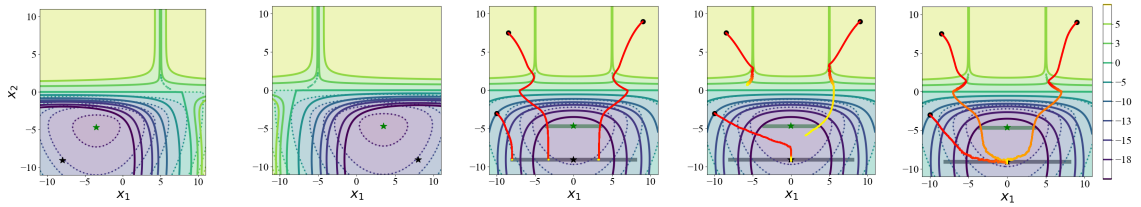
$$\min_{x \in \mathbb{R}^d} \quad F_S(x) \coloneqq [f_{S,1}(x), \ldots, f_{S,M}(x)]. \tag{1.1}$$

One natural method for solving (1.1) is to optimize the (weighted) average of the multiple objectives, also known as *static or unitary weighting* (Kurin et al., 2022; Xin et al., 2022). However, this method may face challenges due to *potential conflicts* among multiple objectives during the optimization process; e.g., conflicting gradient directions $\langle \nabla f_{S,m}(x), \nabla f_{S,m'}(x) \rangle < 0$, if choosing the gradient-based optimizer. A popular alternative is thus to *dynamically weight* gradients from different objectives to avoid conflicts and obtain a direction $d(x)$ that optimizes all objective functions jointly that we call a *conflict-avoidant* (CA) direction. Algorithms in this category include the multi-gradient descent algorithm (MGDA) (Désidéri, 2012), its stochastic variants (Liu and Vicente, 2021; Fernando et al., 2023; Zhou et al., 2022). While the idea of finding CA direction in dynamic weighting-based approaches is very appealing, recent empirical studies reveal that dynamic weighting methods may not outperform static weighting in some MOL benchmarks (Kurin et al., 2022; Xin et al., 2022), especially when it involves stochastic updates and deep models. Specifically, observed by (Kurin et al., 2022), the vanilla stochastic MGDA can be under-optimized, leading to larger optimization error than static weighting. The reason behind this optimization performance degradation has been studied in (Zhou et al., 2022; Fernando et al., 2023), which suggest the vanilla stochastic MGDA has biased update, and propose momentum-based methods to address this issue. Nevertheless, in (Xin et al., 2022), it has been demonstrated that the training errors of MGDA and static weighting are similar, while their main difference lies in the *generalization performance*. Unfortunately, the reason behind this testing performance degradation is not fully understood and remains an open question.

To gain a deeper understanding of the dynamic weighting methods, a natural question is

Q1: *What are the major sources of errors in dynamic weighting-based MOL methods?*

To answer this question theoretically, we first introduce a proper measure of testing performance in MOL – the *Pareto stationary measure* in terms of the population objectives, which will immediately imply stronger measures such as Pareto optimality under strongly convex objectives. We then decompose this measure into *generalization* error and *optimization* error and further introduce a new metric termed *CA distance* that reflects the algorithm's ability to update along CA direction and is unique to MOL; see Sections 2.1 and 2.2.



(a) Objective 1 (b) Objective 2 (c) MGDA (d) Static (e) MoDo (dynamic)

Figure 1: An example from (Liu et al., 2021a) with two objectives (1a and 1b) to show the three-way trade-off in MOL. Figures 1c-1e show the optimization trajectories, where the **black** • marks initializations of the trajectories, colored from **red** (start) to **yellow** (end). The background solid/dotted contours display the landscape of the average empirical/population objectives. The **gray**/**green** bar marks empirical/population Pareto front, and the **black** ⋆/**green** ⋆ marks solution to the average objectives.

To characterize the performance of MOL methods in a unified manner, we introduce a generic dynamic weighting-based MOL method that we term stochastic Multi-Objective gradient with DOuble sampling algorithm (**MoDo**), which uses a step size γ to control the change of dynamic weights. Roughly speaking, by controlling γ, MoDo includes MGDA (large γ) and static weighting algorithm ($\gamma = 0$) as two special cases; see Section 2.3. We first analyze the generalization error of the model learned by MoDo through the lens of algorithmic stability (Bousquet and Elisseeff, 2002; Hardt et al., 2016; Lei and Ying, 2020) in the framework of statistical learning theory. To our best knowledge, this is the *first-ever-known* stability analysis for MOL algorithms. Here the key contributions lie in defining a new notion of stability - MOL uniform stability and then establishing a tight upper bound (matching lower bound) on the MOL uniform stability for the MoDo algorithm that involves two coupled sequences; see Section 3.1. Note that the only other existing stability analysis for two coupled sequences is for minmax problems, where the two sequences are optimizing the opposite objective functions w.r.t. two variables, and the variables are stacked into one to derive similar properties as single objective learning. In contrast, our analysis for two sequences with different objectives is of self-interest and generalizes to other settings such as bilevel and compositional optimization. We then analyze the optimization error of MoDo and its distance to CA directions, where the key contributions lie in relaxing *the bounded function value/gradient assumptions* and significantly improving the convergence rate of state-of-the-art dynamic weighting-based method (Fernando et al., 2023); see Section 3.2.

Different from the stability analysis for single-objective learning (Hardt et al., 2016), the techniques used in our generalization and optimization analysis allow to remove conflicting assumptions such as bounded gradient and bounded function value assumptions in the strongly convex and unconstrained setting, as well as use step sizes larger than $\mathcal{O}(1/t)$ in (Hardt et al., 2016) to ensure both small generalization and optimization errors, which are of independent interest.

Given the test performance degradation of dynamic weighting methods in MOL and the holistic analysis of dynamic weighting methods provided in **Q1**, a follow-up question is

Q2: *What may cause the empirical performance degradation of dynamic weighting methods?*

Visualizing MOL solution concepts. To reason the root cause for this, we first compare different MOL algorithms in a toy example shown in Figure 1. We find MGDA can navigate along CA directions and converge to the empirical Pareto front under all initializations, while static weighting gets stuck in some initializations; at the same time, the empirical Pareto solution obtained by MGDA may incur a larger population risk than the suboptimal empirical solution obtained by the static weighting method; finally, if the step size γ of dynamic weights is carefully tuned, MoDo can converge along CA directions to the empirical Pareto optimal solution that also generalizes well.

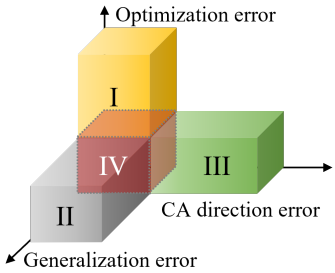
Aligned with this toy example, our theoretical results suggest a novel *three-way trade-off* in the performance of dynamic weighting-based MOL algorithm; see Section 3.3. Specifically, it suggests that the step size for dynamic weighting γ plays a central role in the trade-off among convergence to the CA direction, convergence to empirical Pareto stationarity, and generalization error; see Figure 2. In this sense, MGDA has a relative advantage in convergence to the CA direction to escape suboptimal solutions compared to the static weighting method but it could sacrifice generalization; the static weighting method cannot converge to the CA direction but guarantees convergence to empirical Pareto solutions and their generalization. Our analysis also suggests that MoDo achieves a small population risk under a proper combination of step sizes and the number of iterations.

The major **technical challenges** and how we address them are summarized below.

C1) The definition of testing risk (2.1) is unique in MOL, and the introduction of sampling-determined algorithms overcomes a key challenge brought by the classical function value-based risk measures – the unnecessarily small step size choice. Specifically, prior stability analysis in function values for single objective learning (Hardt et al., 2016) requires $1/t$ step size decay in the nonconvex case, otherwise, the generalization error bound will depend exponentially on the number of iterations. However, such step size choice leads to a very slow convergence of the optimization error. This is addressed by the definitions of gradient-based measures and sampling-determined MOL algorithms, which yield stability bound in $\mathcal{O}(T/n)$; see more discussions below Theorem 3.1.

C2) The stability of the dynamic weighting algorithm in the strongly convex (SC) case is non-trivial compared to single objective learning (Hardt et al., 2016) because it involves two coupled sequences during the update. As a result, the classical contraction property for the update of model parameters that is often used to derive stability does not hold. This is addressed by controlling the change of λ_t by the step size γ, and using mathematical induction to derive a tighter bound; see Appendix A.5.

C3) In the SC case with an unbounded domain, the function is not Lipschitz or the gradients are not uniformly bounded, which violates the commonly used bounded gradient assumption for proving the stability bound and optimization convergence. Different from existing approaches in single-objective learning (Nguyen et al., 2018; Lei and Ying, 2020), which cannot be directly applied to our MOL setting with dynamic weighting, we relax this assumption by proving that the iterates generated by dynamic weighting algorithms in the SC case are bounded on the trajectory in Lemma 3.1.



Region	Opt.	Gen.	Conflict	Step sizes α & γ	Iterations T & data size n
I	↑	↓	↓	$\gamma = \omega(\alpha)$	$T = o(n)$
II	↓	↑	↓	$\gamma = \omega(\alpha)$	$T = \omega(n)$
III	↓	↓	↑	$\gamma = o(T^{-1})$	$T = \omega(n)$
IV	↘	↘	↘	$\gamma = \Theta(\alpha^{\frac{1}{2}}) = \Theta(T^{-\frac{1}{4}})$	$T = \Theta(n^{\frac{4}{5}})$

Figure 2: An illustration of three-way trade-off among optimization, generalization, and conflict avoidance in the strongly convex case; α is the step size for x, γ is the step size for weights λ, where $o(\cdot)$ denotes a strictly slower growth rate, $\omega(\cdot)$ denotes a strictly faster growth rate, and $\Theta(\cdot)$ denotes the same growth rate. Arrows ↓ and ↑ respectively represent diminishing in an optimal rate and growing in a fast rate w.r.t. n, while ↘ represents diminishing w.r.t. n, but not in an optimal rate.

2 Problem Formulation and Target of Analysis

In this section, we first introduce the problem formulation of MOL, the target of analysis, and then present the MGDA algorithm and its stochastic variant.

2.1 Preliminaries of MOL

Denote the vector-valued objective function on datum z as $F_z(x) = [f_{z,1}(x), \ldots, f_{z,M}(x)]$. The training and testing performance of x can then be measured by the empirical objective $F_S(x)$ and the population objective $F(x)$ which are, respectively, defined as $F_S(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} F_{z_i}(x)$ and $F(x) \coloneqq \mathbb{E}_{z\sim\mathcal{D}}[F_z(x)]$. Their gradients are denoted as $\nabla F_S(x)$ and $\nabla F(x) \in \mathbb{R}^{d\times M}$.

Analogous to the stationary and optimal solutions in single-objective learning, we define Pareto stationary and Pareto optimal solutions for MOL problem $\min_{x\in\mathbb{R}^d} F(x)$ as follows.

Definition 2.1 (Pareto stationary and Pareto optimal solutions). *If there exists a convex combination of the gradient vectors that equals to zero, i.e., there exists* $\lambda \in \Delta^M \coloneqq \{\lambda \in \mathbb{R}^M \mid \mathbf{1}^\top\lambda = 1,\ \lambda \geq 0\}$ *such that* $\nabla F(x)\lambda = 0$*, then* $x \in \mathbb{R}^d$ *is Pareto stationary. If there is no* $x \in \mathbb{R}^d$ *and* $x \neq x^*$ *such that, for all* $m \in [M]$ $f_m(x) \leq f_m(x^*)$*, with* $f_{m'}(x) < f_{m'}(x^*)$ *for at least one* $m' \in [M]$*, then* x^* *is Pareto optimal. If there is no* $x \in \mathbb{R}^d$ *such that for all* $m \in [M]$*,* $f_m(x) < f_m(x^*)$*, then* x^* *is weakly Pareto optimal.*

By definition, at a Pareto stationary solution, there is no common descent direction for all objectives. A necessary and sufficient condition for x being Pareto stationary for smooth objectives is that $\min_{\lambda\in\Delta^M} \|\nabla F(x)\lambda\| = 0$ (Tanabe et al., 2019). Therefore, $\min_{\lambda\in\Delta^M} \|\nabla F(x)\lambda\|$ can be used as a measure of Pareto stationarity (PS) (Désidéri, 2012; Fliege et al., 2019; Tanabe et al., 2019; Liu and Vicente, 2021; Fernando et al., 2023). We will denote $R_{\text{pop}}(x) := \min_{\lambda\in\Delta^M} \|\nabla F(x)\lambda\|$ and refer to it as the *PS population risk* henceforth and its empirical version $R_{\text{opt}}(x) := \min_{\lambda\in\Delta^M} \|\nabla F_S(x)\lambda\|$ as *PS empirical risk* or *PS optimization error*. We next introduce the target of our analysis based on the above definitions.

2.2 Target of analysis and error decomposition

In the existing generalization analysis for MOL, measures based on function values have been used to derive generalization guarantees in terms of Pareto optimality (Cortes et al., 2020; Súkeník and Lampert, 2022). However, for general nonconvex smooth MOL problems, it can only be guaranteed for an algorithm to converge to Pareto stationarity of the empirical objective, i.e., a small $\min_{\lambda\in\Delta^M}\|\nabla F_S(x)\lambda\|$ (Désidéri, 2012; Fliege et al., 2019). Thus, it is not reasonable to measure population risk in terms of Pareto optimality in this case. Furthermore, when all the objectives are convex or strongly convex, Pareto stationarity is a sufficient condition for weak Pareto optimality or Pareto optimality, respectively, as stated in Proposition 2.1.

Proposition 2.1. *(Tanabe et al., 2019, Lemma 2.2) If $f_m(x)$ are convex or strongly-convex for all $m \in [M]$, and $x \in \mathbb{R}^d$ is a Pareto stationary point of $F(x)$, then x is weakly Pareto optimal or Pareto optimal.*

Next we proceed to decompose the PS population risk in (2.1).

Error Decomposition. Given a model x, the PS population risk can be decomposed into

$$\underbrace{\min_{\lambda\in\Delta^M}\|\nabla F(x)\lambda\|}_{\text{PS population risk } R_{\text{pop}}(x)} = \underbrace{\min_{\lambda\in\Delta^M}\|\nabla F(x)\lambda\| - \min_{\lambda\in\Delta^M}\|\nabla F_S(x)\lambda\|}_{\text{PS generalization error} R_{\text{gen}}(x)} + \underbrace{\min_{\lambda\in\Delta^M}\|\nabla F_S(x)\lambda\|}_{\text{PS optimization error } R_{\text{opt}}(x)} \tag{2.1}$$

where the optimization error quantifies the training performance, i.e., how well does model x perform on the training data; and the generalization error (gap) quantifies the difference between the testing performance on new data sampled from $\mathcal{D}$ and the training performance, i.e., how well the model x performs on unseen testing data compared to the training data.

Let $A : \mathcal{Z}^n \mapsto \mathbb{R}^d$ denote a randomized MOL algorithm. Given training data S, we are interested in the expected performance of the output model $x = A(S)$, which is measured by $\mathbb{E}_{A,S}[R_{\text{pop}}(A(S))]$. From (2.1) and linearity of expectation, it holds that

$$\mathbb{E}_{A,S}[R_{\text{pop}}(A(S))] = \mathbb{E}_{A,S}[R_{\text{gen}}(A(S))] + \mathbb{E}_{A,S}[R_{\text{opt}}(A(S))]. \tag{2.2}$$

Distance to CA direction. As demonstrated in Figure 1, the key merit of dynamic weighting over static weighting algorithms lies in its ability to navigate through conflicting gradients. Consider an update direction $d = -\nabla F_S(x)\lambda$. To obtain such a steepest CA direction in unconstrained learning, we can reformulate the problem at each iteration (Fliege et al., 2019), with the goal of maximizing the minimum descent (among all objectives) along the update direction d, where the minimum descent given direction d and step size α can be computed by

$$\frac{1}{\alpha}\min_{m\in[M]} f_{S,m}(x) - f_{S,m}(x+\alpha d) \approx \min_{m\in[M]} -\langle \nabla f_{S,m}(x), d\rangle = \min_{\lambda\in\Delta^M} -\langle \nabla F_S(x)\lambda, d\rangle.$$

Since the solutions to λ and d may not necessarily be singletons, we further explicitly regularize the ℓ_2-norm of λ and d so as to put more emphasis on all the objectives instead of focusing on the worst one, and to ensure d does not go to infinity. With the above measurement, the algorithm aims to find an update direction d that maximizes the following

$$\max_{d\in\mathbb{R}^d}\min_{\lambda\in\Delta^M} -\langle \nabla F_S(x)\lambda, d\rangle + \frac{\rho}{2}\|\lambda\|^2 - \frac{1}{2}\|d\|^2 \tag{2.3}$$

Algorithm 1 Regularized MGDA

input Training data S, initial model x_0, and the learning rates $\{\alpha_t\}_{t=0}^T$.
for $t = 0, \ldots, T-1$ **do**
  Compute gradients $\nabla F_S(x_t)$
  Compute CA direction $d(x_t)$ by (2.5)
  Update x_{t+1} via $x_{t+1} = x_t + \alpha_t d(x_t)$
end for
output x_T

Algorithm 2 MoDo - Stochastic MGDA

input Training data S, initial model x_0, initial weight λ_0, and learning rates $\{\alpha_t\}_{t=0}^T$, $\{\gamma_t\}_{t=0}^T$.
for $t = 0, \ldots, T-1$ **do**
  Compute $\nabla F_{z_{t+1,1}}(x_t)$ and $\nabla F_{z_{t+1,2}}(x_t)$
  Update λ_{t+1} by (2.9a)
  Update x_{t+1} by (2.9b)
end for
output x_T

where $\rho \geq 0$ is a regularization constant. We introduce the regularization term with parameter ρ to ensure the optimal dynamic weight does not deviate too much from the uniform weight, with a similar idea as the CAGrad method (Liu et al., 2021a), which could improve the average of the training objectives. By the min-max theorem, this problem can then be reformulated as

$$\max_{\lambda \in \Delta^M} \min_{d \in \mathbb{R}^d} \langle \nabla F_S(x)\lambda, d\rangle - \frac{\rho}{2}\|\lambda\|^2 + \frac{1}{2}\|d\|^2 \tag{2.4}$$

where the optimal solution is $d = -\nabla F_S(x)\lambda_\rho^*(x)$, with $\lambda_\rho^*(x) \in \arg\min_{\lambda \in \Delta^M} \frac{1}{2}\|\nabla F_S(x)\lambda\|^2 + \frac{\rho}{2}\|\lambda\|^2$. Then the CA direction is calculated as

$$\text{CA direction} \qquad d(x) = -\nabla F_S(x)\lambda_\rho^*(x) \;\text{ s.t. }\; \lambda_\rho^*(x) \in \arg\min_{\lambda \in \Delta^M} \|\nabla F_S(x)\lambda\|^2 + \rho\|\lambda\|^2. \tag{2.5}$$

The regularized MGDA adopts $d(x)$ as the update direction at each iteration, as summarized in Algorithm 1. Let $d_\lambda(x) = -\nabla F_Z(x)\lambda$ denote the stochastic update direction with random mini-batch data Z, and $x \in \mathbb{R}^d$, $\lambda \in \Delta^M$ generated by the stochastic algorithm A. We measure the so-termed CA distance via

$$\text{CA direction distance} \qquad \mathcal{E}_{\text{ca}}(x, \lambda) := \|\mathbb{E}_A[d_\lambda(x) - d(x)]\|^2, \tag{2.6}$$

$$\text{CA weight distance} \qquad \mathcal{E}_{\text{caw}}(x, \lambda) := \|\mathbb{E}_A[\lambda - \lambda_\rho^*(x)]\|^2. \tag{2.7}$$

With the above definitions of measures that quantify the performance in different aspects, we then introduce a stochastic gradient algorithm for MOL studied in this work.

2.3 A stochastic algorithm for MOL

MGDA finds $\lambda^*(x)$ in (2.5) using the full-batch gradient $\nabla F_S(x)$, and then constructs $d(x) = -\nabla F_S(x)\lambda^*(x)$, a CA direction for all empirical objectives $f_{S,m}(x)$; see details in Algorithm 1. However, in practical statistical learning settings, the full-batch gradient $\nabla F_S(x)$ may be costly to obtain, and thus one may resort to a stochastic estimate of $\nabla F_S(x)$ instead. The direct stochastic counterpart of MGDA, referred to as the stochastic multi-gradient algorithm in (Liu and Vicente, 2021), replaces the full-batch gradients $\nabla f_{S,m}(x)$ in (2.5) with their stochastic approximations $\nabla f_{z,m}(x)$ for $z \in S$, which, however, introduces a biased stochastic estimate of λ_{t+1}^*, thus a biased CA direction; see Liu and Vicente (2022, Section 4) and Fernando et al. (2023, Section 2.3).

To provide a tight analysis, we introduce a simple yet theoretically grounded stochastic variant of MGDA - stochastic Multi-Objective gradient with DOuble sampling algorithm (MoDo). MoDo obtains an unbiased stochastic estimate of the gradient of problem (2.5) through double (independent) sampling because

$$\mathbb{E}_{z_{t,1},z_{t,2}}[\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t)\lambda_t] = \nabla F_S(x_t)^\top \nabla F_S(x_t)\lambda_t. \tag{2.8}$$

At each iteration t, denote $z_{t,s}$ as an independent sample from S with $s \in [2]$, and $\nabla F_{z_{t,s}}(x_t)$ as a stochastic estimate of $\nabla F_S(x_t)$. MoDo updates x_t and λ_t by

$$\lambda_{t+1} = \Pi_{\Delta^M}\Big(\lambda_t - \gamma_t\big(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I}\big)\lambda_t\Big) \tag{2.9a}$$

$$x_{t+1} = x_t - \alpha_t \nabla F_{Z_{t+1}}(x_t)\lambda_{t+1} \tag{2.9b}$$

where α_t, γ_t are step sizes, $\Pi_{\Delta^M}(\cdot)$ denotes Euclidean projection to the simplex Δ^M, $Z_{t+1} = \{z_{t+1,1}, z_{t+1,2}\}$, and $\nabla F_{Z_{t+1}}(x_t) = \frac{1}{|Z_{t+1}|}\sum_{z\in Z_{t+1}} \nabla F_z(x_t)$. We summarize the MoDo algorithm in Algorithm 2 and will focus on its theoretical analysis subsequently.

3 Optimization, Generalization and Three-Way Trade-Off

This section presents the theoretical analysis of the PS population risk associated with the MoDo algorithm, where the analysis of generalization error is in Section 3.1 and that of optimization error is in Section 3.2. A summary of our main results is given in Table 1.

3.1 Multi-objective generalization and uniform stability

We first bound the expected PS generalization error by the generalization in gradients in Proposition 3.1, then introduce the MOL uniform stability and establish its connection to the generalization in gradients. Finally, we bound the MOL uniform stability.

Proposition 3.1. *With $\|\cdot\|_{\mathrm{F}}$ denoting the Frobenious norm, the PS generalization error $R_{\mathrm{gen}}(A(S))$ in (2.2) is bounded by*

$$\mathbb{E}_{A,S}[R_{\mathrm{gen}}(A(S))] \le \mathbb{E}_{A,S}[\|\nabla F(A(S)) - \nabla F_S(A(S))\|_{\mathrm{F}}]. \tag{3.1}$$

With Proposition 3.1, next, we introduce the concept of MOL uniform stability tailored to MOL problems and show that PS generalization error can be bounded by the MOL uniform stability. Then we analyze their bound in general NC case and SC case, respectively. Note that, the stability in the general NC case cover the convex case. However, it is worse than the stability in single-objective learning in the convex case as shown by (Hardt et al., 2016). This is primarily due to the fact that the non-expansiveness property of the x_t update, commonly observed in single-objective learning with convex objectives, is no longer valid for MOL when employing dynamic weighting algorithms.

Definition 3.1 (MOL uniform stability). *A randomized algorithm $A : \mathcal{Z}^n \mapsto \mathbb{R}^d$, is MOL-uniformly stable with ϵ_{F} if for all neighboring datasets S, S' that differ in at most one sample, we have*

$$\sup_z \ \mathbb{E}_A\big[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2\big] = \epsilon_{\mathrm{F}}^2. \tag{3.2}$$

Table 1: Comparison of optimization error, generalization error, and population risk under different assumptions for static and dynamic weighting. "NC", "SC" represent nonconvex and strongly convex, and "Lip-C", "S" represent Lipschitz continuous and smooth, respectively. Non-dominant terms are omitted. The results are given in $\mathcal{O}(\cdot)$ if not otherwise specified.

Assmp	Method	Optimization	Generalization	Risk	CA weight distance
NC, Lip-C,S	Static	$(\alpha T)^{-\frac{1}{2}} + \alpha^{\frac{1}{2}}$	$T^{\frac{1}{2}} n^{-\frac{1}{2}}$	$n^{-\frac{1}{6}}$	$\Theta(1)$
	Dynamic	$(\alpha T)^{-\frac{1}{2}} + \alpha^{\frac{1}{2}} + \gamma^{\frac{1}{2}}$	$T^{\frac{1}{2}} n^{-\frac{1}{2}}$	$n^{-\frac{1}{6}}$	$\gamma\rho^{-1} + \alpha\gamma^{-1}\rho^{-2}$
SC, S	Static	$(1-\alpha)^{\frac{T}{2}} + \alpha^{\frac{1}{2}}$	$n^{-\frac{1}{2}}$	$n^{-\frac{1}{2}}$	$\Theta(1)$
	Dynamic	$\alpha^{\frac{1}{2}} + \begin{cases} (1-\alpha)^{\frac{T}{2}} + \gamma T, \ \gamma = \mathcal{O}(T^{-2}), \\ (\alpha T)^{-\frac{1}{2}} + \gamma^{\frac{1}{2}} + \rho^{\frac{1}{2}}, \ \text{o.w.} \end{cases}$	$\begin{cases} n^{-\frac{1}{2}}, \ \gamma = \mathcal{O}(T^{-1}) \\ T^{\frac{1}{2}} n^{-\frac{1}{2}}, \ \text{o.w.} \end{cases}$	$\begin{cases} n^{-\frac{1}{2}} \\ n^{-\frac{1}{6}} \end{cases}$	$\gamma\rho^{-1} + \alpha\gamma^{-1}\rho^{-2}$

It is worth noting that the MOL uniform stability (3.2) holds true uniformly for all neighboring datasets S and S', meaning that ϵ_{F} is independent of the choice of S and S'. Next, we show the relation between the upper bound of PS generalization error in (3.1) and MOL uniform stability in (3.2).

Proposition 3.2 (MOL uniform stability and generalization). *Assume for any z, the function $F_z(x)$ is differentiable. If a randomized MOL algorithm $A : \mathcal{Z}^n \mapsto \mathbb{R}^d$ is MOL-uniformly stable with ϵ_{F}, then*

$$\mathbb{E}_{A,S}[\|\nabla F(A(S)) - \nabla F_S(A(S))\|_{\mathrm{F}}] \leq 4\epsilon_{\mathrm{F}} + \sqrt{n^{-1}\mathbb{E}\left[\mathbb{V}_{z\sim\mathcal{D}}(\nabla F_z(A(S)))\right]} \tag{3.3}$$

where the variance is defined as $\mathbb{V}_{z\sim\mathcal{D}}(\nabla F_z(A(S))) = \mathbb{E}_{z\sim\mathcal{D}}\big[\|\nabla F_z(A(S)) - \mathbb{E}_{z\sim\mathcal{D}}[\nabla F_z(A(S))]\|_{\mathrm{F}}^2\big]$.

Proposition 3.2 establishes the connection between the upper bound of the PS generalization error and the MOL uniform stability, where the former can be bounded above by the latter plus the variance of the stochastic gradient over the population data distribution. It is worth noting that the standard arguments of bounding the generalization error measured in function values by the uniform stability measured in function values (Hardt et al., 2016, Theorem 2.2) is not applicable here as the summation and norm operators are not exchangeable. More explanations are provided in the proof in Appendix A.1.

Theorem 3.1 (PS generalization error of MoDo in the NC case). *Let A be the MoDo algorithm. If $\sup_z \mathbb{E}_A\left[\|\nabla F_z(A(S))\|_{\mathrm{F}}^2\right] \leq G^2$ for any S, then the MOL uniform stability ϵ_{F}^2 in Definition 3.1 for MoDo algorithm with step sizes $\alpha = \mathcal{O}(1), \gamma = \mathcal{O}(1)$ satisfies*

a) the MOL uniform stability ϵ_{F}^2 for $A_t(S)$ with $t \in [T]$ is upper bounded by $\epsilon_{\mathrm{F}}^2 = \mathcal{O}(Tn^{-1})$;

b) there exist functions $F_z(x)$, neighboring datasets S, S', and $t \in [T]$ such that the MOL uniform stability ϵ_{F} for $A_t(S)$ is lower bounded by $\epsilon_{\mathrm{F}}^2 = \Omega(Tn^{-1})$.

And the PS generalization error at iteration $t \in [T]$ is $\mathbb{E}_{A,S}[R_{\mathrm{gen}}(A_t(S))] = \mathcal{O}(T^{\frac{1}{2}} n^{-\frac{1}{2}})$.

Proof of Theorem 3.1 is provided in Appendix A.2. Compared to the function value uniform stability upper bound in (Hardt et al., 2016, Theorem 3.12) for nonconvex single-objective learning, Theorem 3.1 does not require a step size decay $\alpha_t = \mathcal{O}(1/t)$, thus can enjoy at least a polynomial convergence rate of optimization errors w.r.t. T. The tightness

of the stability upper bound is verified by providing a matching lower bound of the stability. Combining Theorem 3.1 with Proposition 3.2, to ensure the generalization error is diminishing with n, one needs to choose $T = o(n)$, which lies in the "early stopping" regime and results in potentially large optimization error. While the "early stopping" phenomenon has been indeed observed in practice for general nonconvex settings, we next provide a tighter bound in the strongly convex (SC) case that allows a larger choice of T. Below we list the standard assumptions used to derive the MOL uniform stability in the SC case.

Assumption 1 (Smoothness). *For all $m \in [M]$, $z \in \mathcal{Z}$, $\nabla f_{z,m}(x)$ is $\ell_{f,1}$-Lipschitz continuous, i.e., $\|\nabla f_{z,m}(x) - \nabla f_{z,m}(x')\| \leq \ell_{f,1}\|x - x'\|$. Then $\nabla F_z(x)$ is $\ell_{F,1}$-Lipschitz continuous in Frobenius norm with $\ell_{F,1} = \sqrt{M}\ell_{f,1}$, i.e., $\|\nabla F_z(x) - \nabla F_z(x')\|_{\mathrm{F}} \leq \ell_{F,1}\|x - x'\|$.*

Assumption 2 (Strong convexity). *For all $m \in [M]$, $z \in \mathcal{Z}$, $f_{z,m}(x)$ is μ-strongly convex w.r.t. x, i.e., $f_{z,m}(x') - f_{z,m}(x) \geq \nabla f_{z,m}(x)^\top (x' - x) + \frac{\mu}{2}\|x' - x\|^2$, where $\mu > 0$.*

Note that Assumptions 1 and 2 are only used to derive the MOL uniform stability in the SC case but not in the general NC case. In the SC case, the gradient norm $\|\nabla F_z(x)\|_{\mathrm{F}}$ is generally unbounded in $\mathbb{R}^d$. Therefore, one cannot assume Lipschitz continuity of $f_{z,m}(x)$. We address this challenge by showing that $\{x_t\}_{t=1}^T$ generated by the MoDo algorithm are bounded as stated in Lemma 3.1. Combining with Assumption 1, the gradient norm $\|\nabla F_z(x_t)\|_{\mathrm{F}}$ is also bounded, which serves as a stepping stone to derive the MOL stability bound.

Lemma 3.1 (x_t bounded for SC and smooth objectives). *Suppose Assumptions 1 and 2 hold. For $\{x_t\}, t \in [T]$ generated by the dynamic weighting algorithms such as MoDo and SMG with weight $\lambda \in \Delta^M$, step size $\alpha_t = \alpha$, and $0 \leq \alpha \leq \ell_{f,1}^{-1}$, then*

a) there exists a finite positive constant c_x independent of T such that $\|x_t\| \leq c_x$;

b) there exist finite positive constants ℓ_f, $\ell_F = \sqrt{M}\ell_f$ as functions of c_x, such that for all $\lambda \in \Delta^M$, we have $\|\nabla F(x_t)\lambda\| \leq \ell_f$, and $\|\nabla F(x_t)\|_{\mathrm{F}} \leq \ell_F$.

Proof of Lemma 3.1 is deferred to Appendix A.4. With Lemma 3.1, the MOL uniform stability and the PS generalization error of MoDo are provided below.

Theorem 3.2 (PS generalization error of MoDo in SC case). *Suppose Assumptions 1 and 2 hold. The MOL uniform stability ϵ_{F} in Definition 3.1 for MoDo algorithm with $0 \leq \alpha_t = \alpha \leq \ell_{f,1}^{-1}$, $\gamma_t = \gamma = \mathcal{O}(T^{-1})$ satisfies*

a) ϵ_{F}^2 for $A_t(S)$ with $t \in [T]$ is upper bounded by

$$\epsilon_{\mathrm{F}}^2 = \mathcal{O}\big(Mn^{-1}(\alpha + M\gamma + Mn^{-1})\big); \tag{3.4}$$

b) there exist functions $F_z(x)$ that satisfy Assumptions 1 and 2, neighboring datasets S, S', and $t \in [T]$, $n \in \mathbb{N}$ such that ϵ_{F}^2 for $A_t(S)$ is lower bounded by $\epsilon_{\mathrm{F}}^2 = \Omega(Mn^{-2})$.

The PS generalization error at iteration $t \in [T]$ is $\mathbb{E}_{A,S}[R_{\mathrm{gen}}(A_t(S))] = \mathcal{O}(n^{-\frac{1}{2}})$.

See the proof of Theorem 3.2 in Appendix A.5. The idea of the proof is summarized as follows. 1) To bound the MOL uniform stability is to bound the expected gradient difference evaluated on the output model parameters generated by the algorithm given two

neighboring training datasets S and S'. By the smoothness assumption of the objectives, this can be bounded by the difference of the model parameters generated by the algorithm given the neighboring datasets, i.e., the argument stability. 2) Let $\{x_t\}$, $\{x'_t\}$ be the sequences generated by the MoDo algorithm using S and S', respectively. By the properties of the MoDo update in Section A.5.1, both x_t and λ_t, $x_{t+1} - x'_{t+1}$ can be bounded in terms of $x_t - x'_t$ and $\lambda_t - \lambda'_t$ based on the growth recursion in Lemma A.7. 3) Applying the growth recursion from $t = 0, \ldots, T$, the argument stability $\mathbb{E}_A[\|A(S) - A(S')\|] = \mathbb{E}_A[\|x_T - x'_T\|]$ can be bounded through mathematical induction as detailed in Section A.5.2.

Theorem 3.2 provides both the upper and lower bounds for the MOL uniform stability of MoDo in the step sizes α, γ, and training data size n. Below we provide a remark on how these parameters affect the stability.

Remark 1. *If we choose $\alpha = \Theta(T^{-\frac{1}{2}})$, $\gamma = \mathcal{O}(T^{-1})$, and $T = \Theta(n^2)$, the upper bound of the MOL uniform stability matches its the lower bound in an order of n^{-2}, suggesting that our bound is tight. From Propositions 3.1 and 3.2, we know that the generalization error bound is a direct implication from the MOL uniform stability bound in* (3.4)*. It states that the PS generalization error of MoDo is $\mathcal{O}(n^{-\frac{1}{2}})$, which matches the generalization error of static weighting up to a constant (Lei, 2023). Our result also indicates that when all the objectives are strongly convex, choosing small step sizes α and γ leads to a smaller MOL uniform stability and thus can benefit the generalization error.*

3.2 Multi-objective CA distance and optimization error

In this section, we bound the multi-objective PS optimization error, i.e., $\min_{\lambda \in \Delta^M} \|\nabla F_S(x)\lambda\|$, which has been the main metric in the recent MOL optimization literature such as (Fliege et al., 2019; Désidéri, 2012). As discussed in Section 2.2, this measure being zero implies the model x achieves a Pareto stationarity for the empirical problem.

Below we list an additional standard assumption used to derive the theoretical results.

Assumption 3 (Lipschitz $F_z(x)$). *For all $m \in [M]$, $f_{z,m}(x)$ are ℓ_f-Lipschitz continuous for all z, then $F_z(x)$ are ℓ_F-Lipschitz continuous in Frobenius norm with $\ell_F = \sqrt{M}\ell_f$.*

Note that the constants ℓ_f, ℓ_F used in Lemma 3.1 are derived from Assumptions 1 and 2, depending on $\mu, \ell_{f,1}$, and are different from those in Assumption 3.

We first introduce the theoretical results on the CA direction and CA weight distances, given in Theorems 3.3 and 3.4.

Theorem 3.3 (CA direction distance of MoDo). *Suppose either: 1) Assumptions 1 and 3 hold; or, 2) Assumptions 1 and 2 hold, with ℓ_f and ℓ_F defined in Lemma 3.1. For $\{x_t\}, \{\lambda_t\}$ generated by MoDo with step sizes $\alpha_t = \alpha > 0$, $\gamma_t = \gamma > 0$, and regularization $\rho \geq 0$, given training data S, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathcal{E}_{\text{ca}}(x_t, \lambda_{t+1}) = \mathcal{O}\big(\gamma^{-1}T^{-1} + M^{\frac{1}{2}}\alpha^{\frac{1}{2}}\gamma^{-\frac{1}{2}} + \gamma M + \rho\gamma^{-1} + \rho\big). \tag{3.5}$$

Theorem 3.3 establishes the convergence to the CA direction using the measure introduced in Section 2.2, with proof provided in Appendix B.2. For example, one can choose $\alpha = \Theta(T^{-\frac{3}{4}})$, $\gamma = \Theta(T^{-\frac{1}{4}})$, and $\rho = \mathcal{O}(T^{-\frac{1}{2}})$, then the right hand side of (3.5) is $\mathcal{O}(T^{-\frac{1}{4}})$.

Below we provide a guarantee on the distance to the CA weight when the regularization is strictly enforced, i.e., $\rho > 0$. Note that the weight for static weighting method is predefined and does not involve regularization. Here we derive a lower bound of the CA weight distance in static weighting, which is larger than the CA weight distance of MoDo.

Proposition 3.3 (CA weight distance of static weighting). *Suppose Assumption 1 holds. Then there exists $\lambda \in \Delta^M$ for static weighting such that*

$$\mathcal{E}_{\text{caw}}(x_T, \lambda) = \Theta(1). \tag{3.6}$$

Proof. First, for the upper bound, since both $\lambda, \lambda_\rho^*(x_T) \in \Delta^M$, it holds that

$$\mathcal{E}_{\text{caw}}(x_T, \lambda) \leq \mathbb{E}_A[\|\lambda - \lambda_\rho^*(x_T)\|^2] \leq 4. \tag{3.7}$$

Second, for the lower bound, for a given $\lambda_\rho^*(x_T) \in \Delta^M$ with $M \geq 2$, let $\lambda_{\rho,m}^*(x_T)$ denote the m-th element of $\lambda_\rho^*(x_T)$ with $m \in [M]$. Define $m^* \coloneqq \arg\min_{m\in[M]} \mathbb{E}_A[\lambda_{\rho,m}^*(x_T)]$. Then $\mathbb{E}_A[\lambda_{\rho,m^*}^*(x_T)] \leq \frac{1}{M} \leq \frac{1}{2}$. Then there exists $\lambda \in \Delta^M$ with $\lambda_{m^*} = 1$ such that

$$\mathcal{E}_{\text{caw}}(x_T, \lambda) = \|\lambda - \mathbb{E}_A[\lambda_\rho^*(x_T)]\|^2 \geq \Big(1 - \frac{1}{M}\Big)^2 \geq \frac{1}{4}. \tag{3.8}$$

Combining the upper and lower bounds yields the result. ■

Theorem 3.4 (CA weight distance of MoDo). *Suppose either: 1) Assumptions 1 and 3 hold; or, 2) Assumptions 1 and 2 hold, with ℓ_f and ℓ_F defined in Lemma 3.1. For $\{x_t\}, \{\lambda_t\}$ generated by MoDo with step sizes $\alpha_t = \alpha > 0$, $\gamma_t = \gamma > 0$, and regularization $\rho > 0$, given training data S, it holds that*

$$\mathcal{E}_{\text{caw}}(x_T, \lambda_{T+1}) = \mathcal{O}\big((1-\rho\gamma)^T + \rho^{-1}\gamma(M^{\frac{1}{2}} + \rho)^2 + \rho^{-2}\gamma^{-1}\alpha M\big). \tag{3.9}$$

Proof of Theorem 3.4 is provided in Appendix B.3. Below we provide a remark on the difference between the two measures: CA direction distance, and CA weight distance.

Remark 2. *Compared to the convergence to CA direction, the convergence to CA weight is stronger because it involves last-iterate point convergence instead of average-iterate function value convergence, and can only be guaranteed with $\rho = \omega(\max\{\gamma, \alpha^{\frac{1}{2}}\gamma^{-\frac{1}{2}}, \gamma^{-1}T^{-1}\})$, resulting in a trade-off in convergence to PS stationarity and convergence to the CA weight. Using the distance to the CA weight as a measure, the lower bound of the static weighting method in this measure can be derived, which is a constant, strictly greater than the upper bound of MoDo, further justifying the benefit of MoDo over static weighting in the CA weight distance.*

Below, we state the estimation for the PS optimization error of MoDo in the following theorem.

Theorem 3.5 (PS optimization error of MoDo). *Given training data S, define c_F such that $\mathbb{E}_A[F_S(x_0)\lambda_0] - \min_{x\in\mathbb{R}^d} \mathbb{E}_A[F_S(x)\lambda_0] \le c_F$. Considering $\{x_t\}$ generated by MoDo (Algorithm 2), with $0 < \alpha_t = \alpha \le 1/(2\ell_{f,1})$, $\gamma_t = \gamma \ge 0$. Suppose*

1) Assumptions 1 and 3 hold (NC case), then

$$\mathbb{E}_A\Big[\min_{t\in[T]} R_{\rm opt}(x_t)\Big] = \mathcal{O}\big(\alpha^{-\frac{1}{2}}T^{-\frac{1}{2}} + \gamma^{\frac{1}{2}}M^{\frac{1}{2}} + \alpha^{\frac{1}{2}} + \rho^{\frac{1}{2}}\big); \tag{3.10}$$

2) Assumptions 1, 2 hold (SC case), with ℓ_f defined in Lemma 3.1, then

$$\mathbb{E}_A\Big[\min_{t\in[T]} R_{\rm opt}(x_t)\Big] = \mathcal{O}\big(\min\{\alpha^{-\frac{1}{2}}T^{-\frac{1}{2}}+\gamma^{\frac{1}{2}}M^{\frac{1}{2}}+\alpha^{\frac{1}{2}}+\rho^{\frac{1}{2}}, (1-\alpha)^{\frac{T}{2}}+\alpha^{\frac{1}{2}}+M^{\frac{1}{2}}\gamma T\}\big). \tag{3.11}$$

Proof of Theorem 3.5 is provided in Appendix C.1. The idea of the proof is summarized as follows. 1) At every step t of the algorithm, the weighted descent amount of the loss function $F(x_{t+1})\lambda - F(x_t)\lambda$ given any weight $\lambda \in \Delta^M$ depends on the drift of the dynamic weight $\langle \lambda_t - \lambda, (\nabla F_S(x_t)^\top \nabla F_S(x_t) + \rho \mathrm{I})\lambda_t\rangle$. 2) This drift can be bounded through Lemma B.5, (B.7) according to the update rule of the dynamic weight, given by (2.9a). 3) Combining the inequalities in the previous steps and taking the telescoping sum from $t = 0, \ldots, T$ yield the bound of the optimization error. Below we provide a remark on Theorem 3.5 under different choices of step sizes.

Remark 3. *Note that the original result with the squared PS optimization error in the general nonconvex case is in the form of*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_A\Big[\min_{\lambda\in\Delta^M} \|\nabla F_S(x_t)\lambda\|^2\Big] = \mathcal{O}\big(\alpha^{-1}T^{-1} + \alpha + \gamma + \rho\big). \tag{3.12}$$

And it holds for any choice of α, γ, T as long as $0 < \alpha \le 1/(2\ell_{f,1})$. Based on this result, one optimal choice to ensure the best $\mathcal{O}(T^{-\frac{1}{2}})$ convergence rate of the optimization error in square is $\alpha = \Theta(T^{-\frac{1}{2}})$, $\gamma = \Theta(T^{-\frac{1}{2}})$, $\rho = \mathcal{O}(T^{-\frac{1}{2}})$. However, this results in a constant error bound of the CA distance according to Theorem 3.3. To ensure better convergence to CA direction, one possible choice, $\alpha = \Theta(T^{-\frac{3}{4}})$, $\gamma = \Theta(T^{-\frac{1}{4}})$, and $\rho = \mathcal{O}(T^{-\frac{1}{2}})$, is suboptimal with regard to the convergence to Pareto stationarity, as evidenced by Theorem 3.5. This exhibits a trade-off between convergence to the CA direction and convergence to Pareto stationarity. To ensure a faster convergence rate in the SC case without requiring convergence of CA distance, one can also choose $\gamma = \mathcal{O}(T^{-2})$, $\alpha = \Theta(T^{-1}\ln T)$, then the convergence rate of the optimization error in square is $\mathcal{O}(T^{-1}\ln T)$.

3.3 Optimization, generalization and conflict avoidance trade-off

Combining the results in Sections 3.1 and 3.2, we are ready to analyze and summarize the three-way trade-off of MoDo. With $A_t(S) = x_t$ denoting the output of algorithm A at the t-th iteration, we can decompose the PS population risk $R_{\rm pop}(A_t(S))$ as (cf. (2.1) and (3.1))

$$\mathbb{E}_{A,S}\big[R_{\rm pop}(A_t(S))\big] \le \mathbb{E}_{A,S}\Big[\min_{\lambda\in\Delta^M} \|\nabla F_S(A_t(S))\lambda\|\Big] + \mathbb{E}_{A,S}\Big[\|\nabla F(A_t(S)) - \nabla F_S(A_t(S))\|_{\rm F}\Big].$$

The general nonconvex case. Suppose Assumptions 1 and 3 hold. By the generalization error bound in Theorem 3.1, and the optimization error bound in Theorem 3.5, denote $\hat{t} \in \arg\min_{t\in[T]} R_{\text{opt}}(x_t)$, the PS population risk of the output of MoDo can be bounded by

$$\mathbb{E}_{A,S}\left[R_{\text{pop}}(A_{\hat{t}}(S))\right] = \mathcal{O}\left(\alpha^{-\frac{1}{2}}T^{-\frac{1}{2}} + \alpha^{\frac{1}{2}} + \gamma^{\frac{1}{2}} + T^{\frac{1}{2}}n^{-\frac{1}{2}}\right). \tag{3.13}$$

Discussion of trade-off. Choosing step sizes $\alpha = \Theta(T^{-\frac{1}{2}})$, $\gamma = \Theta(T^{-\frac{1}{2}})$, and number of steps $T = \Theta(n^{\frac{2}{3}})$, then the expected PS population risk is $\mathcal{O}(n^{-\frac{1}{6}})$, which matches the PS population risk upper bound of a general nonconvex single objective in (Lei, 2023). A clear trade-off in this case is between the optimization error and generalization error, controlled by T. Indeed, increasing T leads to smaller optimization errors but larger generalization errors, and vice versa. To satisfy convergence to CA direction, it requires $\gamma = \omega(\alpha)$, i.e., $\frac{\alpha}{\gamma} = o(1)$, based on Theorem 3.3, and the optimization error in turn becomes worse, so does the PS population risk. Specifically, choosing $\alpha = \Theta(T^{-\frac{1}{2}})$, $\gamma = \Theta(T^{-\frac{1}{4}})$, and $T = \Theta(n^{\frac{4}{5}})$ leads to the expected PS population risk in $\mathcal{O}(n^{-\frac{1}{10}})$, and the distance to CA direction in $\mathcal{O}(n^{-\frac{1}{10}})$. This shows another trade-off between conflict avoidance and optimization error.

The strongly convex case. Suppose Assumptions 1 and 2 hold. By the generalization error and the optimization error given in Theorems 3.2 and 3.5, the PS population risk of MoDo can be bounded by

$$\mathbb{E}_{A,S}\left[R_{\text{pop}}(A_{\hat{t}}(S))\right] = \mathcal{O}\left(\alpha^{-\frac{1}{2}}T^{-\frac{1}{2}} + \alpha^{\frac{1}{2}} + \gamma^{\frac{1}{2}} + n^{-\frac{1}{2}}\right). \tag{3.14}$$

Discussion of trade-off. Choosing $\alpha = \Theta(T^{-\frac{1}{2}})$, $\gamma = \Theta(T^{-1})$, and $T = \Theta(n^2)$, the expected PS population risk in gradients is $\mathcal{O}(n^{-\frac{1}{2}})$. However, choosing $\gamma = \Theta(T^{-1})$ leads to large distance to the CA direction according to Theorem 3.3 because the term $4/(\gamma T)$ in (3.5) increases with T. To ensure convergence to the CA direction, it requires $\gamma = \omega(T^{-1})$, under which the tighter bound in Theorem 3.2 does not hold but the bound in Theorem 3.1 still holds. In this case, the PS population risk under the proper choice of α, γ, and T is $\mathcal{O}(n^{-\frac{1}{6}})$ as discussed in the previous paragraph. Therefore, to avoid conflict among gradients, MoDo needs to sacrifice the sample complexity of PS population risk, demonstrating a trade-off between conflict avoidance and PS population risk, as illustrated in Figure 2.

4 Application to Other MOL Algorithms

Our theoretical framework is general and can be applied to other stochastic MOL algorithms to analyze these three errors. To demonstrate this, we apply our framework to analyze other stochastic MOL algorithms including SMG (Liu and Vicente, 2021) and MoCo (Fernando et al., 2023) in this section. Both SMG and MoCo mitigate the bias in the CA direction during optimization. To achieve this, SMG (Liu and Vicente, 2021) increases the batch size during optimization, MoCo (Fernando et al., 2023) uses momentum-based methods. We then describe in detail the updates of SMG and MoCo.

SMG substitutes the full-batch gradient $\nabla F_S(x_t)$ used in MGDA in (2.5) with its stochastic estimate $\nabla F_{Z_t}(x_t)$, where Z_t is a randomly sampled mini-batch of data at the t-th iteration, and its batch size $|Z_t|$ is increasing with the iteration t to mitigate the CA direction bias. The SMG algorithm is summarized in Algorithm 3.

Algorithm 3 SMG (Liu and Vicente, 2021)

Input: Initial model x_0, the learning rates $\{\alpha_t\}_{t=0}^T$, and the regularization $\rho = 0$.
for $t = 0, \ldots, T-1$ **do**
  Compute $\nabla F_{Z_t}(x_t)$ with $|Z_t| = \mathcal{O}(t)$
  Compute CA direction at x_t with (2.5) by replacing $\nabla F_S(x_t)$ with $\nabla F_{Z_t}(x_t)$;
  Update x_{t+1} via $x_{t+1} = x_t + \alpha_t d(x_t)$;
end for
output x_T

Algorithm 4 MoCo (Fernando et al., 2023)

Input: Initial model x_0, the learning rates $\{\alpha_t\}_{t=0}^T$, and the regularization $\rho = 0$.
Set $Y_0 = \nabla F_{z_0}(x_t)$;
for $t = 0, \ldots, T-1$ **do**
  Sample gradients $\nabla F_{z_{t+1}}(x_t)$;
  Update Y_{t+1} by (4.1a);
  Update λ_{t+1} by (4.1b)
  Update x_{t+1} by (4.1c)
end for
output x_T

Different from SMG, MoCo (Fernando et al., 2023) adopts a momentum-based method to mitigate the CA direction bias. It uses the moving average of the stochastic gradient to compute the CA direction. The MoCo algorithm uses the update functions given by

$$Y_{t+1} = Y_t - \beta_t(Y_t - \nabla F_{z_{t+1}}(x_t)) \tag{4.1a}$$

$$\lambda_{t+1} \in \arg\min_{\lambda \in \Delta^M} \|Y_t \lambda\|^2 \tag{4.1b}$$

$$x_{t+1} = x_t - \alpha_t Y_t \lambda_t \tag{4.1c}$$

where Y_t is the moving average of the stochastic gradients, and $Y_t\lambda_t$ is the estimated CA direction. The MoCo algorithm is summarized in Algorithm 4.

Next, we proceed to formally introduce our theoretical results on three errors for SMG and MoCo algorithms in the general nonconvex case.

4.1 Multi-objective generalization

We summarize the PS generalization error bounds of SMG (Algorithm 3) and MoCo (Algorithm 4) in the general nonconvex case in Theorem 4.1. The proof of this theorem follows similar steps as the proof for the PS generalization error of MoDo in Theorem 3.1, by first deriving their MOL uniform stability bounds, and applying Proposition 3.2 to connect their PS generalization errors with their MOL uniform stability bounds.

Theorem 4.1 (Generalization errors of SMG and MoCo). *Let A be the SMG algorithm with batch size $\mathcal{O}(t)$ at the t-th iteration, or the MoCo algorithm. If $\mathbb{E}_A\big[\|\nabla F_z(A(S))\|_{\mathrm{F}}^2\big] \leq G^2$ for any z and S, then the PS generalization errors of SMG and MoCo satisfy*

$$\textit{(SMG)}\ \ \mathbb{E}_{A,S}[R_{\mathrm{gen}}(A(S))] = \mathcal{O}(Tn^{-\frac{1}{2}}); \quad \textit{(MoCo)}\ \ \mathbb{E}_{A,S}[R_{\mathrm{gen}}(A(S))] = \mathcal{O}(T^{\frac{1}{2}}n^{-\frac{1}{2}}). \tag{4.2}$$

The proof is deferred to Appendix A.3. Theorem 4.1 indicates that the PS generalization error of MoCo in the general nonconvex case are in the same rates as MoDo, with $\mathcal{O}(T^{\frac{1}{2}}n^{-\frac{1}{2}})$, while the generation error of SMG is worse with increasing batch sizes. The result further demonstrates the generality of our proposed theoretical framework to analyze the MOL uniform stability and PS generalization errors of various stochastic MOL algorithms.

Table 2: Comparison with prior stochastic MOL algorithms in terms of assumptions and the guarantees of the three errors, where logarithmic dependence is omitted, and Opt., CA dist., and Gen. are short for optimization error, CA distance, and generalization error, respectively. For MoDo, applying Thms 3.3 and 3.5 with $\alpha = \Theta(T^{-\frac{1}{2}}), \gamma = \Theta(T^{-\frac{1}{2}})$ yields the optimization error and the CA distance in the second last row; setting $\alpha = \Theta(T^{-\frac{3}{4}})$, $\gamma = \Theta(T^{-\frac{1}{4}})$, $\rho = \mathcal{O}(T^{-\frac{1}{2}})$ yields the optimization error and the CA distance in the last row.

Algorithm	Batch size	NC	Lipschitz $\lambda^*(x)$	Bounded function	Opt.	CA dist.	Gen.
SMG (Liu and Vicente, 2021, Thm 5.3)	$\mathcal{O}(t)$	✗	✓	✗	$T^{-\frac{1}{8}}$	-	-
CR-MOGM (Zhou et al., 2022, Thm 3)	$\mathcal{O}(1)$	✓	✗	✓	$T^{-\frac{1}{4}}$	-	-
MoCo (Fernando et al., 2023, Thm 2)	$\mathcal{O}(1)$	✓	✗	✗	$T^{-\frac{1}{20}}$	$T^{-\frac{1}{5}}$	-
MoCo (Fernando et al., 2023, Thm 4)	$\mathcal{O}(1)$	✓	✗	✓	$T^{-\frac{1}{4}}$	$\mathcal{O}(1)$	-
SMG (Ours, Thms 4.1-4.3)	$\mathcal{O}(t)$	✓	✗	✗	$T^{-\frac{1}{8}}$	$T^{-\frac{1}{2}}$	$Tn^{-\frac{1}{2}}$
MoCo (Ours, Thms 4.1-4.3)	$\mathcal{O}(1)$	✓	✗	✗	$T^{-\frac{1}{16}}$	$T^{-\frac{1}{4}}$	$T^{\frac{1}{2}}n^{-\frac{1}{2}}$
MoDo (Ours, Thms 3.1,3.3,3.5)	$\mathcal{O}(1)$	✓	✗	✗	$T^{-\frac{1}{4}}$	$\mathcal{O}(1)$	$T^{\frac{1}{2}}n^{-\frac{1}{2}}$
MoDo (Ours, Thms 3.1,3.3,3.5)	$\mathcal{O}(1)$	✓	✗	✗	$T^{-\frac{1}{8}}$	$T^{-\frac{1}{4}}$	$T^{\frac{1}{2}}n^{-\frac{1}{2}}$

Next, we show how to apply our theoretical framework to analyze the CA distances and PS optimization errors of the stochastic MOL algorithms, SMG and MoCo.

4.2 Multi-objective CA distance and optimization error

Notably, in the CA distance and optimization error analysis, we have also developed new techniques to relax the assumptions and/or improve the final convergence rates of different algorithms; see a detailed comparison in Table 2. To obtain the improved analysis, one critical property is that the CA direction is unique and Hölder continuous (cf. Lemmas B.1 and 4.1), despite that $\lambda^*(x)$ is not Lipschitz continuous in general.

For simplicity, we use $Q \in \mathbb{R}^{d\times M}$ to denote either full-batch gradient matrix $\nabla F_S(x)$ or its stochastic estimate. Then the subproblem without regularization, i.e., $\rho = 0$, is

$$\min_{\lambda\in\Delta^M} \|Q\lambda\|^2 \tag{4.3}$$

which is a constrained quadratic programming problem. The estimate of the CA direction used in either SMG with $Q = \nabla F_{Z_t}(x_t)$, or MoCo with $Q = Y_t$, can be computed by $d_Q = Q\lambda_Q^*$, with $\lambda_Q^* \in \arg\min_{\lambda\in\Delta^M} \|Q\lambda\|^2$.

We then proceed to prove the Hölder continuity of d_Q w.r.t. Q in Lemma 4.1, which is essential for deriving the CA direction distance and PS optimization error of SMG and MoCo. This result also generalizes to constrained quadratic programming problems with general compact and convex set constraints.

Lemma 4.1 (Hölder continuity of d_Q w.r.t. Q). *For all $Q, Q' \in \mathbb{R}^{d\times M}$, define $\lambda^* \in \arg\min_{\lambda\in\Delta^M} \|Q\lambda\|^2$, and $\lambda^{*\prime} \in \arg\min_{\lambda\in\Delta^M} \|Q'\lambda\|^2$, and $d_Q = Q\lambda^*$, $d_{Q'} = Q'\lambda^{*\prime}$, then d_Q and $d_{Q'}$ are both unique and satisfy*

$$\|d_Q - d_{Q'}\|^2 \le 4\max\left\{\sup_{\lambda\in\Delta^M}\|Q\lambda\|, \sup_{\lambda\in\Delta^M}\|Q'\lambda\|\right\}\cdot \sup_{\lambda\in\Delta^M}\|(Q-Q')\lambda\|. \tag{4.4}$$

Proof. The uniqueness of $d_{Q,\rho}$ with $\rho \geq 0$ are given in Lemma B.1, which covers the uniqueness of d_Q ($\rho = 0$), we can then rewrite

$$\begin{aligned}\|d_Q - d_{Q'}\|^2 =&\|Q\lambda^* - Q'\lambda^{*\prime}\|^2 = \|Q\lambda^*\|^2 + \|Q'\lambda^{*\prime}\|^2 - 2\langle Q\lambda^*, Q'\lambda^{*\prime}\rangle \\ =&\|Q\lambda^*\|^2 - \|Q'\lambda^{*\prime}\|^2 + 2\langle Q'\lambda^{*\prime}, Q'\lambda^{*\prime} - Q\lambda^*\rangle \\ =&\|Q\lambda^*\|^2 - \|Q'\lambda^{*\prime}\|^2 + \underbrace{2\langle Q'\lambda^{*\prime}, Q'\lambda^{*\prime} - Q'\lambda^*\rangle}_{\leq 0} + 2\langle Q'\lambda^{*\prime}, Q'\lambda^* - Q\lambda^*\rangle\end{aligned}$$

where $\langle Q'\lambda^{*\prime}, Q'\lambda^{*\prime} - Q'\lambda^*\rangle \leq 0$ by Lemma B.2, (B.3a). Then it can be further bounded by

$$\begin{aligned}\|Q\lambda^* - Q'\lambda^{*\prime}\|^2 \overset{(a)}{\leq}& \min_{\lambda\in\Delta^M} \|Q\lambda\|^2 - \min_{\lambda\in\Delta^M} \|Q'\lambda\|^2 + 2\|Q'\lambda^{*\prime}\|\|(Q' - Q)\lambda^*\| \\ =& - \max_{\lambda\in\Delta^M} -\|Q\lambda\|^2 + \max_{\lambda\in\Delta^M} -\|Q'\lambda\|^2 + 2\|Q'\lambda^{*\prime}\|\|(Q' - Q)\lambda^*\| \\ \overset{(b)}{\leq}& \max_{\lambda\in\Delta^M} \big(\|Q\lambda\|^2 - \|Q'\lambda\|^2\big) + 2\|Q'\lambda^{*\prime}\|\|(Q' - Q)\lambda^*\| \\ \overset{(c)}{\leq}& \max_{\lambda\in\Delta^M} \|(Q - Q')\lambda\|\big(\|Q\lambda\| + \|Q'\lambda\|\big) + 2\|Q'\lambda^{*\prime}\|\|(Q' - Q)\lambda^*\| \\ \leq& 4\max\Big\{\sup_{\lambda\in\Delta^M} \|Q\lambda\|, \sup_{\lambda\in\Delta^M} \|Q'\lambda\|\Big\} \cdot \sup_{\lambda\in\Delta^M} \|(Q - Q')\lambda\|\end{aligned}$$

where (a) follows from Cauchy-Schwarz inequality; (b) follows from subadditivity of maximum operator; (c) follows from triangle inequality. The proof is complete. ■

With the property in Lemma 4.1, we are able to derive the CA direction distances of SMG (Algorithm 3) and MoCo (Algorithm 4), as summarized in Theorem 4.2.

Theorem 4.2 (CA direction distances of SMG and MoCo). *Suppose either: 1) Assumptions 1, 3 hold; or, 2) Assumptions 1, 2 hold, with ℓ_f defined in Lemma 3.1. Considering $\{x_t\}$ and $\{\lambda_t\}$ generated by SMG with batch size $\mathcal{O}(t)$ at the t-th iteration or MoCo, both with $0 < \alpha_t = \alpha \leq 1/(2\ell_{f,1})$, then under either condition 1) or 2), their CA direction distances can be bounded by*

$$\textit{(SMG)}\ \ \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{E}_{\mathrm{ca}}(x_t, \lambda_{t+1}) = \mathcal{O}(T^{-\frac{1}{2}}); \qquad \textit{(MoCo)}\ \ \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{E}_{\mathrm{ca}}(x_t, \lambda_{t+1}) = \mathcal{O}(T^{-\frac{1}{4}}). \tag{4.5}$$

Proof of Theorem 4.2 is deferred to Appendix B.4. Theorem 4.2 indicates that increasing the batch size during optimization as in SMG or using momentum-based methods for gradient estimation as in MoCo can both mitigate the bias in the CA direction, and lead to the convergence of the CA direction distances for stochastic MOL algorithms.

Based on Lemma 4.1, we derive improved PS optimization error bounds for SMG and MoCo. In addition, we remove the assumption of bounded function values on the trajectory, by deriving a tighter bound on the inner product term. The PS optimization error bounds of SMG (Algorithm 3) and MoCo (Algorithm 4) are summarized in Theorem 4.3.

Theorem 4.3 (PS optimization errors of SMG and MoCo). *Suppose Assumptions 1 and 3 hold. Define c_F such that $\mathbb{E}_A[F_S(x_0)\lambda_0] - \min_{x\in\mathbb{R}^d} \mathbb{E}_A[F_S(x)\lambda_0] \leq c_F$. Considering $\{x_t\}$ generated by SMG with batch size $\mathcal{O}(t)$ at the t-th iteration or MoCo, both with $\alpha_t = \alpha \leq 1/(2\ell_{f,1})$, and proper choices of α, β depending on T, then their PS optimization errors can be bounded by*

$$\textit{(SMG)}\ \ \mathbb{E}_A\Big[\min_{t\in[T]} R_{\rm opt}(x_t)\Big] = \tilde{\mathcal{O}}\Big(T^{-\frac{1}{8}}\Big);\quad \textit{(MoCo)}\ \ \mathbb{E}_A\Big[\min_{t\in[T]} R_{\rm opt}(x_t)\Big] = \mathcal{O}\Big(T^{-\frac{1}{16}}\Big). \tag{4.6}$$

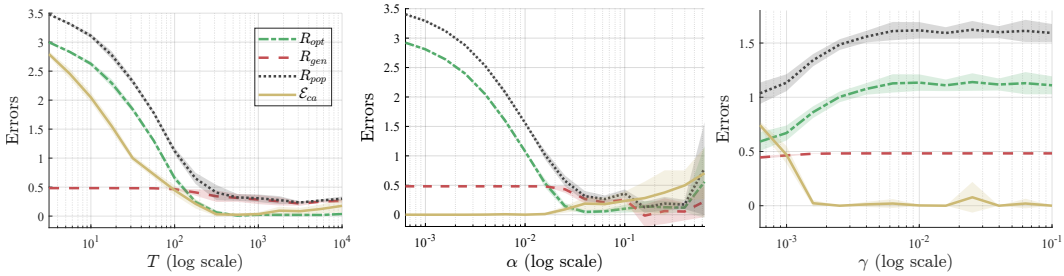
Proof of Theorem 4.3 is deferred to Appendix C.2. Theorem 4.3 provides the PS optimization error guarantees of SMG and MoCo under the same assumptions as Theorem 3.5, which relaxes the assumption of bounded function values on the optimization trajectory as used in (Fernando et al., 2023; Zhou et al., 2022). It also improves the convergence rate of MoCo in PS optimization error without such an assumption, see the comparison in Table 2.

5 Related Works

We review related work from the following three aspects – multi-task learning, theory of MOL, and generalization based on algorithm stability.

Multi-task learning (MTL). As one application of MOL, MTL leverages shared information among various tasks to train models to perform multiple tasks, and has been widely applied to natural language processing, computer vision, and robotics (Zhang and Yang, 2021). A simple method for MTL is to take the weighted average of the per-task losses as the objective. The weights can be static or dynamic during optimization. Weights for different tasks can be chosen based on different criteria such as gradient norms (Chen et al., 2018) or task difficulty (Guo et al., 2018). These methods are often heuristic and designed for specific applications. Another line of work tackles MTL through MOL (Sener and Koltun, 2018; Liu et al., 2021a). A foundational algorithm in this regard is MGDA (Désidéri, 2012), which takes dynamic weighting of gradients to obtain a CA direction for all objectives. Stochastic variants of MGDA with optimization convergence guarantees have been proposed in (Liu and Vicente, 2021; Zhou et al., 2022; Fernando et al., 2023). Algorithms for finding a set of Pareto optimal models have been proposed in (Navon et al., 2020; Liu et al., 2021b; Momma et al., 2022), to name a few.

Theory of MOL. Optimization analysis for the deterministic MGDA algorithm has been provided in (Fliege et al., 2019). Later on, stochastic variants of MGDA were introduced (Liu and Vicente, 2021; Zhou et al., 2022; Fernando et al., 2023) with bias reduction schemes and theoretical guarantees of PS optimization error. However, this can also be achieved by the simplest static weighting method. Therefore, although the community has a rich history of investigating the optimization of MOL algorithms, their theoretical benefits over static weighting, and their generalization guarantees remain open. Not until recently, generalization guarantees for MOL were theoretically analyzed. In (Cortes et al., 2020), a min-max formulation to solve the MOL problem is analyzed, where the weights are chosen based on the maximum function values, rather than the CA direction. More recently, (Súkeník and Lampert, 2022) provides generalization guarantees for MOL for a more general class of weighting. These two works analyze generalization based on Rademacher complexity of the



(a) Number of iterations T. (b) Different step size α. (c) Different step size γ.

Figure 3: Optimization, generalization, and CA direction distances of MoDo in the SC case under various T, α, γ, ($T = 100$, $\alpha = 0.01$, $\gamma = 0.001$ by default).

hypothesis class of the learner, with generalization error bound independent of the training process. Different from these works, we use the algorithm stability framework to derive the first algorithm-dependent generalization error bounds, highlighting the effect of the training dynamics. In addition, in contrast to prior works for MOL theory, which focus solely on either optimization (Zhou et al., 2022; Fernando et al., 2023) or generalization (Cortes et al., 2020; Súkeník and Lampert, 2022), we propose a holistic framework to analyze the three types of errors, namely, optimization, generalization, and CA distance in MOL with an instantiation of the proposed MoDo algorithm. This allows us to study how the theoretical test performance depends on hyperparameters such as the number of iterations and step sizes, and how to choose hyperparameters to achieve the best trade-off among the errors.

Algorithm stability and generalization. Stability analysis dates back to the work (Devroye and Wagner, 1979) in 1970s. Uniform stability and its relationship with generalization were studied in (Bousquet and Elisseeff, 2002) for the exact minimizer of the ERM problem with strongly convex objectives. The work (Hardt et al., 2016) pioneered the stability analysis for stochastic gradient descent (SGD) algorithms with convex and smooth objectives. The results were extended and refined in (Kuzborskij and Lampert, 2018) with data-dependent bounds, in (Charles and Papailiopoulos, 2018; Richards and Kuzborskij, 2021; Lei et al., 2022) for non-convex objectives, and in (Bassily et al., 2020; Lei and Ying, 2020) for SGD with non-smooth and convex losses. However, all these studies mainly focus on single-objective learning problems. To our best knowledge, there is no existing work on the stability and generalization analysis for multi-objective learning problems and our results on its stability and generalization are the *first-ever-known* ones.

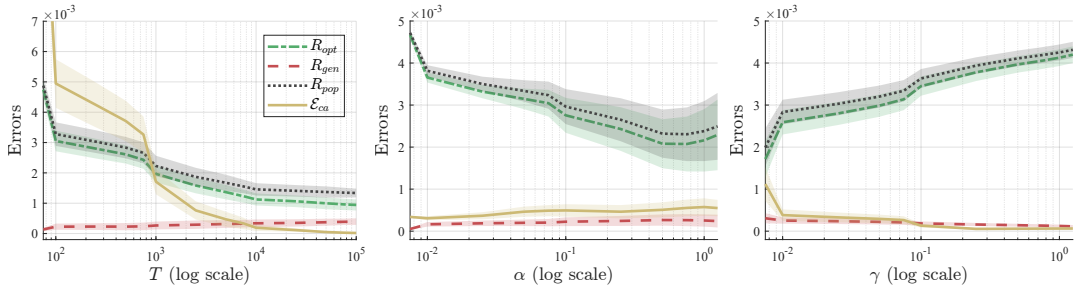
6 Experiments

In this section, we conduct experiments to further demonstrate the three-way trade-off among the optimization, generalization, and conflict avoidance of the MoDo algorithm. An average over 10 random seeds with 0.5 standard deviation is reported.

6.1 Synthetic experiments

6.1.1 Experiments on toy strongly-convex objectives

Our theory in the SC case is first verified through a synthetic experiment; see the details in Appendix D.1. Figure 3a shows the PS optimization error and PS population risk, as well



(a) Number of iterations T (b) Different step size α (c) Different step size γ

Figure 4: Optimization, generalization, and CA direction distances of MoDo on MNIST classification under various T, α, γ, ($T = 1000$, $\alpha = 0.1$, $\gamma = 0.01$ by default).

as the distance to CA direction, decreases as T increases, which corroborates Lemma 3.3, and Theorem 3.5. In addition, the generalization error, in this case, does not vary much with T, verifying Theorems 3.2. In Figure 3b, the optimization error first decreases and then increases as α increases, which is consistent with Theorem 3.5. Notably, we observe a threshold for α below which the distance to the CA direction converges even when the optimization error does not converge, while beyond which the distance to the CA direction becomes larger, verifying Lemma 3.3. Additionally, Figure 3c demonstrates that increasing γ enlarges the PS optimization error, PS generalization error, and thus the PS population risk, but decreases the distance to CA direction, which supports Lemma 3.3.

6.1.2 Multi-objective MNIST Experiments

We further verify our theory in the NC case on MNIST image classification (LeCun, 1998) using a multi-layer perceptron and three objectives: cross-entropy, mean squared error (MSE), and Huber loss. Following Section 2.2, we evaluate the performance in terms of $R_{\text{pop}}(x)$, $R_{\text{opt}}(x)$, $R_{\text{gen}}(x)$, and $\mathcal{E}_{\text{ca}}(x, \lambda)$. The exact PS population risk $R_{\text{pop}}(x)$ is not accessible without the true data distribution. To estimate the PS population risk, we evaluate $\min_{\lambda \in \Delta^M} \|\nabla F_{S_{\text{te}}}(x)\lambda\|$ on the testing data set S_{te} that is independent of training data set S. The PS optimization error $R_{\text{opt}}(x)$ is obtained by $\min_{\lambda \in \Delta^M} \|\nabla F_S(x)\lambda\|$, and the PS generalization error $R_{\text{gen}}(x)$ is estimated by $\min_{\lambda \in \Delta^M} \|\nabla F_{S_{\text{te}}}(x)\lambda\| - R_{\text{opt}}(x)$.

We examine the impact of different T, α, γ on the errors in Figure 4. Figure 4a shows that increasing T reduces optimization error and CA direction distance but increases generalization error, aligning with Theorems 3.1, 3.3, and 3.5. Figure 4b shows that increasing α leads to an initial decrease and subsequent increase in PS optimization error and population risk. which aligns with Theorem 3.5 and (3.13). On the other hand, there is an overall increase in CA direction distance with α, which aligns with Theorem 3.3. Figure 4c shows that increasing γ increases both the PS population and optimization errors but decreases CA direction distance. This matches our bounds for PS optimization error in Theorem 3.5, PS population risk in (3.13), and CA direction distance in Theorem 3.3.

7 Conclusions, Limitations, and Future Work

This work studied the three-way trade-off in MOL – among optimization, generalization, and conflict avoidance. Our results showed that, in the general nonconvex setting, the well-known

trade-off between optimization and generalization controlled by the number of iterations also exists in MOL. Moreover, dynamic weighting algorithms like MoDo introduced a new dimension of trade-off in terms of conflict avoidance compared to static weighting. We demonstrated that this three-way trade-off can be controlled by the step size γ for updating the dynamic weighting parameter λ and the number of iterations T. Proper choice of these parameters led to decent performance on all three metrics. We also demonstrated the power of this new analytical framework by applying it to analyze SMG and MoCo.

Limitations and future work. This work focuses on MOL with smooth objectives in unconstrained settings. Future research could explore the theory of non-smooth objectives or constrained learning. In addition, there is still room for improvement on the complexity of the proposed MoDo algorithm and the corresponding trade-off by adopting variance reduction techniques or implementing objective sampling, which we leave for future work. Our work has broad implications in advancing both the theory and practice of multi-objective optimization, with potential future applications as follows. *1) Theoretical Applications:* Our theoretical framework extends its utility beyond our proposed MoDo algorithm, allowing analysis of various MOL algorithms like CAGrad and PCGrad. Additionally, it aids in the investigation of the advantages of MOL algorithms over static weighting in reducing CA distance. This validates their use when CA distance reduction is crucial. *2) Practical Applications:* Our theory is crucial for optimizing hyperparameters (e.g., step size, iterations) to minimize testing risks effectively. It also enables informed algorithm selection based on the trade-off among three errors. Lastly, our theory may inspire the development of MOL algorithms that balance these errors more effectively.

Appendix

Appendix A. Bounding the PS Generalization Error

A.1 Proof of Propositions 3.1 and 3.2

Proof. [Proof of Proposition 3.1] For a given model x, it holds that

$$\begin{aligned} R_{\text{gen}}(x) &= \min_{\lambda\in\Delta^M} \|\nabla F(x)\lambda\| - \min_{\lambda\in\Delta^M} \|\nabla F_S(x)\lambda\| = -\max_{\lambda\in\Delta^M} -\|\nabla F(x)\lambda\| + \max_{\lambda\in\Delta^M} -\|\nabla F_S(x)\lambda\| \\ &\overset{(a)}{\leq} \max_{\lambda\in\Delta^M} (\|\nabla F(x)\lambda\| - \|\nabla F_S(x)\lambda\|) \overset{(b)}{\leq} \max_{\lambda\in\Delta^M} (\|(\nabla F(x) - \nabla F_S(x))\lambda\|) \\ &\overset{(c)}{\leq} \max_{\lambda\in\Delta^M} (\|\nabla F(x) - \nabla F_S(x)\|_{\text{F}}\|\lambda\|_{\text{F}}) \leq \|\nabla F(x) - \nabla F_S(x)\|_{\text{F}} \end{aligned} \tag{A.1}$$

where (a) follows from the subadditivity of max operator, (b) follows from triangle inequality, (c) follows from Cauchy-Schwartz inequality. Setting $x = A(S)$, and taking expectation over A, S on both sides of the above inequality proves the result. ■

Proof. [Proof of Proposition 3.2] The proof extends that of (Lei, 2023) for single objective learning to our MOL setting. Recall that $S = \{z_1, \ldots, z_n\}$, which are drawn i.i.d. from the data distribution $\mathcal{D}$. Define the perturbed dataset $S^{(i)} = \{z_1, \ldots, z_i', \ldots, z_n\}$ sampled i.i.d. from $\mathcal{D}$ with z_i' independent of z_j, for all $i, j \in [n]$. Let $\tilde{z}$ be an independent sample of z_j, z_j' , for

all $j \in [n]$, and from the same distribution $\mathcal{D}$. We first decompose the difference of population gradient and empirical gradient on the algorithm output $n(\nabla F(A(S)) - \nabla F_S(A(S)))$ as follows using the gradient on $A(S^{(i)})$. Since $\mathbb{E}_{\tilde{z}}[\nabla F_{\tilde{z}}(A(S))] = \nabla F(A(S))$, it holds that

$$
\begin{aligned}
&n\big(\nabla F(A(S)) - \nabla F_S(A(S))\big) = n\mathbb{E}_{\tilde{z}}[\nabla F_{\tilde{z}}(A(S))] - n\nabla F_S(A(S)) \\
=&n\mathbb{E}_{\tilde{z}}[\nabla F_{\tilde{z}}(A(S))] - \left(\sum_{i=1}^n \nabla F_{z_i}(A(S))\right) + \sum_{i=1}^n \Big(\mathbb{E}_{z_i'}[\nabla F(A(S^{(i)}))] - \mathbb{E}_{z_i'}[\nabla F(A(S^{(i)}))]\Big) \\
&+ \sum_{i=1}^n \Big(\mathbb{E}_{z_i'}[\nabla F_{z_i}(A(S^{(i)}))] - \mathbb{E}_{z_i'}[\nabla F_{z_i}(A(S^{(i)}))]\Big) \\
=&\sum_{i=1}^n \mathbb{E}_{\tilde{z},z_i'}[\nabla F_{\tilde{z}}(A(S)) - \nabla F_{\tilde{z}}(A(S^{(i)}))] + \sum_{i=1}^n \underbrace{\mathbb{E}_{z_i'}[\mathbb{E}_{\tilde{z}}[\nabla F_{\tilde{z}}(A(S^{(i)}))] - \nabla F_{z_i}(A(S^{(i)}))]}_{\xi_i(S)} \\
&+ \sum_{i=1}^n \mathbb{E}_{z_i'}\big[\nabla F_{z_i}(A(S^{(i)})) - \nabla F_{z_i}(A(S))\big]
\end{aligned} \tag{A.2}
$$

where the last equality follows from rearranging and that $z_i, z_i', \tilde{z}$ are mutually independent. Applying triangle inequality to (A.2), it then follows that

$$
\begin{aligned}
n\|\nabla F(A(S)) - \nabla F_S(A(S))\|_{\mathrm{F}} \le &\sum_{i=1}^n \mathbb{E}_{\tilde{z},z_i'}[\|\nabla F_{\tilde{z}}(A(S)) - \nabla F_{\tilde{z}}(A(S^{(i)}))\|_{\mathrm{F}}] + \Big\|\sum_{i=1}^n \xi_i(S)\Big\|_{\mathrm{F}} \\
&+ \sum_{i=1}^n \mathbb{E}_{z_i'}[\|\nabla F_{z_i}(A(S^{(i)})) - \nabla F_{z_i}(A(S))\|_{\mathrm{F}}].
\end{aligned} \tag{A.3}
$$

Note S and $S^{(i)}$ differ by a single sample. By Definition 3.1, the MOL uniform stability ϵ_{F}, and Jensen's inequality, we further get

$$
n\mathbb{E}\left[\|\nabla F(A(S)) - \nabla F_S(A(S))\|_{\mathrm{F}}\right] \le 2n\epsilon_{\mathrm{F}} + \mathbb{E}\Big[\Big\|\sum_{i=1}^n \xi_i(S)\Big\|_{\mathrm{F}}\Big]. \tag{A.4}
$$

We then proceed to bound $\mathbb{E}\big[\big\|\sum_{i=1}^n \xi_i(S)\big\|_{\mathrm{F}}\big]$, which satisfies

$$
\left(\mathbb{E}\Big[\Big\|\sum_{i=1}^n \xi_i(S)\Big\|_{\mathrm{F}}\Big]\right)^2 \le \mathbb{E}\Big[\Big\|\sum_{i=1}^n \xi_i(S)\Big\|_{\mathrm{F}}^2\Big] = \sum_{i=1}^n \underbrace{\mathbb{E}\big[\|\xi_i(S)\|_{\mathrm{F}}^2\big]}_{J_{1,i}} + \sum_{i,j\in[n]:i\neq j} \underbrace{\mathbb{E}[\langle \xi_i(S), \xi_j(S)\rangle]}_{J_{2,i,j}}. \tag{A.5}
$$

For $J_{1,i}$, according to the definition of $\xi_i(S)$ in (A.2) and Jensen inequality, it holds that

$$
\begin{aligned}
J_{1,i} = \mathbb{E}[\|\xi_i(S)\|_{\mathrm{F}}^2] &= \mathbb{E}\left[\big\|\mathbb{E}_{z_i'}\big[\mathbb{E}_{\tilde{z}}[\nabla F_{\tilde{z}}(A(S^{(i)}))] - \nabla F_{z_i}(A(S^{(i)}))\big]\big\|_{\mathrm{F}}^2\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\big\|\mathbb{E}_{\tilde{z}}[\nabla F_{\tilde{z}}(A(S^{(i)}))] - \nabla F_{z_i}(A(S^{(i)}))\big\|_{\mathrm{F}}^2\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\big\|\mathbb{E}_{\tilde{z}}[\nabla F_{\tilde{z}}(A(S))] - \nabla F_{z_i'}(A(S))\big\|_{\mathrm{F}}^2\right] = \mathbb{E}\left[\mathbb{V}_{\tilde{z}}(\nabla F_{\tilde{z}}(A(S)))\right],
\end{aligned} \tag{A.6}
$$

where (a) follows from Jensen's inequality, (b) follows from the symmetry between z_i and z_i'. To bound $J_{2,i,j}$ with $i \neq j$, introduce $S'' = \{z_1'', \ldots, z_n''\}$ drawn i.i.d. from the data distribution $\mathcal{D}$. For each $i \neq j \in [n]$, introduce S_j as a neighboring dataset of S by replacing z_j with z_j'', and $S_j^{(i)}$ as a neighboring dataset of $S^{(i)}$ by replacing z_j with z_j'', i.e.,

$$S_j = \{z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n\}, \tag{A.7a}$$

$$S_j^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n\}. \tag{A.7b}$$

Then the idea is to bound $J_{2,i,j}$ using the newly introduced neighboring datasets S_j and $S_j^{(i)}$, so as to connect to the definition of the stability ϵ_{F}. We first show that $\mathbb{E}\left[\langle \xi_i(S), \xi_j(S)\rangle\right] = \mathbb{E}\left[\langle \xi_i(S) - \xi_i(S_j), \xi_j(S) - \xi_j(S_i)\rangle\right]$ because for $i \neq j$,

$$\mathbb{E}\left[\langle \xi_i(S_j), \xi_j(S)\rangle\right] \overset{(c)}{=} 0, \quad \mathbb{E}\left[\langle \xi_i(S_j), \xi_j(S_i)\rangle\right] \overset{(d)}{=} 0, \quad \mathbb{E}\left[\langle \xi_i(S_j), \xi_j(S_i)\rangle\right] \overset{(e)}{=} 0. \tag{A.8}$$

For $i \neq j$, (c) follows from

$$\mathbb{E}\left[\langle \xi_i(S_j), \xi_j(S)\rangle\right] = \mathbb{E}\mathbb{E}_{z_j}\left[\langle \xi_i(S_j), \xi_j(S)\rangle\right] = \mathbb{E}\left[\left\langle \xi_i(S_j), \mathbb{E}_{z_j}\left[\xi_j(S)\right]\right\rangle\right] = 0, \tag{A.9}$$

where the second identity holds since $\xi_i(S_j)$ is independent of z_j and the last identity follows from $\mathbb{E}_{z_j}\left[\xi_j(S)\right] = 0$ due to the symmetry between $\tilde{z}$ and z_i, and their independence with $S^{(i)}$, derived as

$$\mathbb{E}_{z_i}\left[\xi_i(S)\right] = \mathbb{E}_{z_i}\Big[\mathbb{E}_{z_i'}[\mathbb{E}_{\tilde{z}}[\nabla F_{\tilde{z}}(A(S^{(i)}))] - \nabla F_{z_i}(A(S^{(i)}))]\Big] = 0, \quad \forall i \in [n]. \tag{A.10}$$

In a similar way, for $i \neq j$, (d) and (e) follow from

$$\mathbb{E}\left[\langle \xi_i(S), \xi_j(S_i)\rangle\right] = \mathbb{E}\mathbb{E}_{z_i}\left[\langle \xi_i(S), \xi_j(S_i)\rangle\right] = \mathbb{E}\left[\langle \xi_j(S_i), \mathbb{E}_{z_i}\left[\xi_i(S)\right]\rangle\right] = 0, \tag{A.11}$$

$$\mathbb{E}\left[\langle \xi_i(S_j), \xi_j(S_i)\rangle\right] = \mathbb{E}\mathbb{E}_{z_i}\left[\langle \xi_i(S_j), \xi_j(S_i)\rangle\right] = \mathbb{E}\left[\langle \xi_j(S_i), \mathbb{E}_{z_i}\left[\xi_i(S_j)\right]\rangle\right] = 0. \tag{A.12}$$

Based on (A.8), for $i \neq j$ we have

$$\begin{aligned} J_{2,i,j} =& \mathbb{E}\left[\langle \xi_i(S), \xi_j(S)\rangle\right] = \mathbb{E}\left[\langle \xi_i(S) - \xi_i(S_j), \xi_j(S) - \xi_j(S_i)\rangle\right] \\ \leq& \mathbb{E}\left[\|\xi_i(S) - \xi_i(S_j)\|_{\mathrm{F}} \|\xi_j(S) - \xi_j(S_i)\|_{\mathrm{F}}\right] \\ \leq& \frac{1}{2}\mathbb{E}\left[\|\xi_i(S) - \xi_i(S_j)\|_{\mathrm{F}}^2\right] + \frac{1}{2}\mathbb{E}\left[\|\xi_j(S) - \xi_j(S_i)\|_{\mathrm{F}}^2\right] \end{aligned} \tag{A.13}$$

where we have used $ab \leq \frac{1}{2}\left(a^2 + b^2\right)$. According to the definition of $\xi_i(S)$ and $\xi_i(S_j)$ we know the following identity for $i \neq j$

$$\begin{aligned} \mathbb{E}\left[\|\xi_i(S) - \xi_i(S_j)\|_{\mathrm{F}}^2\right] =& \mathbb{E}\Big[\Big\|\mathbb{E}_{z_i'}\mathbb{E}_{\tilde{z}}\big[\nabla F_{\tilde{z}}(A(S^{(i)})) - \nabla F_{\tilde{z}}(A(S_j^{(i)}))\big] \\ &+ \mathbb{E}_{z_i'}\big[\nabla F_{z_i}(A(S_j^{(i)})) - \nabla F_{z_i}(A(S^{(i)}))\big]\Big\|_{\mathrm{F}}^2\Big]. \end{aligned} \tag{A.14}$$

It then follows from the inequality $(a+b)^2 \leq 2\left(a^2 + b^2\right)$ and the Jensen's inequality that

$$\mathbb{E}[\|\xi_i(S) - \xi_i(S_j)\|_{\mathrm{F}}^2] \leq 2\mathbb{E}[\|\nabla F_{\tilde{z}}(A(S^{(i)})) - \nabla F_{\tilde{z}}(A(S_j^{(i)}))\|_{\mathrm{F}}^2]$$

$$+ 2\mathbb{E}[\|\nabla F_{z_i}(A(S_j^{(i)})) - \nabla F_{z_i}(A(S^{(i)}))\|_{\mathrm{F}}^2]. \tag{A.15}$$

Since $S^{(i)}$, $S_j^{(i)}$ and $S^{(j)}$, $S_i^{(j)}$ are two pairs of neighboring datasets, it follows from the definition of stability that

$$\mathbb{E}\left[\|\xi_i(S) - \xi_i(S_j)\|_{\mathrm{F}}^2\right] \leq 4\epsilon_{\mathrm{F}}^2, \text{ and } \mathbb{E}\left[\|\xi_j(S) - \xi_j(S_i)\|_{\mathrm{F}}^2\right] \leq 4\epsilon_{\mathrm{F}}^2, \quad \forall i \neq j. \tag{A.16}$$

We can plug the above inequalities back into (A.13) and bound $J_{2,i,j}$ by

$$J_{2,i,j} = \mathbb{E}\left[\langle \xi_i(S), \xi_j(S)\rangle\right] \leq 4\epsilon_{\mathrm{F}}^2, \quad \forall i \neq j. \tag{A.17}$$

Combining the bound for $J_{1,i}$ in (A.6) and $J_{2,i,j}$ in (A.17) and substituting them back into (A.5), it then follows that

$$\begin{aligned}\mathbb{E}\Big[\Big\|\sum_{i=1}^n \xi_i(S)\Big\|_{\mathrm{F}}^2\Big] &= \mathbb{E}\Big[\sum_{i=1}^n \|\xi_i(S)\|_{\mathrm{F}}^2\Big] + \sum_{i,j\in[n]:i\neq j} \mathbb{E}[\langle \xi_i(S), \xi_j(S)\rangle] \\ &\leq n\mathbb{E}\left[\mathbb{V}_{\tilde{z}}(\nabla F_{\tilde{z}}(A(S)))\right] + 4n(n-1)\epsilon_{\mathrm{F}}^2. \end{aligned} \tag{A.18}$$

We can plug the above inequality back into (A.4), use the subadditivity of square root function, and get

$$n\mathbb{E}[\|\nabla F(A(S)) - \nabla F_S(A(S))\|_{\mathrm{F}}] \leq 4n\epsilon_{\mathrm{F}} + \sqrt{n\mathbb{E}\left[\mathbb{V}_{\tilde{z}}(\nabla F_{\tilde{z}}(A(S)))\right]}. \tag{A.19}$$

The proof is complete. ■

A.2 Proof of Theorem 3.1 – Generalization of MoDo in the NC case

In this subsection, we prove Theorem 3.1, which establishes the PS generalization error of MoDo, SMG, and MoCo in the nonconvex case.

Organization of proof. We first give the concept of sampling-determined MOL algorithms in Definition A.1, which generalizes the concept in (Lei, 2023) for single-objective learning. Then we show that MoDo is sampling-determined in Proposition A.1. Combining Propositions 3.1 and A.1, we are able to prove the upper bound of the MOL uniform stability. A matching lower bound of the MOL uniform stability is provided in Lemma A.2. Combining the upper and lower bounds, the proof for Theorem 3.1 is complete.

Definition A.1 (Sampling-determined algorithm (Lei, 2023)). *Let A be a randomized algorithm that randomly chooses an index sequence $I(A) = \{i_{t,s}\}$ to compute stochastic gradients. We say a symmetric algorithm A is sampling-determined if the output model is fully determined by $\{z_i : i \in I(A)\}$.*

Proposition A.1 (MoDo, SMG, MoCo, are sampling-determined). *MoDo (Algorithms 2, 3, and 4) are sampling determined. In other words, Let $I(A) = \{i_t\}$ be the sequence of index chosen by these algorithms from training set $S = \{z_1, \ldots, z_n\}$, and $z_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}$ for all $i \in [n]$ to build stochastic gradients, the output $A(S)$ is determined by $\{z_j \mid j \in I(A)\}$. To be precise, $A(S)$ is independent of z_j if $j \notin I(A)$.*

Proof. [Proposition A.1| Let $I(A) = \{I_1, \ldots, I_T\}$, $I_t = \{i_{t,s}\}_{s=1}^2$ and $i_{t,s} \in [n]$ for all $1 \le t \le T$. Let $Z_{I(A)} = \{z_{i_{t,s}} \mid t \in [T], s \in [2]\}$. By the description in Algorithm 2, $A(S) = G_{Z_{I_T}} \circ \cdots \circ G_{Z_{I_1}}(x_0)$, where $G_Z(\cdot)$ is the stochastic update function of the model parameter given random mini-batch Z. Therefore, for all possible random mini-batch Z selected by A, we have

$$\begin{aligned}\mathbb{P}(A(S) = x \mid z_j = z, j \notin I(A)) =& \mathbb{P}(G_{Z_{I_T}} \circ \cdots \circ G_{Z_{I_1}}(x_0) = x \mid z_j = z, j \notin I(A)) \\ =& \mathbb{P}(G_{Z_{I_T}} \circ \cdots \circ G_{Z_{I_1}}(x_0) = x \mid j \notin I(A)) \\ =& \mathbb{P}(A(S) = x \mid j \notin I(A)) \end{aligned} \tag{A.20}$$

where the last equality holds because $z_j \notin S_{I(A)}$, and z_j is independent of all elements in $S_{I(A)}$ by i.i.d. sampling. Therefore, $A(S)$ is independent of z_j if $j \notin I(A)$, MoDo is sampling-determined.

Similarly, for SMG, let $I_t = \{i_{t,s}\}_{s=1}^{|Z_t|}$, and $i_{t,s} \in [n]$ for all $t \in [T]$, then (A.20) still holds for A being the SMG algorithm. Therefore, SMG is also sampling-determined.

Finally, for MoCo, its update at iteration t depends on the stochastic sample selected at iteration t, as well as all the stochastic samples at previous iterations. Denote the update function at each iteration as $G_{Z_{I_{1:t}}}(x_t)$, where $Z_{I_{1:t}} = \{z_{I_1}, z_{I_2}, \ldots, z_{I_t}\}$, then we have

$$\begin{aligned}\mathbb{P}(A(S) = x \mid z_j = z, j \notin I(A)) =& \mathbb{P}(G_{Z_{I_{1:T}}} \circ \cdots \circ G_{Z_{I_{1:1}}}(x_0) = x \mid z_j = z, j \notin I(A)) \\ =& \mathbb{P}(G_{Z_{I_{1:T}}} \circ \cdots \circ G_{Z_{I_{1:1}}}(x_0) = x \mid j \notin I(A)) \\ =& \mathbb{P}(A(S) = x \mid j \notin I(A)). \end{aligned} \tag{A.21}$$

which proves MoCo is sampling-determined. ■

Lemma A.1. *(Lei, 2023, Theorem 5 (b)) Let A be a sampling-determined random algorithm (Definition A.1) and S, S' be neighboring datasets with n data points that differ only in the i-th data point. If $\sup_z \mathbb{E}_A\left[\|\nabla F_z(A(S))\|_{\mathrm{F}}^2 \mid i \in I(A)\right] \le G^2$ for any S, then*

$$\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2] \le 4G^2 \cdot \mathbb{P}\{i \in I(A)\}. \tag{A.22}$$

Lemma A.2 (Lower bound of MOL uniform stability in the NC case). *There exists a vector-valued objective function $F_z(x)$, where for each $m \in [M]$, $z \in \mathcal{Z}$, the scalar-valued function $f_{z,m}(x)$ is nonconvex and smooth, and there exists neighboring datasets S and S' with $|S| = |S'| = n$, which differ with at most one sample, and a randomized algorithm MoDo, denoted as A, such that the MOL uniform stability of the t-th iteration output with $t \in [T]$ is lower bounded by*

$$\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2] = \Omega\Big(\frac{t}{n}\Big).$$

Proof. From the definition of the sampling-determined algorithms, and that MoDo selects two samples at each iteration, we can compute the probability of $i^* \in I(A)$ as

$$\mathbb{P}(i^* \in I(A)) = 1 - \Big(\frac{n-1}{n}\Big)^{2T} \tag{A.23}$$

whose lower bound can be computed by

$$\mathbb{P}(i^* \in I(A)) = 1 - \Big(\frac{n-1}{n}\Big)^{2T} \overset{(a)}{\geq} 1 - 1 + \frac{1}{1+\frac{2T-1}{n}} \cdot \frac{2T}{n} = \frac{T}{n(1+\frac{2T-1}{n})} \overset{(b)}{\geq} \frac{2T}{3n} \tag{A.24}$$

where (a) follows from the inequality that $(1+c)^r \leq 1 + \frac{rc}{1-(r-1)c}$ for $c \in [-1, \frac{1}{r-1}), r > 1$, plugging in $r = 2T > 1$, $c = -1/n < 0 < 1/(r-1)$; and (b) follows from $T \leq n$.

The MOL uniform stability of a sampling-determined algorithm in the general non-convex case can then be lower bounded by

$$\begin{aligned}
&\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2] \\
=&\sup_z \Big(\mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2 \mid i^* \in I(A)] \cdot \mathbb{P}(i^* \in I(A)) \\
&\qquad + \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2 \mid i^* \notin I(A)] \cdot \mathbb{P}(i^* \notin I(A))\Big) \\
\geq&\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2 \mid i^* \in I(A)] \cdot \mathbb{P}(i^* \in I(A)) \\
\geq&\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2 \mid i^* \in I(A)] \cdot \frac{2T}{3n}.
\end{aligned} \tag{A.25}$$

We proceed to bound the term $\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2 \mid i^* \in I(A)]$ in the above inequality by constructing the following simple example with $M = 2$, $|S| = |S'| = n > 10$, $S = \{0, \ldots, 0\}$, $S' = \{0, \ldots, 0, -\frac{1}{8}\pi\}$.

$$\begin{aligned}
f_{z,1}(x) = f_{z,2}(x) &= \sin(x+z) \\
\nabla f_{z,1}(x) = \nabla f_{z,2}(x) &= \cos(x+z)
\end{aligned}$$

For algorithm A, choose $2 \leq T \leq 10 < n$, [1] step size $\alpha_t = \alpha = \frac{\pi}{80}$, initialization $x_0 = x_0' = \frac{1}{4}\pi$. Let $x_t = A_t(S)$, and $x_t' = A_t(S')$. Since $|\nabla f_{z,1}(x)| \leq 1$, we have

$$|x_0 - x_T| \leq \alpha \left| \sum_{t=0}^{T-1} \nabla F_z(x_t)\lambda_t \right| \leq \alpha \sum_{t=0}^{T-1} |\nabla F_z(x_t)\lambda_t| \leq \alpha T \leq \frac{1}{8}\pi. \tag{A.26}$$

Similarly, we have

$$|x_0' - x_T'| \leq \frac{1}{8}\pi. \tag{A.27}$$

Therefore, for all $t \in [T]$, it holds that

$$\frac{1}{8}\pi \leq x_t \leq \frac{3}{8}\pi, \quad \frac{1}{8}\pi \leq x_t' \leq \frac{3}{8}\pi. \tag{A.28}$$

We need to bound $\mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2 \mid i^* \in I(A)]$ in (A.25). Considering the case $i^* \in I(A)$, let $t_0 \in [T-1]$ denote the first iteration to select i^*, then

$$\nabla f_{z'}(x_{t_0}') - \nabla f_z(x_{t_0}) = \cos(x_{t_0}' + z') - \cos(x_{t_0} + z) = \cos\Big(x_{t_0}' - \frac{1}{8}\pi\Big) - \cos(x_{t_0})$$

1. This choice of T simplifies the analysis. Other choices are possible depending on the choice of x_0 and α_t.

$$= -2\sin\Big(x_{t_0} - \frac{1}{16}\pi\Big)\sin\Big(-\frac{1}{16}\pi\Big) \geq 0.076$$

which implies

$$\begin{aligned} x_{t_0+1} - x'_{t_0+1} =& x_{t_0} - x'_{t_0} + \alpha(\nabla f_{z'}(x'_{t_0}) - \nabla f_z(x_{t_0})) \\ =& \alpha(\nabla f_{z'}(x'_{t_0}) - \nabla f_z(x_{t_0})) \geq 0.076\alpha > 0. \end{aligned} \tag{A.29}$$

We then prove by induction that $x_T - x'_T \geq 0.076\alpha$ using the statements below:
1) $x_{t_0+1} - x'_{t_0+1} \geq 0.076\alpha$;
2) $x_{t+1} - x'_{t+1} \geq x_t - x'_t \geq 0.076\alpha$ if $x_t - x'_t \geq 0.076\alpha > 0$.

The first statement is proved in (A.29). The second statement can be proved by

$$\begin{aligned} x_{t+1} - x'_{t+1} =& x_t - x'_t + \alpha(\nabla f_{z'}(x'_t) - \nabla f_z(x_t)) \\ =& x_t - x'_t + \alpha(\cos(x'_t + z') - \cos(x_t + z)) \geq 0. \end{aligned}$$

The last inequality follows from that for $t_0 < t \leq T$, $\frac{1}{8}\pi \leq x'_t < x_t \leq \frac{3}{8}\pi$ as (A.28), where $\nabla f_z(x) = \cos(x + z)$ is monotonically decreasing with $x + z \in [0, \frac{1}{2}\pi]$. And since $x_t - x'_t > 0$, $z' \leq z$, $x'_t + z' \leq x_t + z$, therefore $\cos(x'_t + z') - \cos(x_t + z) \geq 0$. Then we arrive at

$$x_T - x'_T \geq 0.076\alpha \geq \frac{7\pi}{800}. \tag{A.30}$$

By the Mean value theorem, there exists $\bar{x} \in [x'_T, x_T] \subset [\frac{1}{8}\pi, \frac{3}{8}\pi]$ such that

$$|\nabla f_z(A(S)) - \nabla f_z(A(S'))| = |\nabla f_z(x_T) - \nabla f_z(x'_T)| = |\nabla^2 f_z(\bar{x})||x_T - x'_T| \geq \sin\Big(\frac{1}{8}\pi\Big)\frac{7\pi}{800}. \tag{A.31}$$

Therefore, combining (A.25) and (A.31) yields

$$\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2] = \Omega\big(T/n\big). \tag{A.32}$$

The proof of the lower bound in the nonconvex case is complete. ■

The following remark discusses the application of the above MOL uniform stability lower bound to single-objective learning (SOL).

Remark 4. *Our lower bound in the NC case can be easily reduced to the SOL problems with sampling-determined algorithms since our proof is based on the construction of a special case with identical multi-objectives where the update of λ does not affect the update of x. The reduction to the SOL setting is also the first lower bound with sampling-determined algorithms for SOL in the NC case that matches the upper bound in (Lei, 2023).*

A.2.1 Proof of Theorem 3.1

Proof. [Theorem 3.1| From Proposition A.1, MoDo algorithm is sampling-determined. Then based on Lemma A.1, its MOL uniform stability in Definition 3.1 can be bounded by

$$\epsilon_{\mathrm{F}}^2 \leq 4G^2 \cdot \mathbb{P}\{i \in I(A)\}. \tag{A.33}$$

Let i_t be the index of the sample selected by A at t-th iteration, $I_t(A)$ be the indices of the samples selected by A up to the t-th iteration, with $t \in [T]$, and i^* be the index of the data point that is different in S and S'. Then

$$\mathbb{P}\{i^* \in I_t(A)\} \leq \sum_{k=1}^{t} \mathbb{P}\{i_k = i^*\} \leq \frac{t}{n}. \tag{A.34}$$

Combining (A.33) and (A.34) gives the MOL uniform stability of the t-th iterate with $t \in [T]$ is upper bounded by

$$\epsilon_{\mathrm{F}}^2 \leq \frac{4G^2 t}{n} \leq \frac{4G^2 T}{n}. \tag{A.35}$$

This proves the upper bound in a). The proof of the lower bound in b) is given in Lemma A.2 in Appendix A.2. Then based on Propositions 3.1-3.2, the PS generalization error is upper bounded by

$$\begin{aligned}
\mathbb{E}_{A,S}[R_{\mathrm{gen}}(A_t(S))] &\leq \mathbb{E}_{A,S}[\|\nabla F(A_t(S)) - \nabla F_S(A_t(S))\|_{\mathrm{F}}] && \text{by Proposition 3.1} \\
&\leq 4\epsilon_{\mathrm{F}} + \sqrt{n^{-1}\mathbb{E}_S\left[\mathbb{V}_{z\sim\mathcal{D}}(\nabla F_z(A_t(S)))\right]} && \text{by Proposition 3.2} \\
&= \mathcal{O}(T^{\frac{1}{2}} n^{-\frac{1}{2}}). && \text{by (A.35)}
\end{aligned}$$

The proof of the upper bound is complete. Lemma A.2 provides the lower bound. Combining both completes the proof. ■

A.3 Proof of Theorem 4.1 – Generalization of SMG and MoCo in the NC case

Proof. [Theorem 4.1] The proof follows similar steps as the proof for Theorem 3.1. First, Proposition A.1 states that SMG and MoCo are sampling-determined, and thus their MOL uniform stability depends on the probability $\mathbb{P}\{i \in I(A)\}$ by Lemma A.1.

For MoCo, following similar proof steps as MoDo in Theorem 3.1, we have the MOL uniform stability of MoCo is $\epsilon_{\mathrm{F}}^2 = \mathcal{O}(Tn^{-1})$. Then combining with Proposition 3.2 which connects the MOL uniform stability and PS generalization error, it yields that their PS generalization errors are $\mathbb{E}_{A,S}[R_{\mathrm{gen}}(A_t(S))] = \mathcal{O}(T^{\frac{1}{2}} n^{-\frac{1}{2}})$ for all $t \in [T]$.

For SMG, suppose we choose t as the batch size at iteration t. Let i_t be the set of indices of the samples selected by SMG at t-th iteration with $|i_t| = t$, $I_t(A)$ be the indices of the samples selected by SMG up to the t-th iteration, with $t \in [T]$, and i^* be the index of the data point that is different in S and S'. Then

$$\mathbb{P}\{i^* \in I_t(A)\} \leq \sum_{k=1}^{t} \mathbb{P}\{i^* \in i_k\} \leq \sum_{k=1}^{t} \frac{k}{n} = \frac{(1+t)t}{2n}. \tag{A.36}$$

Combining (A.36) with Lemma A.1, we have the MOL uniform stability of SMG is upper bounded by

$$\epsilon_{\mathrm{F}}^2 \leq \frac{2G^2(1+T)T}{n}. \tag{A.37}$$

Then based on Propositions 3.1-3.2, the PS generalization error of SMG is upper bounded by

$$\mathbb{E}_{A,S}[R_{\text{gen}}(A_t(S))] \leq 4\epsilon_{\text{F}} + \sqrt{n^{-1}\mathbb{E}_S\left[\mathbb{V}_{z\sim\mathcal{D}}(\nabla F_z(A_t(S)))\right]} = \mathcal{O}(Tn^{-\frac{1}{2}}). \tag{A.38}$$

The proof is complete. ■

A.4 Proof of Lemma 3.1 – x_t bounded in the SC case

Organization of proof. Without loss of generality, we assume $\inf_{x\in\mathbb{R}^d} f_{z,m}(x) < \infty$ for all $m \in [M]$ and $z \in \mathcal{Z}$ in the SC case. Lemma A.3 shows that the optimal solution of $F_z(x)\lambda$ given any stochastic sample $z \in \mathcal{Z}$, and weighting parameter $\lambda \in \Delta^M$, is bounded. Lemma A.4 shows that if the argument parameter is bounded, then the updated parameter by MoDo at each iteration is also bounded by exploiting the co-coerciveness of strongly convex and smooth objectives. Finally, based on Lemma A.3 and Lemma A.4, we first prove that with a bounded initialization x_0, the model parameter $\{x_t\}_{t=1}^T$ generated by MoDo algorithm is bounded on the trajectory. Then by the smoothness assumption of $f_{z,m}(x)$, we immediately have that $\|\nabla f_{z,m}(x)\|$ is bounded for $x \in \{x_t\}_{t=1}^T$ generated by MoDo algorithm, which completes the proof of Lemma 3.1.

A.4.1 Auxiliary Lemmas

This section provides the auxiliary lemmas to prove Lemma 3.1. The proofs can be found at (Chen et al., 2023a, Appendix A.4.1)

Lemma A.3. *Suppose Assumptions 1, 2 hold. W.l.o.g., assume* $\inf_{x\in\mathbb{R}^d} f_{z,m}(x) < \infty$ *for all* $m \in [M]$ *and* $z \in \mathcal{Z}$. *For any given* $\lambda \in \Delta^M$, *and stochastic sample* $z \in \mathcal{Z}$, *define* $x^*_{\lambda,z} = \arg\min_{x\in\mathbb{R}^d} F_z(x)\lambda$, *then* $\inf_{x\in\mathbb{R}^d} F_z(x)\lambda < \infty$ *and* $\|x^*_{\lambda,z}\| < \infty$, *i.e., there exist finite positive constants* c_{F^*} *and* c_{x^*} *such that*

$$\inf_{x\in\mathbb{R}^d} F_z(x)\lambda \leq c_{F^*} \quad \text{and} \quad \|x^*_{\lambda,z}\| \leq c_{x^*}. \tag{A.39}$$

Lemma A.4. *Suppose Assumptions 1, 2 hold, and define* $\kappa = 3\ell_{f,1}/\mu \geq 3$. *For any given* $\lambda \in \Delta^M$, *and a stochastic sample* $z \in \mathcal{Z}$, *define* $x^*_{\lambda,z} = \arg\min_x F_z(x)\lambda$. *Then by Lemma A.3, there exists a positive finite constant* $c_{x,1} \geq c_{x^*}$ *such that* $\|x^*_{\lambda,z}\| \leq c_{x^*} \leq c_{x,1}$. *Recall the multi-objective gradient update is*

$$G_{\lambda,z}(x) = x - \alpha\nabla F_z(x)\lambda \tag{A.40}$$

with step size $0 \leq \alpha \leq \ell_{f,1}^{-1}$. *Defining* $c_{x,2} = (1+\sqrt{2\kappa})c_{x,1}$, *we have that*

$$\text{if} \quad \|x\| \leq c_{x,2}, \quad \text{then} \quad \|G_{\lambda,z}(x)\| \leq c_{x,2}. \tag{A.41}$$

A.4.2 Proof of Lemma 3.1

Proof. [Lemma 3.1] We first prove (a), i.e., $\{x_t\}$ generated by the MoDo algorithm are bounded. Define $\kappa = 3\ell_{f,1}/\mu$ and $x^*_{\lambda,z} = \arg\min_x F_z(x)\lambda$ with $\lambda \in \Delta^M$. Under Assumptions 1, 2, by Lemma A.3, $\|x^*_{\lambda,z}\| < \infty$, i.e., there exists a finite positive constant c_{x^*} such

that $\|x^*_{\lambda,z}\| \leq c_{x^*}$. Choose the initial iterate to be bounded, i.e., there exists a finite positive constant c_{x_0} such that $\|x_0\| \leq c_{x_0}$. Then we will prove that for $\{x_t\}$ generated by MoDo algorithm with $\alpha_t = \alpha$ and $0 \leq \alpha \leq \ell_{f,1}^{-1}$, we have

$$\|x_t\| \leq c_x, \quad \text{with} \quad c_x = \max\{(1+\sqrt{2\kappa})c_{x^*}, c_{x_0}\}. \tag{A.42}$$

To prove (A.42), we rely on Lemma A.4, which states that if the current iterate x_t is bounded, then with MoDo update, the next iterate x_{t+1} is also bounded. Let $c_{x,1} = \max\{(1+\sqrt{2\kappa})^{-1}c_{x_0}, c_{x^*}\}$, and $c_{x,2} = (1+\sqrt{2\kappa})c_{x,1} = \max\{c_{x_0}, (1+\sqrt{2\kappa})c_{x^*}\}$ in Lemma A.4. We then consider the following two cases:

1) If $(1+\sqrt{2\kappa})c_{x^*} \leq c_{x_0}$, then $\|x^*_{\lambda,z}\| \leq c_{x^*} \leq (1+\sqrt{2\kappa})^{-1}c_{x_0}$. Then it satisfies the condition in Lemma A.4 that $\|x^*_{\lambda,z}\| \leq c_{x,1}$ and $\|x_0\| \leq c_{x,2}$. Applying Lemma A.4 yields $\|x_1\| \leq c_{x,2}$.

2) If $(1+\sqrt{2\kappa})c_{x^*} > c_{x_0}$, then $\|x_0\| \leq c_{x_0} < (1+\sqrt{2\kappa})c_{x^*}$. Then it satisfies the condition in Lemma A.4 that $\|x^*_{\lambda,z}\| \leq c_{x,1}$ and $\|x_0\| \leq c_{x,2}$. Applying Lemma A.4 yields $\|x_1\| \leq c_{x,2}$.

Therefore, (A.42) holds for $t = 1$. We then prove by induction that (A.42) also holds for $t \in [T]$. Assume (A.42) holds at $1 \leq k \leq T-1$, i.e.,

$$\|x_k\| \leq c_x = c_{x,2} \tag{A.43}$$

Then by Lemma A.4, at $k+1$,

$$\|x_{k+1}\| = \|G_{\lambda_{k+1},Z_{k+1}}(x_k)\| \leq c_{x,2}. \tag{A.44}$$

Since $\|x_1\| \leq c_{x,2}$, for $t = 0, \ldots, T-1$, we have

$$\|x_{t+1}\| = \|G_{\lambda_{t+1},Z_{t+1}}(x_t)\| \leq c_{x,2}. \tag{A.45}$$

Therefore, by mathematical induction, $\|x_t\| \leq c_{x,2} = c_x$, for all $t \in [T]$. The proof of (a) is thus complete.

We then prove (b). This result follows directly from (a), Assumption 1, i.e., the $\ell_{f,1}$-smoothness assumption for all objectives, and boundedness of the Pareto optimal solutions given in Lemma A.3. Specifically, by Lemma A.3, there exist finite positive constant c_{x^*} such that $\|x^*_{\lambda,z}\| \leq c_{x^*}$. Then by Assumption 1, the $\ell_{f,1}$-Lipschitz continuity of the gradient $\nabla F_z(x)\lambda$, we have

$$\begin{aligned}\|\nabla F_z(x)\lambda\| =& \|\nabla F_z(x)\lambda - \nabla F_z(x^*_{\lambda,z})\lambda\| \\ \leq& \ell_{f,1}\|x - x^*_{\lambda,z}\| \leq \ell_{f,1}(\|x\| + \|x^*_{\lambda,z}\|) \leq \ell_{f,1}(c_x + c_{x^*})\end{aligned} \tag{A.46}$$

where the first equality uses the fact that $\nabla F_z(x^*_{\lambda,z})\lambda = 0$. Define $\ell_f \coloneqq \ell_{f,1}(c_x + c_{x^*})$, and $\ell_F \coloneqq \sqrt{M}\ell_f$, and then it holds for all $\lambda \in \Delta^M$ that

$$\|\nabla F(x_t)\lambda\| \leq \ell_f \quad \text{and} \quad \|\nabla F(x_t)\| \leq \|\nabla F(x_t)\|_{\mathrm{F}} \leq \sqrt{M}\ell_f = \ell_F. \tag{A.47}$$

The proof of (b) is thus complete. ■

A.5 Proof of Theorem 3.2 – Generalization of MoDo in the SC case

Organization of proof. In Section A.5.1, we introduce the properties of the MoDo update. Building upon these properties, in Section A.5.2, we prove the upper bound of argument stability in Theorem A.1. To show the tightness of the upper bound, in Section A.5.3, Theorem A.2, we provide a matching lower bound of the argument stability. Combining the upper bound in Section A.5.2 and the lower bound in Section A.5.3 leads to the results in Theorem 3.2, whose proof is in Section A.5.4.

A.5.1 Expansiveness and boundedness of MoDo update

In this section, we list the properties of the update function of MoDo at each iteration, including boundedness and approximate expansiveness, whose proofs can be found at (Chen et al., 2023a, Appendix B.5.1). These properties are then used to derive the algorithm stability. For $z, z_1, z_2 \in S$, $\lambda \in \Delta^M$, recall that the update functions of MoDo is

$$G_{x,z_1,z_2}(\lambda) = \Pi_{\Delta^M}\Big(\lambda - \gamma(\nabla F_{z_1}(x)^\top \nabla F_{z_2}(x) + \rho \mathrm{I})\lambda\Big)$$
$$G_{\lambda,z}(x) = x - \alpha \nabla F_z(x)\lambda.$$

Lemma A.5 (Boundedness of MoDo update). *Let ℓ_f be a positive constant. If $\|\nabla F_z(x)\lambda\| \le \ell_f$ for all $\lambda \in \Delta^M$, $z \in S$ and $x \in \{x_t\}_{t=1}^T$ generated by the MoDo algorithm with step size $\alpha_t \le \alpha$, then $G_{\lambda,z}(x)$ is $(\alpha\ell_f)$-bounded on the trajectory of MoDo, i.e.,*

$$\sup_{x \in \{x_t\}_{t=1}^T} \|G_{\lambda,z}(x) - x\| \le \alpha \ell_f. \tag{A.48}$$

Lemma A.6 (Properties of MoDo update in SC case). *Suppose Assumptions 1, 2 hold. Let ℓ_f be a positive constant. If for all $\lambda, \lambda' \in \Delta^M$, $z \in S$, and $x \in \{x_t\}_{t=1}^T$, $x' \in \{x'_t\}_{t=1}^T$ generated by the MoDo algorithm on datasets S and S', respectively, we have $\|\nabla F_z(x)\lambda\| \le \ell_f$, $\|\nabla F_z(x')\lambda'\| \le \ell_f$, and $\|\nabla F_z(x)\| \le \ell_F$, $\|\nabla F_z(x')\| \le \ell_F$, and step sizes of MoDo satisfy $\alpha_t \le \alpha$, $\gamma_t \le \gamma$, it holds that*

$$\begin{aligned}\|G_{\lambda,z}(x) - G_{\lambda',z}(x')\|^2 \le& (1 - 2\alpha\mu + 2\alpha^2\ell_{f,1}^2)\|x - x'\|^2 \\ &+ 2\alpha\ell_F\|x - x'\|\|\lambda - \lambda'\| + 2\alpha^2\ell_F^2\|\lambda - \lambda'\|^2 \qquad \text{(A.49)}\\ \|G_{x,z_1,z_2}(\lambda) - G_{x',z_1,z_2}(\lambda')\|^2 \le& \Big((1 + \ell_F^2\gamma)^2 + (1 + \ell_F^2\gamma)\ell_{g,1}\gamma\Big)\|\lambda - \lambda'\|^2 \\ &+ \Big((1 + \ell_F^2\gamma)\ell_{g,1}\gamma + \ell_{g,1}^2\gamma^2\Big)\|x - x'\|^2. \qquad \text{(A.50)}\end{aligned}$$

Lemma A.7 (Growth recursion with approximate expansiveness). *Fix an arbitrary sequence of updates $G_1, \ldots, G_T$ and another sequence $G'_1, \ldots, G'_T$. Let $x_0 = x'_0$ be a starting point in Ω and define $\delta_t = \|x'_t - x_t\|$ where x_t, x'_t are defined recursively through*

$$x_{t+1} = G_t(x_t), \quad x'_{t+1} = G'_t(x'_t) \quad (t > 0).$$

Let $\eta_t > 0, \nu_t \ge 0$, and $\varsigma_t \ge 0$. Then, for any $p > 0$, and $t \in [T]$, we have the recurrence relation (with $\delta_0 = 0$)

$$\delta_{t+1}^2 \le \begin{cases} \eta_t\delta_t^2 + \nu_t, & G_t = G'_t \text{ is } (\eta_t, \nu_t)\text{-approximately expansive in square;} \\ (1+p)\min\{\eta_t\delta_t^2 + \nu_t, \delta_t^2\} + (1 + \frac{1}{p})4\varsigma_t^2 & G_t \text{ and } G'_t \text{ are } \varsigma_t\text{-bounded,} \\ & G_t \text{ is } (\eta_t, \nu_t)\text{-approximately expansive in square.} \end{cases}$$

A.5.2 Upper bound of MOL uniform stability

In Theorem A.1 we bound the argument stability, which is then used to derive the MOL uniform stability and PS generalization error in Theorem 3.2.

Theorem A.1 (Argument stability of MoDo in the SC case). *Suppose Assumptions 1, 2, hold. Let A be the MoDo algorithm in Algorithm 2. Choose the step sizes $\alpha_t \le \alpha \le \min\{1/(2\ell_{f,1}), \mu/(2\ell_{f,1}^2)\}$, and $\gamma_t \le \gamma \le \min\Big\{\frac{\mu^2}{120\ell_f^2\ell_{g,1}}, \frac{1}{8(3\ell_f^2+2\ell_{g,1})}\Big\}/T$. Then it holds for all $t \in [T]$ that*

$$\mathbb{E}_A[\|A_t(S) - A_t(S')\|^2] \le \frac{48}{\mu n}\ell_f^2\Big(\alpha + \frac{12 + 4M\ell_f^2}{\mu n} + \frac{10M\ell_f^4\gamma}{\mu}\Big). \tag{A.51}$$

Proof. [Theorem A.1| Under Assumptions 1, 2, Lemma 3.1 implies that for $\{x_t\}$ generated by the MoDo algorithm, and for all $\lambda \in \Delta^M$, and for all $m \in [M]$,

$$\|\nabla F_z(x_t)\lambda\| \le \ell_{f,1}(c_x + c_{x^*}) = \ell_f. \quad \text{and} \quad \|\nabla F_z(x_t)\| \le \|\nabla F_z(x_t)\|_{\mathrm{F}} \le \sqrt{M}\ell_f = \ell_F. \tag{A.52}$$

For notation simplicity, denote $\delta_t = \|x_t - x_t'\|$, $\zeta_t = \|\lambda_t - \lambda_t'\|$, $x_T = A_T(S)$ and $x_T' = A_T(S')$. Denote the index of the different sample in S and S' as i^*, and the set of indices selected at the t-th iteration as I_t, i.e., $I_t = \{i_{t,s}\}_{s=1}^3$. When $i^* \notin I_t$, for any $c_1 > 0$, based on Lemma A.6,

$$\begin{aligned}
\delta_{t+1}^2 \le& (1 - 2\alpha_t\mu + 2\alpha_t^2\ell_{f,1}^2)\delta_t^2 + 2\alpha_t\ell_F\delta_t\zeta_{t+1} + 2\alpha_t^2\ell_F^2\zeta_{t+1}^2 \\
\le& (1 - 2\alpha_t\mu + 2\alpha_t^2\ell_{f,1}^2)\delta_t^2 + \alpha_t\ell_F(c_1\delta_t^2 + c_1^{-1}\zeta_{t+1}^2) + 2\alpha_t^2\ell_F^2\zeta_{t+1}^2 \\
\le& (1 - \alpha_t\mu)\delta_t^2 + \alpha_t\ell_F(c_1\delta_t^2 + c_1^{-1}\zeta_{t+1}^2) + 2\alpha_t^2\ell_F^2\zeta_{t+1}^2
\end{aligned} \tag{A.53}$$

where the second last inequality is due to Young's inequality; the last inequality is due to choosing $\alpha_t \le \mu/(2\ell_{f,1}^2)$.

When $i^* \in I_t$, from Lemma A.5, the $(\alpha_t\ell_f)$-boundedness of the update at t-th iteration, and Lemma A.7, the growth recursion, for a given constant $p > 0$, we have

$$\delta_{t+1}^2 \le (1+p)\delta_t^2 + (1 + 1/p)4\alpha_t^2\ell_f^2. \tag{A.54}$$

Taking expectation of δ_{t+1}^2 over I_t, we have

$$\begin{aligned}
\mathbb{E}_{I_t}[\delta_{t+1}^2] \le& \mathbb{P}(i^* \notin I_t)\Big((1 - \alpha_t\mu)\delta_t^2 + \alpha_t\ell_F c_1\delta_t^2 + (\alpha_t\ell_F c_1^{-1} + 2\alpha_t^2\ell_F^2)\mathbb{E}_{I_t}[\zeta_{t+1}^2 \mid i^* \notin I_t]\Big) \\
&+ \mathbb{P}(i^* \in I_t)\Big((1+p)\delta_t^2 + (1+1/p)4\alpha_t^2\ell_f^2\Big) \\
\le& \Big(1 - \alpha_t(\mu - \ell_F c_1)\mathbb{P}(i^* \notin I_t) + p\mathbb{P}(i^* \in I_t)\Big)\delta_t^2 \\
&+ \alpha_t \underbrace{(\ell_F c_1^{-1} + 2\alpha\ell_F^2)}_{c_2}\mathbb{E}_{I_t}[\zeta_{t+1}^2 \mid i^* \notin I_t]\mathbb{P}(i^* \notin I_t) + \Big(1 + \frac{1}{p}\Big)\mathbb{P}(i^* \in I_t)4\alpha_t^2\ell_f^2.
\end{aligned} \tag{A.55}$$

At each iteration of MoDo, we randomly select three independent samples (instead of one) from the training set S. Then the probability of selecting the different sample from S and S' at the t-th iteration, $\mathbb{P}(i^* \in I_t)$ in the above equation, can be computed as follows

$$\mathbb{P}(i^* \in I_t) = 1 - \Big(\frac{n-1}{n}\Big)^2 \leq \frac{2}{n}. \tag{A.56}$$

Consequently, the probability of selecting the same sample from S and S' at the t-th iteration is $\mathbb{P}(i^* \notin I_t) = 1 - \mathbb{P}(i^* \in I_t)$.

Let $\ell_{g,1} = \ell_f \ell_{F,1} + \ell_F \ell_{f,1}$. Recalling when $i^* \notin I_t$, $\zeta_{t+1} \leq (1 + \ell_F^2 \gamma_t)\zeta_t + 2\gamma_t \ell_{g,1} \delta_t$ from Lemma A.6, it follows that

$$\begin{aligned}\zeta_{t+1}^2 \leq& \Big((1+\ell_F^2\gamma_t)^2 + (1+\ell_F^2\gamma_t)\ell_{g,1}\gamma_t\Big)\zeta_t^2 + \Big((1+\ell_F^2\gamma_t)\ell_{g,1}\gamma_t + \ell_{g,1}^2\gamma_t^2\Big)\delta_t^2 \\ \leq& \big(1 + \underbrace{(3\ell_F^2 + 2\ell_{g,1})}_{c_3}\gamma_t\big)\zeta_t^2 + 3\ell_{g,1}\gamma_t\delta_t^2 \end{aligned} \tag{A.57}$$

where the last inequality follows from $\ell_{g,1}\gamma_t \leq 1$, and $\ell_F^2\gamma_t \leq 1$.
And since ζ_t and δ_t are independent of I_t, it follows that

$$\mathbb{E}_{I_t}[\zeta_{t+1}^2 \mid i^* \notin I_t] \leq \big(1 + c_3\gamma_t\big)\zeta_t^2 + 3\ell_{g,1}\gamma_t\delta_t^2. \tag{A.58}$$

Combining (A.55) and (A.58), we have

$$\begin{aligned}\mathbb{E}_{I_t}[\delta_{t+1}^2] \leq& \Big(1 - \alpha_t(\mu - \ell_F c_1)\mathbb{P}(i^* \notin I_t) + p\mathbb{P}(i^* \in I_t)\Big)\delta_t^2 + \Big(1 + \frac{1}{p}\Big)\mathbb{P}(i^* \in I_t)4\alpha_t^2\ell_f^2 \\ &+ \alpha_t c_2\Big(\big(1 + c_3\gamma_t\big)\zeta_t^2 + 3\ell_{g,1}\gamma_t\delta_t^2\Big)\mathbb{P}(i^* \notin I_t) \\ =& \Big(\eta_t + p\mathbb{P}(i^* \in I_t)\Big)\delta_t^2 + \alpha_t c_2\big(1 + c_3\gamma_t\big)\zeta_t^2\mathbb{P}(i^* \notin I_t) + \Big(1 + \frac{1}{p}\Big)\mathbb{P}(i^* \in I_t)4\alpha_t^2\ell_f^2 \end{aligned} \tag{A.59}$$

where we define $\eta_t = 1 - \alpha_t(\mu - \ell_F c_1 - 3c_2\ell_{g,1}\gamma_t)\mathbb{P}(i^* \notin I_t)$.
While when $i^* \in I_t$, for a given constant $p_2 > 0$, we have

$$\begin{aligned}\zeta_{t+1} =& \|\Pi_{\Delta^M}(\lambda_t - \gamma_t h_{t,1}(x_t)^\top h_{t,2}(x_t)\lambda_t) - \Pi_{\Delta^M}(\lambda_t' - \gamma_t h_{t,1}'(x_t')^\top h_{t,2}'(x_t')\lambda_t')\| \\ \leq& \|\lambda_t - \lambda_t' - \gamma_t(h_{t,1}(x_t)^\top h_{t,2}(x_t)\lambda_t - h_{t,1}'(x_t')^\top h_{t,2}'(x_t')\lambda_t')\| \\ \leq& \|\lambda_t - \lambda_t'\| + 2\gamma_t\ell_F\ell_f \leq \zeta_t + 2\gamma_t\sqrt{M}\ell_f^2 \\ \zeta_{t+1}^2 \leq& (1+p_2)\zeta_t^2 + (1+1/p_2)4\gamma_t^2 M\ell_f^4. \end{aligned} \tag{A.60}$$

Taking expectation of ζ_{t+1}^2 over I_t gives

$$\begin{aligned}\mathbb{E}_{I_t}[\zeta_{t+1}^2] =& \mathbb{E}_{I_t}[\zeta_{t+1}^2 \mid i^* \in I_t]\mathbb{P}(i^* \in I_t) + \mathbb{E}_{I_t}[\zeta_{t+1}^2 \mid i^* \notin I_t]\mathbb{P}(i^* \notin I_t) \\ \leq& \Big((1+p_2)\zeta_t^2 + (1+1/p_2)4\gamma_t^2 M\ell_f^4\Big)\mathbb{P}(i^* \in I_t) + \Big((1+c_3\gamma_t)\zeta_t^2 + 3\ell_{g,1}\gamma_t\delta_t^2\Big)\mathbb{P}(i^* \notin I_t) \\ \leq& \Big(1 + c_3\gamma_t + \frac{3}{n}p_2\Big)\zeta_t^2 + (1 + \frac{1}{p_2})4\gamma_t^2 M\ell_f^4\frac{3}{n} + 3\ell_{g,1}\gamma_t\delta_t^2. \end{aligned} \tag{A.61}$$

Based on linearity of expectation and applying (A.61) recursively yields

$$\mathbb{E}[\zeta_{t+1}^2] \leq \sum_{t'=0}^{t}\Big((1+\frac{1}{p_2})4\gamma^2 M\ell_f^4\frac{3}{n} + 3\ell_{g,1}\gamma\mathbb{E}[\delta_{t'}^2]\Big)\Bigg(\prod_{k=t'+1}^{t}\Big(1 + c_3\gamma + \frac{3}{n}p_2\Big)\Bigg)$$

$$
\begin{aligned}
&= \sum_{t'=0}^{t} \Big((1+\frac{1}{p_2}) 4\gamma^2 M \ell_f^4 \frac{3}{n} + 3\ell_{g,1}\gamma \mathbb{E}[\delta_{t'}^2] \Big) \Big(1 + c_3\gamma + \frac{3}{n} p_2 \Big)^{t-t'} \\
&\overset{(a)}{\leq} \sum_{t'=0}^{t} \Big((1+\frac{8T}{n}) 4\gamma^2 M \ell_f^4 \frac{3}{n} + 3\ell_{g,1}\gamma \mathbb{E}[\delta_{t'}^2] \Big) \Big(1 + \frac{1}{2T} \Big)^{t-t'} \\
&\overset{(b)}{\leq} \sum_{t'=0}^{t} \Big((1+\frac{8T}{n}) 4\gamma^2 M \ell_f^4 \frac{3}{n} + 3\ell_{g,1}\gamma \mathbb{E}[\delta_{t'}^2] \Big) e^{\frac{1}{2}} \\
&\overset{(c)}{\leq} 2\gamma \sum_{t'=0}^{t} \Big((1+\frac{8T}{n}) 4\gamma M \ell_f^4 \frac{3}{n} + 3\ell_{g,1} \mathbb{E}[\delta_{t'}^2] \Big)
\end{aligned} \tag{A.62}
$$

where (a) follows from choosing $\gamma_t \leq \gamma \leq 1/(8c_3T)$, $p_2 = n/(8T)$, (b) follows from $t - t' \leq T$, and $(1 + \frac{a}{T})^T \leq e^a$, and the inequality (c) follows from $e^{\frac{1}{2}} < 2$. Note that $\delta_0 = 0, \zeta_1 = 0$. Applying (A.55) at $t = 0$ gives

$$
\mathbb{E}[\delta_1^2] \leq \frac{3}{n} \left(1 + \frac{1}{p} \right) 4\alpha^2 \ell_f^2
$$

which together with (A.61) gives

$$
\mathbb{E}[\zeta_2^2] \leq 3\ell_{g,1}\gamma_1 \delta_1^2 + \left(1 + \frac{1}{p_2} \right) 4\gamma_1^2 M \ell_f^4 \frac{3}{n}.
$$

Therefore, for $0 \leq t \leq 1$, it satisfies that

$$
\begin{aligned}
\mathbb{E}[\delta_t^2] \leq & \left(\frac{3}{n}(1+\frac{1}{p}) 4\alpha^2 \ell_f^2 + 24 M \ell_f^4 c_2 (\frac{8\gamma T}{n} + \gamma) \frac{\alpha}{n} \right) \underbrace{\Big(\sum_{t'=0}^{t-1} (1 - \frac{1}{2}\alpha\mu + \frac{3p}{n})^{t-t'-1} \Big)}_{\beta_t} \\
= & \left(\frac{3}{n}(1+\frac{1}{p}) 4\alpha^2 \ell_f^2 + 24 M \ell_f^4 c_2 \Big(\frac{8\gamma T}{n} + \gamma \Big) \frac{\alpha}{n} \right) \beta_t.
\end{aligned} \tag{A.63}
$$

Next, we will prove by induction that (A.63) also holds for $t > 1$. Assuming that (A.63) holds for all $0 \leq t \leq k \leq T-1$, we apply (A.59) to the case where $t = k$ to obtain

$$
\begin{aligned}
\mathbb{E}[\delta_{k+1}^2] \leq & \Big(\eta_k + \frac{3p}{n} \Big) \mathbb{E}[\delta_k^2] + \alpha_k c_2 \Big(1 + c_3\gamma_k \Big) \mathbb{E}[\zeta_k^2] \mathbb{P}(i^* \notin I_t) + \frac{3}{n} \Big(1 + \frac{1}{p} \Big) 4\alpha_k^2 \ell_f^2 \\
\overset{(a)}{\leq} & \Big(\eta_k + \frac{3p}{n} \Big) \mathbb{E}[\delta_k^2] \\
& + 2\alpha_k c_2 \gamma \left(\sum_{t'=1}^{k} \left((1+\frac{8T}{n}) \frac{12\gamma M \ell_f^4}{n} + 3\ell_{g,1} \mathbb{E}[\delta_{t'}^2] \right) \right) \mathbb{P}(i^* \notin I_t) + \frac{3}{n} \left(1 + \frac{1}{p} \right) 4\alpha_k^2 \ell_f^2 \\
\overset{(b)}{\leq} & \underbrace{\left(\Big(\eta_k + \frac{3p}{n} \Big) \beta_k + 1 + 6\alpha_k c_2 \ell_{g,1} \gamma \Big(\sum_{t'=1}^{k} \beta_{t'} \Big) \mathbb{P}(i^* \notin I_t) \right)}_{J_1}
\end{aligned}
$$

$$\times \left(\frac{3}{n}\left(1+\frac{1}{p}\right)4\alpha^2\ell_f^2 + 24M\ell_f^4 c_2\left(\frac{8\gamma T}{n}+\gamma\right)\frac{\alpha}{n}\right) \tag{A.64}$$

where (a) follows from (A.62), and (b) follows from (A.63) for $0 \le t \le k$ and that $\gamma k \le \gamma T \le 1$. The coefficient J_1 in (A.64) can be further bounded by

$$\begin{aligned} J_1 =& \Big(\eta_k + \frac{3p}{n}\Big)\beta_k + 1 + 6\alpha_k c_2 \ell_{g,1}\gamma(\sum_{t'=1}^{k} c_{t'})\mathbb{P}(i^* \notin I_t) \\ \overset{(c)}{\le}& \Big(\eta_k + \frac{3p}{n}\Big)\beta_k + 1 + 6\alpha_k c_2 \ell_{g,1} k\gamma\beta_k \mathbb{P}(i^* \notin I_t) \\ \overset{(d)}{\le}& \Big(1 - \alpha_k(\mu - \ell_F c_1 - 3c_2\ell_{g,1}\gamma(1+2k))\mathbb{P}(i^* \notin I_t) + \frac{3p}{n}\Big)\beta_k + 1 \overset{(e)}{\le} \Big(1 - \frac{1}{2}\alpha\mu + \frac{3p}{n}\Big)\beta_k + 1 \end{aligned} \tag{A.65}$$

where (c) is from $\beta_t \le \beta_{t+1}$, $\gamma_t \le \gamma$ for all $t = 0, \dots, T$; (d) is from the definition of η_k; (e) is because $\gamma \le \mu^2/(120\ell_F^2\ell_{g,1}T)$, $\alpha \le 1/(2\ell_{f,1}) \le 1/(2\mu)$ and choosing $c_1 = \mu/(4\ell_F)$ leads to

$$\begin{aligned} \ell_F c_1 + 3c_2\ell_{g,1}\gamma(1+2k)\gamma \le& \ell_F c_1 + 6(\ell_F c_1^{-1} + 2\alpha\ell_F^2)\ell_{g,1}(k+1)\gamma \\ \le& \frac{1}{4}\mu + 6(4\mu^{-1} + 2\alpha)\ell_F^2\ell_{g,1}\frac{k+1}{T}\frac{\mu^2}{120\ell_F^2\ell_{g,1}} \le \frac{1}{2}\mu. \end{aligned}$$

Combining (A.64) and (A.65) implies

$$\begin{aligned} \mathbb{E}[\delta_{k+1}^2] \le& \Big(\big(1 - \frac{1}{2}\alpha\mu + \frac{3p}{n}\big)\beta_k + 1\Big)\Big(\frac{3}{n}(1+\frac{1}{p})4\alpha^2\ell_f^2 + 24M\ell_f^4 c_2\big(\frac{8\gamma T}{n}+\gamma\big)\frac{\alpha}{n}\Big) \\ =& c_{k+1}\Big(\frac{3}{n}(1+\frac{1}{p})4\alpha^2\ell_f^2 + 24M\ell_f^4 c_2\big(\frac{8\gamma T}{n}+\gamma\big)\frac{\alpha}{n}\Big) \end{aligned} \tag{A.66}$$

where the equality follows by the definition of β_t given in (A.63). The above statements from (A.64)-(A.66) show that if (A.63) holds for all t such that $0 \le t \le k \le T-1$, it also holds for $t = k+1$. Therefore, we can conclude that for all $t \in [T]$, it follows

$$\begin{aligned} \mathbb{E}[\delta_t^2] \le& \beta_T\Big(\frac{3}{n}(1+\frac{1}{p})4\alpha^2\ell_f^2 + 24M\ell_f^4 c_2\big(\frac{8\gamma T}{n}+\gamma\big)\frac{\alpha}{n}\Big) \\ =& \Big(\frac{3}{n}(1+\frac{1}{p})4\alpha^2\ell_f^2 + 24M\ell_f^4 c_2\big(\frac{8\gamma T}{n}+\gamma\big)\frac{\alpha}{n}\Big)\left(\sum_{k=0}^{T-1}\Big(1 - \frac{1}{2}\alpha\mu + \frac{3p}{n}\Big)^{T-k-1}\right) \\ =& \Big(\frac{3}{n}(1+\frac{12}{\alpha\mu n})4\alpha^2\ell_f^2 + 24M\ell_f^4 c_2\big(\frac{8\gamma T}{n}+\gamma\big)\frac{\alpha}{n}\Big)\Big(\frac{1}{4}\alpha\mu\Big)^{-1}\left(1 - \Big(1 - \frac{1}{4}\alpha\mu\Big)^T\right) \end{aligned} \tag{A.67}$$

where the first inequality follows from $\beta_t \le \beta_T$ for all $t \in [T]$; the last equality follows from taking $p = \alpha\mu n/12$, and computing the sum of geometric series. By plugging in $c_1 = \mu/(4\ell_F)$, $c_2 = \ell_F c_1^{-1} + 2\alpha\ell_F^2$, $c_3 = 3\ell_F^2 + 2\ell_{g,1}$, for all $t \in [T]$,we have that

$$\begin{aligned} \mathbb{E}[\delta_t^2] \le& \Big(\frac{3}{n}(1+\frac{12}{\alpha\mu n})4\alpha^2\ell_f^2 + 24M\ell_f^4 c_2 c_3^{-1}\frac{\alpha}{n^2} + 24M\ell_f^4 c_2\frac{\alpha\gamma}{n}\Big)(\frac{1}{4}\alpha\mu)^{-1} \\ \le& \frac{48}{\mu n}\ell_f^2\Big(\alpha + \frac{12}{\mu n} + \frac{2M\ell_f^2 c_2 c_3^{-1}}{n} + 2M\ell_f^2 c_2\gamma\Big) \le \frac{48}{\mu n}\ell_f^2\Big(\alpha + \frac{12 + 4M\ell_f^2}{\mu n} + \frac{10M\ell_f^4\gamma}{\mu}\Big) \end{aligned} \tag{A.68}$$

where the last inequality follows from $c_2 = \ell_F^2(4\mu^{-1} + 2\alpha) \leq 5M\ell_f^2\mu^{-1}$, and $c_2 c_3^{-1} \leq 5\ell_F^2\mu^{-1}/(3\ell_F^2) \leq 2\mu^{-1}$. ■

A.5.3 Lower bound of MOL uniform stability

This section provides a lower bound of the MOL uniform stability in the SC case. The full proof is available at (Chen et al., 2023a, Section B.5.3).

Theorem A.2. *(Chen et al., 2023a, Theorem B.2) Suppose Assumptions 1 and 2 hold. Under Example 1 in (Chen et al., 2023a) with $M = 2$, choose $\lambda_0 = \frac{1}{M}\mathbf{1}$, $x_0 = x_0' = 7v$, $\alpha = \frac{1}{4\mu T}$, $0 < \gamma \leq \frac{1}{2MT\ell_F\ell_f}$, $\rho = 0$, and $T \leq 4n^{\frac{2}{3}}$ for the MoDo algorithm. Denote $\{x_t\}$, $\{\lambda_t\}$ and $\{x_t'\}$, $\{\lambda_t'\}$ as the sequences generated by the MoDo algorithm with dataset S and S', respectively. Then it holds that*

$$\mathbb{E}[\|x_T - x_T'\|] \geq \frac{\gamma T}{2n^2} + \frac{1}{16n}. \tag{A.69}$$

A.5.4 Proof of Theorem 3.2

Proof. [Theorem 3.2| Combining the argument stability in Theorem A.1, and Assumption 1, the MOL uniform stability can be bounded by

$$\begin{aligned}\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2] \leq & \mathbb{E}_A[\ell_{F,1}^2\|A(S) - A(S')\|^2] && \text{by Assumption 1}\\ \leq & \frac{48}{\mu n}\ell_f^2\ell_{F,1}^2\Big(\alpha + \frac{12 + 4M\ell_f^2}{\mu n} + \frac{10M\ell_f^4\gamma}{\mu}\Big). && \text{(A.70)}\end{aligned}$$

Then based on Propositions 3.1-3.2, we have

$$\begin{aligned}\mathbb{E}_{A,S}[R_{\mathrm{gen}}(A(S))] \leq & \mathbb{E}_{A,S}[\|\nabla F(A(S)) - \nabla F_S(A(S))\|_{\mathrm{F}}] && \text{by Proposition 3.1}\\ \leq & 4\epsilon_{\mathrm{F}} + \sqrt{n^{-1}\mathbb{E}_S\left[\mathbb{V}_{z\sim\mathcal{D}}(\nabla F_z(A(S)))\right]} && \text{by Proposition 3.2}\\ = & \mathcal{O}(n^{-\frac{1}{2}}). && \text{by (A.70)}\end{aligned}$$

The proof of the upper bound is complete. We then prove the MOL uniform stability lower bound based on the argument uniform stability lower bound in Theorem A.2. By the strong convexity of the function $f_{z,m}(x)$, for all $m \in [M]$

$$\begin{aligned}&\sup_z \mathbb{E}_A[\|\nabla F_z(A(S)) - \nabla F_z(A(S'))\|_{\mathrm{F}}^2] \geq \mathbb{E}_A[M\mu^2\|A(S) - A(S')\|^2] && \text{by Assumption 2}\\ \geq & M\mu^2\Big(\frac{\gamma T}{2n^2} + \frac{1}{16n}\Big)^2 && \text{by Theorem A.2 and Jensen's inequality}\\ \geq & \frac{M\mu^2}{64n^2}. && \text{by choosing } n \geq \tfrac{4}{M\ell_F\ell_f} \geq 8\gamma T\end{aligned}$$

The proof of the lower bound is complete. ■

Appendix B. Bounding the CA Distance

B.1 Auxiliary lemmas

This section summarizes properties of the generalized subproblem $\min_{\lambda\in\Delta^M} \|Q\lambda\|^2 + \rho\|\lambda\|^2$ with $\rho \geq 0$ and properties of the update function of the MoDo algorithm, where Q can be the full-batch gradient $\nabla F_S(x)$ or its stochastic estimate. Proofs of these auxiliary lemmas can be found in (Chen et al., 2023a, Appendix C.1). Before proceeding, we first define a few notations we will use repeatedly in this section.

$$\text{The CA weight} \qquad \lambda^*_{Q,\rho} \in \arg\min_{\lambda\in\Delta^M} \|Q\lambda\|^2 + \rho\|\lambda\|^2 \tag{B.1a}$$

$$\text{The CA direction} \qquad d_{Q,\rho} \coloneqq Q\lambda^*_{Q,\rho} \tag{B.1b}$$

Lemma B.1 (Uniqueness of CA direction $d_{Q,\rho}$). *Given $Q \in \mathbb{R}^{d\times M}$, $\rho \geq 0$, then $d_{Q,\rho} \coloneqq Q\lambda^*_{Q,\rho}$ with $\lambda^*_{Q,\rho} \in \arg\min_{\lambda\in\Delta^M} \|Q\lambda\|^2 + \rho\|\lambda\|^2$ exists, and $d_{Q,\rho}$ is unique.*

Proof. When $\rho = 0$, the proof is given in (Désidéri, 2012, Section 2). When $\rho > 0$, it is a standard result for strictly convex problems with a unique $\lambda^*_{Q,\rho}$, thus unique $d_{Q,\rho}$. ■

Lemma B.2. *Given $Q \in \mathbb{R}^{d\times M}$, recall $\lambda^*_{Q,\rho}$ with $\rho \geq 0$ is defined as*

$$\lambda^*_{Q,\rho} \in \arg\min_{\lambda\in\Delta^M} \|Q\lambda\|^2 + \rho\|\lambda\|^2. \tag{B.2}$$

Then, for any $\lambda \in \Delta^M$, it holds that

$$\langle Q\lambda^*_{Q,\rho}, Q\lambda\rangle \geq \|Q\lambda^*_{Q,\rho}\|^2 - \rho, \tag{B.3a}$$

$$\text{and} \quad \|Q\lambda - Q\lambda^*_{Q,\rho}\|^2 \leq \|Q\lambda\|^2 - \|Q\lambda^*_{Q,\rho}\|^2 + 2\rho. \tag{B.3b}$$

Lemma B.3 (Continuity of $\lambda^*_{Q,\rho}$ with $\rho > 0$). *Given $Q \in \mathbb{R}^{d\times M}$, $\rho > 0$ and $x \in \mathbb{R}^d$, for $\lambda^*_{Q,\rho}$ defined in* (B.2)*, the following inequality holds*

$$\|\lambda^*_{Q,\rho} - \lambda^*_{Q',\rho}\| \leq \rho^{-1}\|Q^\top Q - Q'^\top Q'\|. \tag{B.4}$$

Furthermore, suppose either 1) Assumptions 1, 3 hold, or 2) Assumptions 1, 2 hold, with ℓ_F defined in Lemma 3.1. Then for $x \in \{x_t\}_{t=1}^T$, $x' \in \{x'_t\}_{t=1}^T$ generated by MoDo algorithm on training dataset S and S', respectively, let $\lambda^*_\rho(x) = \lambda^*_{\nabla F_S(x),\rho}$, $\lambda^*_\rho(x') = \lambda^*_{\nabla F_S(x'),\rho}$, it implies

$$\|\lambda^*_\rho(x) - \lambda^*_\rho(x')\| \leq 2\rho^{-1}\ell_{F,1}\ell_F\|x - x'\|. \tag{B.5}$$

Lemma B.4. *Given $Q \in \mathbb{R}^{d\times M}$, $\rho \geq 0, \bar\rho > 0$, with $\lambda^*_{Q,\rho}$ defined in* (B.1a)*, then we have*

$$-\bar\rho\Big(1 - \frac{1}{M}\Big) \leq \|Q\lambda^*_{Q,\rho}\|^2 - \|Q\lambda^*_{Q,\bar\rho}\|^2 \leq \rho\left(1 - \frac{1}{M}\right). \tag{B.6}$$

Lemma B.5 (Properties of MoDo update of λ_t). *Consider $\{x_t\}$, $\{\lambda_t\}$ generated by the MoDo algorithm. For all $\lambda \in \Delta^M$, $\rho \geq 0$, it holds that*

$$\begin{aligned}
&2\gamma_t \mathbb{E}_A \langle \lambda_t - \lambda, (\nabla F_S(x_t)^\top \nabla F_S(x_t) + \rho \mathrm{I})\lambda_t \rangle \\
\leq &\mathbb{E}_A[\|\lambda_t - \lambda\|^2] - \mathbb{E}_A[\|\lambda_{t+1} - \lambda\|^2] + \gamma_t^2 \mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2], \quad \text{(B.7)} \\
\textit{and} \quad &\gamma_t \mathbb{E}_A(\|\nabla F_S(x_t)\lambda_t\|^2 - \|\nabla F_S(x_t)\lambda\|^2 + \rho\|\lambda_t\|^2 - \rho\|\lambda\|^2 + \rho\|\lambda_t - \lambda\|^2) \\
\leq &\mathbb{E}_A[\|\lambda_t - \lambda\|^2] - \mathbb{E}_A[\|\lambda_{t+1} - \lambda\|^2] + \gamma_t^2 \mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2]. \quad \text{(B.8)}
\end{aligned}$$

B.2 Proof of Theorem 3.3 – CA direction distance of MoDo

Organization of proof. In Lemma B.6, we prove the upper bound of the CA direction distance in terms of two averages of sequences, $S_{1,T}$, and $S_{2,T}$. Then under either Assumptions 1, 3, or Assumptions 1, 2, we prove the upper bound of $S_{1,T}$, and $S_{2,T}$, and thus the CA direction distance in Theorem 3.3.

Lemma B.6. *Suppose Assumption 1 holds. Let $\{x_t\}, \{\lambda_t\}$ be the sequences produced by the MoDo algorithm with step sizes $\alpha_t = \alpha > 0$, $\gamma_t = \gamma > 0$, and regularization $\rho \geq 0$. With a positive constant $\bar{\rho} > 0$, define*

$$S_{1,T} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2] \tag{B.9a}$$

$$S_{2,T} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_A[\|\nabla F_S(x_{t+1}) + \nabla F_S(x_t)\|\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|]. \tag{B.9b}$$

Then it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2 - \|\nabla F_S(x_t)\lambda_\rho^*(x_t)\|^2] \leq \bar{\rho} + \frac{4}{\gamma T}(1 + \bar{\rho}^{-1}\alpha \ell_{F,1} T S_{2,T}) + \frac{\rho}{\gamma} + \gamma S_{1,T}. \tag{B.10}$$

Proof. Define $\lambda_{\bar{\rho}}^*(x_t) = \arg\min_{\lambda \in \Delta^M} \frac{1}{2}\|\nabla F_S(x_t)\lambda\|^2 + \frac{\bar{\rho}}{2}\|\lambda\|^2$ with $\bar{\rho} > 0$. Note that different from $\rho \geq 0$, $\bar{\rho} > 0$ is strictly positive, and used as an intermediate parameter only for analysis but not for algorithm update.

Substituting $\lambda = \lambda_{\bar{\rho}}^*(x_t)$ in Lemma B.5, (B.8), we have

$$\begin{aligned}
&\gamma_t \mathbb{E}_A(\|\nabla F_S(x_t)\lambda_t\|^2 - \|\nabla F_S(x_t)\lambda_{\bar{\rho}}^*(x_t)\|^2 + \rho\|\lambda_t\|^2 - \rho\|\lambda_{\bar{\rho}}^*(x_t)\|^2 + \rho\|\lambda_t - \lambda_{\bar{\rho}}^*(x_t)\|^2) \\
\leq &\mathbb{E}_A[\|\lambda_t - \lambda_{\bar{\rho}}^*(x_t)\|^2] - \mathbb{E}_A[\|\lambda_{t+1} - \lambda_{\bar{\rho}}^*(x_t)\|^2] + \gamma_t^2 \mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2].
\end{aligned} \tag{B.11}$$

Setting $\gamma_t = \gamma > 0$, and telescoping the above inequality gives

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2 - \|\nabla F_S(x_t)\lambda_{\bar{\rho}}^*(x_t)\|^2] \\
\leq &\frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{\gamma}\mathbb{E}_A[\|\lambda_t - \lambda_{\bar{\rho}}^*(x_t)\|^2 - \|\lambda_{t+1} - \lambda_{\bar{\rho}}^*(x_t)\|^2] + \frac{1}{T}\sum_{t=0}^{T-1} \gamma \mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2]
\end{aligned}$$

$$
\begin{aligned}
&-\frac{\rho}{\gamma}\mathbb{E}_A[\|\lambda_t\|^2-\|\lambda^*_{\bar\rho}(x_t)\|^2+\|\lambda_t-\lambda^*_{\bar\rho}(x_t)\|^2]\\
=&\frac{1}{\gamma T}\underbrace{\Big(\sum_{t=0}^{T-1}\mathbb{E}_A[\|\lambda_t-\lambda^*_{\bar\rho}(x_t)\|^2-\|\lambda_{t+1}-\lambda^*_{\bar\rho}(x_t)\|^2]\Big)}_{I_1}+\frac{1}{T}\sum_{t=0}^{T-1}\gamma\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t)+\rho\mathrm{I})\lambda_t\|^2]\\
&-\frac{\rho}{\gamma}\mathbb{E}_A[\|\lambda_t\|^2-\|\lambda^*_{\bar\rho}(x_t)\|^2+\|\lambda_t-\lambda^*_{\bar\rho}(x_t)\|^2] \qquad \text{(B.12)}
\end{aligned}
$$

where I_1 can be further derived as

$$
\begin{aligned}
I_1=&\sum_{t=0}^{T-1}\mathbb{E}_A[\|\lambda_t-\lambda^*_{\bar\rho}(x_t)\|^2]-\mathbb{E}_A[\|\lambda_{t+1}-\lambda^*_{\bar\rho}(x_t)\|^2]\\
=&\mathbb{E}_A[\|\lambda_0-\lambda^*_{\bar\rho}(x_0)\|^2]-\mathbb{E}_A[\|\lambda_T-\lambda^*_{\bar\rho}(x_{T-1})\|^2]+\sum_{t=0}^{T-2}\mathbb{E}_A[\|\lambda_{t+1}-\lambda^*_{\bar\rho}(x_{t+1})\|^2-\|\lambda_{t+1}-\lambda^*_{\bar\rho}(x_t)\|^2]\\
\le&\mathbb{E}_A[\|\lambda_0-\lambda^*_{\bar\rho}(x_0)\|^2]-\mathbb{E}_A[\|\lambda_T-\lambda^*_{\bar\rho}(x_{T-1})\|^2]\\
&+\sum_{t=0}^{T-2}\mathbb{E}_A[\|2\lambda_{t+1}-\lambda^*_{\bar\rho}(x_{t+1})-\lambda^*_{\bar\rho}(x_t)\|\|\lambda^*_{\bar\rho}(x_{t+1})-\lambda^*_{\bar\rho}(x_t)\|]\le 4+4\sum_{t=0}^{T-2}\mathbb{E}_A[\|\lambda^*_{\bar\rho}(x_{t+1})-\lambda^*_{\bar\rho}(x_t)\|]
\end{aligned}
\qquad \text{(B.13)}
$$

where $\|\lambda^*_{\bar\rho,t+1}(x_{t+1})-\lambda^*_{\bar\rho}(x_t)\|$, by Lemma B.3, can be bounded by

$$
\begin{aligned}
\|\lambda^*_{\bar\rho,t+1}(x_{t+1})-\lambda^*_{\bar\rho}(x_t)\|\le&\bar\rho^{-1}\|\nabla F_S(x_{t+1})+\nabla F_S(x_t)\|\|\nabla F_S(x_{t+1})-\nabla F_S(x_t)\|\\
\le&\bar\rho^{-1}\ell_{F,1}\|\nabla F_S(x_{t+1})+\nabla F_S(x_t)\|\|x_{t+1}-x_t\|\\
\le&\bar\rho^{-1}\alpha\ell_{F,1}\|\nabla F_S(x_{t+1})+\nabla F_S(x_t)\|\|\nabla F_{Z_{t+1}}\lambda_{t+1}\|. \qquad \text{(B.14)}
\end{aligned}
$$

Hence, it follows that

$$
I_1\le 4+4\bar\rho^{-1}\alpha\ell_{F,1}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_{t+1})+\nabla F_S(x_t)\|\|\nabla F_{Z_{t+1}}\lambda_{t+1}\|]=4+4\bar\rho^{-1}\alpha\ell_{F,1}TS_{2,T} \qquad \text{(B.15)}
$$

plugging which into (B.12) gives

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2-\|\nabla F_S(x_t)\lambda^*_{\bar\rho}(x_t)\|^2]\le\frac{4}{\gamma T}(1+\bar\rho^{-1}\alpha\ell_{F,1}TS_{2,T})+\gamma S_{1,T}+\frac{\rho}{\gamma}. \qquad \text{(B.16)}
$$

Recall $\lambda^*_\rho(x_t)=\arg\min_{\lambda\in\Delta^M}\|\nabla F_S(x_t)\lambda\|^2+\rho\|\lambda\|^2$. Then

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2-\|\nabla F_S(x_t)\lambda^*_\rho(x_t)\|^2]\\
=&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2-\|\nabla F_S(x_t)\lambda^*_{\bar\rho}(x_t)\|^2+\|\nabla F_S(x_t)\lambda^*_{\bar\rho}(x_t)\|^2-\|\nabla F_S(x_t)\lambda^*_\rho(x_t)\|^2]
\end{aligned}
$$

$$\overset{\text{(B.16)}}{\leq} \frac{4}{\gamma T}(1+\bar{\rho}^{-1}\alpha\ell_{F,1}TS_{2,T})+\gamma S_{1,T}+\frac{\rho}{\gamma}+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda^*_{\bar{\rho}}(x_t)\|^2-\|\nabla F_S(x_t)\lambda^*_{\rho}(x_t)\|^2]$$

$$\leq \frac{4}{\gamma T}(1+\bar{\rho}^{-1}\alpha\ell_{F,1}TS_{2,T})+\gamma S_{1,T}+\frac{\rho}{\gamma}+\bar{\rho} \tag{B.17}$$

where the last inequality follows from Lemma B.4. The proof is complete. ■

Proof. [Theorem 3.3] Building on the result in Lemma B.6, and by the convexity of the subproblem, $\min_{\lambda\in\Delta^M}\frac{1}{2}\|\nabla F_S(x_t)\lambda\|^2+\rho\|\lambda\|^2$, and Lemma B.2, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t-\nabla F_S(x_t)\lambda^*_{\rho}(x_t)\|^2]\leq\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2-\|\nabla F_S(x_t)\lambda^*_{\rho}(x_t)\|^2]+2\rho$$

$$\leq\bar{\rho}+2\rho+\frac{4}{\gamma T}(1+\bar{\rho}^{-1}\alpha T\ell_{F,1}S_{2,T})+\frac{\rho}{\gamma}+\gamma S_{1,T}. \tag{B.18}$$

By Assumptions 1, 3 or Assumptions 1, 2 and Lemma 3.1, we have

$$S_{1,T}=\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t)+\rho\mathrm{I})\lambda_t\|^2]\leq(\ell_f\ell_F+\rho)^2=(M^{\frac{1}{2}}\ell_f^2+\rho)^2 \tag{B.19}$$

$$S_{2,T}=\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_{t+1})+\nabla F_S(x_t)\|\|\nabla F_{Z_{t+1}}\lambda_{t+1}\|]\leq 2\ell_f\ell_F=2M^{\frac{1}{2}}\ell_f^2. \tag{B.20}$$

Substituting $S_{1,T}$, $S_{2,T}$ in (B.18) with the above bound yields

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t-\nabla F_S(x_t)\lambda^*_{\rho}(x_t)\|^2]\\ \leq&\bar{\rho}+2\rho+\frac{4}{\gamma T}(1+2\bar{\rho}^{-1}\alpha T\ell_{F,1}\ell_f\ell_F)+\frac{\rho}{\gamma}+\gamma(M^{\frac{1}{2}}\ell_f^2+\rho)^2.\end{aligned} \tag{B.21}$$

Based on the definition of the CA direction distance, we have

$$\begin{aligned}&\mathcal{E}_{\mathrm{ca}}(x_t,\lambda_{t+1})=\|\mathbb{E}_A[\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}-d(x_t)]\|^2=\|\mathbb{E}_A[\nabla F_S(x_t)\lambda_{t+1}-\nabla F_S(x_t)\lambda^*_{\rho}(x_t)]\|^2\\ \leq&\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_{t+1}-\nabla F_S(x_t)\lambda^*_{\rho}(x_t)\|^2]\\ \leq&2\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t-\nabla F_S(x_t)\lambda^*_{\rho}(x_t)\|^2]+2\mathbb{E}_A[\|\nabla F_S(x_t)(\lambda_{t+1}-\lambda_t)\|^2]\\ \leq&2\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t-\nabla F_S(x_t)\lambda^*_{\rho}(x_t)\|^2]+2\gamma^2\ell_F^2\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t)+\rho\mathrm{I})\lambda_t\|^2].\end{aligned} \tag{B.22}$$

Because $\ell_{F,1}\ell_F\leq M\ell_{f,1}\ell_f$, choosing $\bar{\rho}=2(\alpha M\ell_{f,1}\ell_f^2/\gamma)^{\frac{1}{2}}$ yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{E}_{\mathrm{ca}}(x_t,\lambda_{t+1})\leq&\frac{1}{T}\sum_{t=0}^{T-1}2\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t-\nabla F_S(x_t)\lambda^*_{\rho}(x_t)\|^2]+2M\gamma^2\ell_f^2S_{1,T}\\ \leq&2\bar{\rho}+4\rho+\frac{8}{\gamma T}(1+2\bar{\rho}^{-1}\alpha TM\ell_{f,1}\ell_f^2)+\frac{2\rho}{\gamma}+2\gamma(1+M\gamma)(M^{\frac{1}{2}}\ell_f^2+\rho)^2\end{aligned}$$

$$=\frac{8}{\gamma T}+12\sqrt{M\ell_{f,1}\ell_f^2\frac{\alpha}{\gamma}}+\frac{2\rho}{\gamma}+4\rho+2\gamma(1+M\gamma)(M^{\frac{1}{2}}\ell_f^2+\rho)^2. \quad \text{(B.23)}$$

This proves the result. ■

B.3 Proof of Theorem 3.4 – CA weight distance of MoDo

In this section, we consider the regularization $\rho > 0$, and prove Theorem 3.4, the guarantee of CA weight distance, which is stronger than the guarantee of CA direction distance.

Proof. [Theorem 3.4| Consider the function $g(\lambda; \nabla F_S(x), \rho) \coloneqq \frac{1}{2}\|\nabla F_S(x)\lambda\|^2 + \frac{1}{2}\rho\|\lambda\|^2 \geq 0$, which is ρ-strongly convex. Based on Lemma B.5, (B.8), the property of the update of λ,

$$\begin{aligned}&\gamma_t\mathbb{E}_A[g(\lambda_t; \nabla F_S(x_t), \rho) - g(\lambda; \nabla F_S(x_t), \rho) + \rho\|\lambda_t - \lambda\|^2]\\ \leq&\mathbb{E}_A[\|\lambda_t - \lambda\|^2] - \mathbb{E}_A[\|\lambda_{t+1} - \lambda\|^2] + \gamma_t^2\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t) + \rho\mathrm{I})\lambda_t\|^2]\end{aligned}$$

where setting $\gamma_t = \gamma > 0$ and rearranging yields

$$\begin{aligned}\mathbb{E}_A[\|\lambda_{t+1} - \lambda\|^2] \leq&(1-\rho\gamma)\mathbb{E}_A[\|\lambda_t - \lambda\|^2] + \gamma^2\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t) + \rho\mathrm{I})\lambda_t\|^2]\\ &- \gamma\mathbb{E}_A[g(\lambda_t; \nabla F_S(x_t), \rho) - g(\lambda; \nabla F_S(x_t), \rho)]. \quad \text{(B.24)}\end{aligned}$$

Substituting $\lambda = \lambda_\rho^*(x_t) = \arg\min_{\lambda\in\Delta^M} g(\lambda; \nabla F_S(x_t), \rho)$, we have

$$\begin{aligned}\mathbb{E}_A[\|\lambda_{t+1} - \lambda_\rho^*(x_t)\|^2] \leq&(1-\rho\gamma)\mathbb{E}_A[\|\lambda_t - \lambda_\rho^*(x_t)\|^2] + \gamma^2\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t) + \rho\mathrm{I})\lambda_t\|^2]\\ &- \gamma\mathbb{E}_A[g(\lambda_t; \nabla F_S(x_t), \rho) - g(\lambda_\rho^*(x_t); \nabla F_S(x_t), \rho)]\\ \leq&(1-\rho\gamma)\mathbb{E}_A[\|\lambda_t - \lambda_\rho^*(x_t)\|^2] + \gamma^2(M^{\frac{1}{2}}\ell_f^2 + \rho)^2 \quad \text{(B.25)}\end{aligned}$$

where the last inequality holds because $g(\lambda_\rho^*(x_t); \nabla F_S(x_t), \rho) = \min_{\lambda\in\Delta^M} g(\lambda; \nabla F_S(x_t), \rho)$, and $\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t) + \rho\mathrm{I})\lambda_t\| \leq M^{\frac{1}{2}}\ell_f^2 + \rho$.
Then $\mathbb{E}_A[\|\lambda_t - \lambda_\rho^*(x_t)\|^2]$ can be further bounded by

$$\begin{aligned}\mathbb{E}_A[\|\lambda_t - \lambda_\rho^*(x_t)\|^2] \leq&\mathbb{E}_A[\|\lambda_t - \lambda_\rho^*(x_{t-1})\|^2] + \mathbb{E}_A(\|\lambda_t - \lambda_\rho^*(x_t)\|^2 - \|\lambda_t - \lambda_\rho^*(x_{t-1})\|^2)\\ \leq&\mathbb{E}_A\|\lambda_t - \lambda_\rho^*(x_{t-1})\|^2 + 4\mathbb{E}_A\|\lambda_\rho^*(x_t) - \lambda_\rho^*(x_{t-1})\| \quad \text{(B.26)}\end{aligned}$$

where $\|\lambda_\rho^*(x_t) - \lambda_\rho^*(x_{t-1})\|$, by Lemma B.3, can be bounded by

$$\begin{aligned}\|\lambda_\rho^*(x_t) - \lambda_\rho^*(x_{t-1})\| \leq&\rho^{-1}\|\nabla F_S(x_{t-1}) + \nabla F_S(x_t)\|\|\nabla F_S(x_{t-1}) - \nabla F_S(x_t)\|\\ \leq&\rho^{-1}\ell_{F,1}\|\nabla F_S(x_{t-1}) + \nabla F_S(x_t)\|\|x_{t-1} - x_t\|\\ \leq&2\rho^{-1}\alpha\ell_{F,1}\|\nabla F_S(x_{t-1}) + \nabla F_S(x_t)\|\|\nabla F_{Z_t}\lambda_t\| \leq 2\rho^{-1}\alpha\ell_{F,1}\ell_F\ell_f \quad \text{(B.27)}\end{aligned}$$

where the last inequality follows from either: 1) Assumptions 1, 3; or 2) Assumptions 1, 2, with ℓ_f and ℓ_F defined in Lemma 3.1.
Combining (B.25), (B.26) and (B.27) gives

$$\mathbb{E}_A[\|\lambda_{t+1} - \lambda_\rho^*(x_t)\|^2] \leq(1-\rho\gamma)\mathbb{E}_A[\|\lambda_t - \lambda_\rho^*(x_{t-1})\|^2] + \gamma^2(M^{\frac{1}{2}}\ell_f^2 + \rho)^2 + 8\rho^{-1}\alpha M\ell_{f,1}\ell_f^2.$$

Applying the above inequality recursively yields

$$\begin{aligned}\mathbb{E}_A[\|\lambda_{T+1}-\lambda^*_\rho(x_T)\|^2] \leq & (1-\rho\gamma)^T\mathbb{E}_A[\|\lambda_1-\lambda^*_\rho(x_0)\|^2]+\rho^{-1}\gamma(M^{\frac{1}{2}}\ell_f^2+\rho)^2+8\rho^{-2}\gamma^{-1}\alpha M\ell_{f,1}\ell_f^2\\ \leq & 4(1-\rho\gamma)^T+\rho^{-1}\gamma(M^{\frac{1}{2}}\ell_f^2+\rho)^2+8\rho^{-2}\gamma^{-1}\alpha M\ell_{f,1}\ell_f^2.\end{aligned}$$

The proof is complete. ■

B.4 Proof of Theorem 4.2 – CA direction distances of SMG and MoCo

Proof. [Theorem 4.2] As a direct consequence of Lemma 4.1 by plugging in $Q' = \nabla F_S(x)$, given $x \in \mathbb{R}^d$ and $Q \in \mathbb{R}^{d\times M}$, define $\lambda^*(x) \in \arg\min_{\lambda\in\Delta^M}\|\nabla F_S(x)\lambda\|^2$, and $\lambda^*_Q \in \arg\min_{\lambda\in\Delta^M}\|Q\lambda\|^2$, then it holds that

$$\|\nabla F_S(x)\lambda^*(x)-Q\lambda^*_Q\|^2 \leq 4\max\Big\{\sup_{\lambda\in\Delta^M}\|Q\lambda\|,\sup_{\lambda\in\Delta^M}\|\nabla F_S(x)\lambda\|\Big\}\cdot\|Q-\nabla F_S(x)\|. \tag{B.28}$$

If $\max\{\sup_{\lambda\in\Delta^M}\|Q\lambda\|,\sup_{\lambda\in\Delta^M}\|\nabla F_S(x)\lambda\|\}\leq\ell_f$, then the CA direction distance can be further bounded by

$$\|\nabla F_S(x)\lambda^*(x)-\mathbb{E}_A[Q\lambda^*_Q]\|^2 \leq \mathbb{E}_A[\|\nabla F_S(x)\lambda^*(x)-Q\lambda^*_Q\|^2] \overset{\text{(B.28)}}{\leq} 4\ell_f\mathbb{E}_A[\|Q-\nabla F_S(x)\|]. \tag{B.29}$$

For the *SMG* algorithm, plugging in $Q=\nabla F_Z(x)$, with Z denoting a subset or mini-batch randomly sampled from S. Then it holds that

$$\|\nabla F_S(x)\lambda^*(x)-\mathbb{E}_A[\nabla F_Z(x)\lambda^*_{\nabla F_Z(x)}]\|^2 \leq 4\ell_f\mathbb{E}_A[\|\nabla F_Z(x)-\mathbb{E}_A[\nabla F_Z(x)]\|] = \mathcal{O}\big(1/\sqrt{|Z|}\big). \tag{B.30}$$

This suggests when the size $|Z|$ increases, $\mathbb{E}_Z[\|\nabla F_Z(x)-\nabla F_S(x)\|]$ decreases, then the upper bound of $\|\nabla F_S(x)\lambda^*(x)-\mathbb{E}_Z[\nabla F_Z(x)\lambda^*_{\nabla F_Z(x)}]\|^2$ also decreases. This proves the bias to the CA direction is mitigated by increasing the batch size. With $\{x_t\}$, $\{Z_t\}$ denoting the sequence of models and the stochastic mini-batch of data generated by the SMG algorithm, and $|Z_t|=\mathcal{O}(t)$, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda^*(x_t)-\mathbb{E}_A[\nabla F_{Z_t}(x_t)\lambda^*_{\nabla F_{Z_t}(x_t)}]\|^2] = \mathcal{O}(T^{-\frac{1}{2}}). \tag{B.31}$$

Similarly, for the *MoCo* algorithm, $Q=Y_t=(1-\beta_{t-1})Y_{t-1}+\beta_{t-1}\nabla F_{z_{t-1}}(x_{t-1})$, denotes its moving average gradient at iteration t. Let $\beta_t=\beta>0$ be a constant given T, then by (46) in (Fernando et al., 2023), set $\alpha=\Theta(T^{-\frac{3}{4}})$, and $\beta=\Theta(T^{-\frac{1}{2}})$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|Y_t-\nabla F_S(x_t)\|^2] = \mathcal{O}(\beta^{-1}T^{-1}+\beta+\alpha^2\beta^{-2}) = \mathcal{O}(T^{-\frac{1}{2}}). \tag{B.32}$$

This states that $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|Y_t - \nabla F_S(x_t)\|]$ is converging, so is the CA direction distance of MoCo, given by

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda^*(x_t) - \mathbb{E}_A[Y_t\lambda^*_{Y_t}]\|^2] \leq \frac{1}{T}\sum_{t=0}^{T-1} 4\ell_f \mathbb{E}_A[\|\nabla F_S(x_t) - \mathbb{E}_A[Y_t]\|]$$
$$\leq \frac{1}{T}\sum_{t=0}^{T-1} 4\ell_f \mathbb{E}_A[\|\nabla F_S(x_t) - Y_t\|] \leq \left(\frac{1}{T}\sum_{t=0}^{T-1} 4\ell_f \mathbb{E}_A[\|\nabla F_S(x_t) - Y_t\|^2]\right)^{\frac{1}{2}} = \mathcal{O}(T^{-\frac{1}{4}}).$$

The proof is complete. ■

Appendix C. Bounding the PS Optimization Error

C.1 Proof of Theorems 3.5 – Optimization of MoDo

Organization of proof. In Lemma C.1, we prove the upper bound of the PS optimization error in terms of three averages of sequences, $S_{1,T}$, $S_{3,T}$, and $S_{4,T}$. Then we prove the upper bound of $S_{1,T}$, $S_{3,T}$, and $S_{4,T}$, and thus the PS optimization error either in the NC case under Assumptions 1, 3 or in the SC case under Assumptions 1, 2. In Lemma C.2, we prove the last-iterate convergence in the SC case, which can be tighter than Lemma C.1 in the SC case with $\gamma = \mathcal{O}(T^{-\frac{3}{2}})$. Combining the results leads to Theorem 3.5.

C.1.1 Auxiliary Lemmas

Lemma C.1. *Suppose Assumption 1 holds. Consider the sequence $\{x_t\}, \{\lambda_t\}$ generated by MoDo in unbounded domain for x. Define*

$$S_{1,T} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2] \tag{C.1a}$$

$$S_{3,T} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|\|\nabla F_S(x_t)^\top \nabla F_S(x_t)\lambda_0\|] \tag{C.1b}$$

$$S_{4,T} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2]. \tag{C.1c}$$

Then it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2] \leq \frac{1}{\alpha T}\mathbb{E}_A[F_S(x_0) - F_S(x_T)]\lambda_0 + \frac{1}{2}\gamma S_{1,T} + \gamma S_{3,T} + \frac{1}{2}\alpha\ell_{f,1}S_{4,T} + \rho.$$

Proof. By the $\ell_{f,1}$-smoothness of $F_S(x)\lambda$ for all $\lambda \in \Delta^M$, we have

$$F_S(x_{t+1})\lambda - F_S(x_t)\lambda \leq \langle \nabla F_S(x_t)\lambda, x_{t+1} - x_t\rangle + \frac{\ell_{f,1}}{2}\|x_{t+1} - x_t\|^2$$
$$\leq -\alpha_t\langle \nabla F_S(x_t)\lambda, \nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\rangle + \frac{\ell_{f,1}}{2}\alpha_t^2\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2. \tag{C.2}$$

Taking expectation over Z_{t+1} on both sides of the above inequality gives

$$\mathbb{E}_{Z_{t+1}}[F_S(x_{t+1})]\lambda - F_S(x_t)\lambda \leq -\alpha_t\langle \nabla F_S(x_t)\lambda, \nabla F_S(x_t)\lambda_{t+1}\rangle + \frac{\ell_{f,1}}{2}\alpha_t^2\mathbb{E}_{Z_{t+1}}[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2]. \tag{C.3}$$

By Lemma B.5, (B.7), we have

$$\begin{aligned}
&2\gamma_t\mathbb{E}_A\langle \lambda_t - \lambda, (\nabla F_S(x_t)^\top \nabla F_S(x_t) + \rho \mathrm{I})\lambda_t\rangle \\
\leq &\mathbb{E}_A[\|\lambda_t - \lambda\|^2] - \mathbb{E}_A[\|\lambda_{t+1} - \lambda\|^2] + \gamma_t^2\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2].
\end{aligned} \tag{C.4}$$

Rearranging the above inequality and letting $\gamma_t = \gamma > 0$ gives

$$\begin{aligned}
&-\mathbb{E}_A\langle \lambda, (\nabla F_S(x_t)^\top \nabla F_S(x_t))\lambda_t\rangle \leq -\mathbb{E}_A\langle \lambda_t, (\nabla F_S(x_t)^\top \nabla F_S(x_t) + \rho \mathrm{I})\lambda_t\rangle + \rho\mathbb{E}_A[\lambda^\top \lambda_t] \\
&\qquad + \frac{1}{2\gamma}\mathbb{E}_A[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2] + \frac{1}{2}\gamma\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2] \\
\leq &-\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2] + \rho\mathbb{E}_A[\lambda^\top \lambda_t - \|\lambda_t\|^2] + \frac{1}{2\gamma}\mathbb{E}_A[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2] \\
&+ \frac{1}{2}\gamma\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2].
\end{aligned} \tag{C.5}$$

Plugging the above inequality into (C.3), and setting $\alpha_t = \alpha > 0$, we have

$$\begin{aligned}
&\mathbb{E}_A[F_S(x_{t+1})\lambda - F_S(x_t)\lambda] \leq -\alpha\mathbb{E}_A\langle \nabla F_S(x_t)\lambda, \nabla F_S(x_t)\lambda_{t+1}\rangle + \frac{\ell_{f,1}}{2}\alpha^2\mathbb{E}_A[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2] \\
\leq &-\alpha\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2] + \frac{\alpha}{2\gamma}\mathbb{E}_A[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2] + \alpha\rho \\
&+ \alpha\mathbb{E}_A\langle \nabla F_S(x_t)\lambda, \nabla F_S(x_t)(\lambda_t - \lambda_{t+1})\rangle + \frac{1}{2}\alpha^2\ell_{f,1}\mathbb{E}_A[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2] \\
&+ \frac{1}{2}\alpha\gamma\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2].
\end{aligned} \tag{C.6}$$

Taking telescope sum and rearranging yields, for all $\lambda \in \Delta^M$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2] \leq &\frac{1}{2\gamma T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2] + \frac{1}{\alpha T}\sum_{t=0}^{T-1}\mathbb{E}_A[F_S(x_t) - F_S(x_{t+1})]\lambda \\
&+ \frac{1}{2T}\sum_{t=0}^{T-1}\Big(\gamma\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top \nabla F_{z_{t,2}}(x_t) + \rho \mathrm{I})\lambda_t\|^2] + \alpha\ell_{f,1}\mathbb{E}_A[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2] \\
&+ 2\mathbb{E}_A\langle \nabla F_S(x_t)\lambda, \nabla F_S(x_t)(\lambda_t - \lambda_{t+1})\rangle\Big) + \rho \\
\leq &\rho + \frac{1}{2\gamma T}\mathbb{E}_A[\|\lambda_0 - \lambda\|^2 - \|\lambda_T - \lambda\|^2] + \frac{1}{\alpha T}\mathbb{E}_A[F_S(x_0) - F_S(x_T)]\lambda + \frac{1}{2}\gamma S_{1,T} + \gamma S_{3,T} + \frac{1}{2}\alpha\ell_{f,1}S_{4,T}.
\end{aligned} \tag{C.7}$$

Setting $\lambda = \lambda_0$ in the above inequality yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2] \leq \frac{1}{\alpha T}\mathbb{E}_A[F_S(x_0) - F_S(x_T)]\lambda_0 + \frac{1}{2}\gamma S_{1,T} + \gamma S_{3,T} + \frac{1}{2}\alpha\ell_{f,1}S_{4,T} + \rho.$$

Finally, the results follow from the definition of $\lambda_t^*(x_t)$ that $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t\|^2]$. ■

Lemma C.2. *Suppose Assumptions 1, 2 hold, with ℓ_f defined in Lemma 3.1. Define c_F such that $\mathbb{E}_A[F_S(x_0)\lambda_0] - \inf_{x\in\mathbb{R}^d}\mathbb{E}_A[F_S(x)\lambda_0] \leq c_F$. Considering $\{x_t\}$ generated by MoDo, with $\alpha_t = \alpha \leq 1/(2\ell_{f,1})$, $\gamma_t = \gamma$, then it holds that*

$$\mathbb{E}_A\Big[\min_{\lambda\in\Delta^M}\|\nabla F_S(x_T)\lambda\|^2\Big] = \mathcal{O}\Big((1-\alpha\mu)^T + \alpha + M(\gamma T)^2\Big).$$

Proof. By the $\ell_{f,1}$-smoothness of $F_S(x)\lambda$ for all $\lambda\in\Delta^M$, we have

$$\begin{aligned} F_S(x_{t+1})\lambda - F_S(x_t)\lambda &\leq \langle\nabla F_S(x_t)\lambda, x_{t+1}-x_t\rangle + \frac{\ell_{f,1}}{2}\|x_{t+1}-x_t\|^2 \\ &\leq -\alpha_t\langle\nabla F_S(x_t)\lambda, \nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\rangle + \frac{\ell_{f,1}}{2}\alpha_t^2\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2. \end{aligned} \tag{C.8}$$

Taking expectation over Z_{t+1} on both sides of the above inequality gives

$$\begin{aligned} &\mathbb{E}_{Z_{t+1}}[F_S(x_{t+1})]\lambda - F_S(x_t)\lambda \leq -\alpha_t\langle\nabla F_S(x_t)\lambda, \nabla F_S(x_t)\lambda_{t+1}\rangle + \frac{\ell_{f,1}}{2}\alpha_t^2\mathbb{E}_{Z_{t+1}}[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2] \\ =& -\frac{1}{2}\alpha_t\big(\|\nabla F_S(x_t)\lambda\|^2 + \|\nabla F_S(x_t)\lambda_{t+1}\|^2 - \|\nabla F_S(x_t)(\lambda-\lambda_{t+1})\|^2\big) + \frac{\ell_{f,1}}{2}\alpha_t^2\mathbb{E}_{Z_{t+1}}[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2] \\ \leq& -\frac{1}{2}\alpha_t 2\mu(F_S(x_t)\lambda - \inf_x F_S(x)\lambda) + \frac{1}{2}\alpha_t\|\nabla F_S(x_t)(\lambda-\lambda_{t+1})\|^2 + \frac{\ell_{f,1}}{2}\alpha_t^2\mathbb{E}_{Z_{t+1}}[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2]. \end{aligned} \tag{C.9}$$

Let $\alpha_t = \alpha$ and rearranging the above inequality yields

$$\begin{aligned} &\mathbb{E}_{Z_{t+1}}[F_S(x_{t+1})]\lambda - \inf_x F_S(x)\lambda \\ \leq&(1-\alpha\mu)(F_S(x_t)\lambda - \inf_x F_S(x)\lambda) + \frac{1}{2}\alpha\|\nabla F_S(x_t)(\lambda-\lambda_{t+1})\|^2 + \frac{\ell_{f,1}}{2}\alpha^2\mathbb{E}_{Z_{t+1}}[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2] \\ \leq&(1-\alpha\mu)(F_S(x_t)\lambda - \inf_x F_S(x)\lambda) + \frac{1}{2}\alpha\|\nabla F_S(x_t)(\lambda-\lambda_{t+1})\|^2 + \frac{1}{2}\ell_{f,1}\alpha^2\ell_f^2 \\ =&(1-\alpha\mu)(F_S(x_t)\lambda - \inf_x F_S(x)\lambda) + \frac{1}{2}\alpha s_t + \frac{1}{2}\alpha^2\ell_{f,1}\ell_f^2 \end{aligned} \tag{C.10}$$

where we let $s_t = \|\nabla F_S(x_t)(\lambda-\lambda_{t+1})\|^2$. Apply the above inequality recursively, we get

$$\begin{aligned} &\mathbb{E}_A[F_S(x_T)\lambda - \inf_x F_S(x)\lambda] \\ \leq&(1-\alpha\mu)^{T-1}\mathbb{E}_A[F_S(x_1)\lambda - \inf_x F_S(x)\lambda] + \frac{1}{2}\alpha^2\ell_{f,1}\ell_f^2\sum_{t=1}^{T}(1-\alpha\mu)^{t-1} + \frac{1}{2}\alpha\sum_{t=1}^{T}(1-\alpha\mu)^{T-t}s_t \\ \leq&(1-\alpha\mu)^{T-1}\mathbb{E}_A[F_S(x_1)\lambda - \inf_x F_S(x)\lambda] + \frac{1}{2}\alpha\mu^{-1}\ell_{f,1}\ell_f^2 + \frac{1}{2}\alpha\sum_{t=1}^{T}(1-\alpha\mu)^{T-t}s_t \end{aligned} \tag{C.11}$$

where let $\lambda = \lambda_0$, then $s_t \leq (\gamma t\ell_F)^2$, then

$$\mathbb{E}_A[F_S(x_T)\lambda_0 - \inf_x F_S(x)\lambda_0] \leq (1-\alpha\mu)^{T-1}c_F + \frac{1}{2}\alpha\mu^{-1}\ell_{f,1}\ell_f^2 + \frac{1}{2}\alpha\sum_{t=1}^{T}(1-\alpha\mu)^{T-t}(\gamma t\ell_F)^2$$

$$\leq (1-\alpha\mu)^{T-1}c_F + \frac{1}{2}\alpha\mu^{-1}\ell_{f,1}\ell_f^2 + \frac{1}{2}\alpha(\gamma T\ell_F)^2\sum_{t=1}^{T}(1-\alpha\mu)^{T-t}$$
$$\leq (1-\alpha\mu)^{T-1}c_F + \frac{1}{2}\alpha\mu^{-1}\ell_{f,1}\ell_f^2 + \frac{1}{2}\mu^{-1}M(\gamma T\ell_f)^2. \tag{C.12}$$

And by the smoothness of the functions $F_S(x)\lambda_0$, it holds that

$$\begin{aligned}\mathbb{E}_A\Big[\min_{\lambda\in\Delta^M}\|\nabla F_S(x_T)\lambda\|^2\Big] \leq& \mathbb{E}_A[\|\nabla F_S(x_T)\lambda_0\|^2] \leq 2\ell_{f,1}\mathbb{E}_A[F_S(x_T)\lambda_0 - \inf_x F_S(x)\lambda_0]\\ =&\mathcal{O}\Big((1-\alpha\mu)^T + \alpha + M(\gamma T)^2\Big).\end{aligned} \tag{C.13}$$

The proof is complete. ■

C.1.2 Proof of Theorem 3.5

Proof. [Theorem 3.5| Lemma C.1 states that, under Assumption 1, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2] \leq \frac{1}{\alpha T}\mathbb{E}_A[F_S(x_0) - F_S(x_T)]\lambda_0 + \rho + \frac{1}{2}\gamma S_{1,T} + \gamma S_{3,T} + \frac{1}{2}\alpha\ell_{f,1}S_{4,T}.$$

Then we proceed to bound $S_{1,T}, S_{3,T}, S_{4,T}$. Under either Assumptions 1, 3, or Assumptions 1, 2 with ℓ_f, ℓ_F defined in Lemma 3.1, we have that for all $z \in S$ and $\lambda \in \Delta^M$, $\|\nabla F_z(x_t)\lambda\| \leq \ell_f$, and $\|\nabla F_z(x_t)\| \leq \ell_F$. Then $S_{1,T}, S_{3,T}, S_{4,T}$ can be bounded below

$$S_{1,T} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t) + \rho\mathrm{I})\lambda_t\|^2] \leq (M^{\frac{1}{2}}\ell_f^2 + \rho)^2 \tag{C.14a}$$

$$S_{3,T} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|(\nabla F_{z_{t,1}}(x_t)^\top\nabla F_{z_{t,2}}(x_t) + \rho\mathrm{I})\lambda_t\|\|\nabla F_S(x_t)^\top\nabla F_S(x_t)\lambda_0\|] \leq (M^{\frac{1}{2}}\ell_f^2)(M^{\frac{1}{2}}\ell_f^2 + \rho) \tag{C.14b}$$

$$S_{4,T} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_{Z_{t+1}}(x_t)\lambda_{t+1}\|^2] \leq \ell_f^2 \tag{C.14c}$$

which proves that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2] \leq \frac{1}{\alpha T}c_F + \rho + \frac{1}{2}\gamma(M^{\frac{1}{2}}\ell_f^2 + \rho)(3M^{\frac{1}{2}}\ell_f^2 + \rho) + \frac{1}{2}\alpha\ell_{f,1}\ell_f^2. \tag{C.15}$$

Then by $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|] \leq \Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2]\Big)^{\frac{1}{2}}$ from the Jensen's inequality and the convexity of the square function, as well as the subadditivity of the square root function, under either Assumptions 1 and 3 or Assumptions 1 and 2, it holds that

$$\mathbb{E}_A\Big[\min_{t\in[T]}\ R_{\mathrm{opt}}(x_t)\Big] \leq \min_{t\in[T]}\ \mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|]$$

$$\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|] = \mathcal{O}\big(\alpha^{-\frac{1}{2}}T^{-\frac{1}{2}} + \gamma^{\frac{1}{2}}M^{\frac{1}{2}} + \alpha^{\frac{1}{2}} + \rho^{\frac{1}{2}}\big). \tag{C.16}$$

This proves the first part of Theorem 3.5. And by Lemma C.2, in the SC case, under Assumptions 1 and 2, it additionally holds that

$$\begin{aligned}&\mathbb{E}_A\Big[\min_{t\in[T]} R_{\text{opt}}(x_t)\Big] \leq \min_{t\in[T]} \mathbb{E}_A\big[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|\big]\\ \leq&\mathbb{E}_A\Big[\min_{\lambda\in\Delta^M}\|\nabla F_S(x_T)\lambda\|\Big] = \mathcal{O}\Big((1-\alpha\mu)^{\frac{T}{2}} + \alpha^{\frac{1}{2}} + M^{\frac{1}{2}}\gamma T\Big).\end{aligned} \tag{C.17}$$

Combining (C.16) and (C.17) proves the second part of Theorem 3.5. ■

C.2 Proof of Theorem 4.3 – Optimization of SMG and MoCo

Proof. [Theorem 4.3] We use the general notation $Q_t \in \mathbb{R}^{d\times M}$ to represent the gradient estimate for SMG or MoCo at iteration t. Recall that $Q_t = \nabla F_{Z_t}(x_t)$ for SMG update, and $Q_t = Y_t$ for MoCo update. We first derive the results with the general Q_t which holds for both SMG and MoCo. Then we derive the bounds for SMG and MoCo separately by substituting Q_t with their actual gradient estimate, i.e., $\nabla F_{Z_t}(x_t)$ or Y_t.

By the $\ell_{f,1}$-smoothness of $F_S(x)\lambda$ for all $\lambda \in \Delta^M$, we have

$$F_S(x_{t+1})\lambda - F_S(x_t)\lambda \leq \langle \nabla F_S(x_t)\lambda, x_{t+1}-x_t\rangle + \frac{\ell_{f,1}}{2}\|x_{t+1}-x_t\|^2 \tag{C.18}$$

where $x_{t+1}-x_t = \alpha_t d_{Q_t}$, with $d_{Q_t} \coloneqq Q_t\lambda_{Q_t}^*$, s.t. $\lambda_{Q_t}^* \in \arg\min_{\lambda\in\Delta^M}\|Q_t\lambda\|^2$. Then,

$$F_S(x_{t+1})\lambda - F_S(x_t)\lambda \leq -\alpha_t\langle \nabla F_S(x_t)\lambda, Q_t\lambda_{Q_t}^*\rangle + \frac{\ell_{f,1}}{2}\alpha_t^2\|Q_t\lambda_{Q_t}^*\|^2. \tag{C.19}$$

The inner product term can be bounded as

$$\begin{aligned}-\langle \nabla F_S(x_t)\lambda, Q_t\lambda_{Q_t}^*\rangle =&\langle \nabla F_S(x_t)\lambda, \nabla F_S(x_t)\lambda_t^*(x_t) - Q_t\lambda_{Q_t}^*\rangle - \langle \nabla F_S(x_t)\lambda, \nabla F_S(x_t)\lambda_t^*(x_t)\rangle\\ \overset{(a)}{\leq}&\langle \nabla F_S(x_t)\lambda, \nabla F_S(x_t)\lambda_t^*(x_t) - Q_t\lambda_{Q_t}^*\rangle - \|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2\\ \leq&\ell_f\|\nabla F_S(x_t)\lambda_t^*(x_t) - Q_t\lambda_{Q_t}^*\| - \|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2\\ \overset{(b)}{\leq}&2\ell_f^{\frac{3}{2}}\|Q_t - \nabla F_S(x_t)\|^{\frac{1}{2}} - \|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2\end{aligned} \tag{C.20}$$

where (a) follows from Lemma B.2, (B.3a), (b) follows from Lemma 4.1. Plugging (C.20) into (C.19), taking expectations on both sides and rearranging yield

$$\alpha_t\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2] \leq \mathbb{E}_A[F_S(x_t) - F_S(x_{t+1})]\lambda + 2\ell_f^{\frac{3}{2}}\alpha_t\mathbb{E}_A[\|Q_t - \nabla F_S(x_t)\|^{\frac{1}{2}}] + \frac{\ell_{f,1}}{2}\ell_f^2\alpha_t^2.$$

For all $t \in [T]$, plugging in $\alpha_t = \alpha$, and taking the telescope sum yield

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2]$$

$$\leq \frac{1}{\alpha T}\mathbb{E}_A[F_S(x_t) - F_S(x_{t+1})]\lambda + 2\ell_f^{\frac{3}{2}}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|Q_t - \nabla F_S(x_t)\|^{\frac{1}{2}}] + \frac{\ell_{f,1}}{2}\ell_f^2\alpha$$

$$\leq \frac{1}{\alpha T}\mathbb{E}_A[F_S(x_t) - F_S(x_{t+1})]\lambda + 2\ell_f^{\frac{3}{2}}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|Q_t - \nabla F_S(x_t)\|^2]\Big)^{\frac{1}{4}} + \frac{\ell_{f,1}}{2}\ell_f^2\alpha. \quad \text{(C.21)}$$

For SMG, by increasing the batch size during optimization with $|Z_t| = \mathcal{O}(t)$, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_{Z_t}(x_t) - \nabla F_S(x_t)\|^2] = \mathcal{O}\left(\frac{1}{T}\sum_{t=0}^{T-1}t^{-1}\right) = \mathcal{O}(T^{-1}\ln T). \quad \text{(C.22)}$$

Therefore, for SMG, plugging (C.22) back into (C.21), its PS optimization error is

$$\mathbb{E}_A\Big[\min_{t\in[T],\lambda\in\Delta^M}\|\nabla F_S(x_t)\lambda\|^2\Big] \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2] = \mathcal{O}\Big(\alpha^{-1}T^{-1} + \alpha + (T^{-1}\ln T)^{\frac{1}{4}}\Big) \quad \text{(C.23)}$$

where by applying Jensen's inequality, subadditivity of the square root function, and choosing $\alpha = \Theta(T^{-\frac{1}{2}})$, it holds that

$$\mathbb{E}_A\Big[\min_{t\in[T]}R_{\text{opt}}(x_t)\Big] = \tilde{\mathcal{O}}(T^{-\frac{1}{8}}).$$

For MoCo, $Q_t = Y_t = (1-\beta_{t-1})Y_{t-1} + \beta_{t-1}\nabla F_{z_{t-1}}(x_{t-1})$. Let $\beta_t = \beta > 0$ be a constant given T, then by (46) in (Fernando et al., 2023), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|Y_t - \nabla F_S(x_t)\|^2] = \mathcal{O}(\beta^{-1}T^{-1} + \beta + \alpha^2\beta^{-2}) \quad \text{(C.24)}$$

where by setting $\alpha = \Theta(T^{-\frac{3}{4}})$, and $\beta = \Theta(T^{-\frac{1}{2}})$, and plugging back into (C.21), we obtain

$$\mathbb{E}_A\Big[\min_{t\in[T]}R_{\text{opt}}(x_t)\Big] \leq \left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_A[\|\nabla F_S(x_t)\lambda_t^*(x_t)\|^2]\right)^{\frac{1}{2}} = \mathcal{O}(T^{-\frac{1}{16}}). \quad \text{(C.25)}$$

The proof is complete. ■

Appendix D. Implementation Details

Compute. Experiments are done on a machine with GPU NVIDIA RTX A5000. We use MATLAB R2021a for the synthetic experiments in strongly convex case, and Python 3.8, CUDA 11.7, Pytorch 1.8.0 for other experiments. Unless otherwise specified, all experiments are repeated 5 times with average performances and standard deviations reported.

D.1 Synthetic experiments

D.1.1 Experiments on strongly convex objectives

Below we provide the details of experiments that generate Figure 3. We use the following synthetic example for the experiments in the strongly convex case. The m-th objective function with stochastic data sample z is specified as

$$f_{z,m}(x) = \frac{1}{2} b_{1,m} x^\top A x - b_{2,m} z^\top x \tag{D.1}$$

where $b_{1,m} > 0$ for all $m \in [M]$, and $b_{2,m}$ is another scalar. We set $M = 3$, $b_1 = [b_{1,1}; b_{1,2}; b_{1,3}] = [1; 2; 1]$, and $b_2 = [b_{2,1}; b_{2,2}; b_{2,3}] = [1; 3; 2]$. The default parameters are $T = 100$, $\alpha = 0.01$, $\gamma = 0.001$. In other words, in Figure 3a, we fix $\alpha = 0.01, \gamma = 0.001$, and vary T; in Figure 3b, we fix $T = 100, \gamma = 0.001$, and vary α; and in Figure 3c, we fix $T = 100, \alpha = 0.01$, and vary γ.

D.1.2 Experiments on non-convex objectives

The toy example used in Figure 1 is modified from (Liu et al., 2021a) to consider stochastic data. Denote the model parameter as $x = [x_1, x_2]^\top \in \mathbb{R}^2$, stochastic data as $z = [z_1, z_2]^\top \in \mathbb{R}^2$ sampled from the standard multi-variate Gaussian distribution. The individual empirical objectives are defined as:

$$\begin{aligned}
&f_{z,1}(x) = c_1(x)h_1(x) + c_2(x)g_{z,1}(x) \text{ and } f_{z,2}(x) = c_1(x)h_2(x) + c_2(x)g_{z,2}(x), \text{ where} \\
&h_1(x) = \ln(\max(|0.5(-x_1 - 7) - \tanh(-x_2)|, 0.000005)) + 6, \\
&h_2(x) = \ln(\max(|0.5(-x_1 + 3) - \tanh(-x_2) + 2|, 0.000005)) + 6, \\
&g_{z,1}(x) = ((-x_1 + 3.5)^2 + 0.1 * (-x_2 - 1)^2)/10 - 20 - 2 * z_1 x_1 - 5.5 * z_2 x_2, \\
&g_{z,2}(x) = ((-x_1 - 3.5)^2 + 0.1 * (-x_2 - 1)^2)/10 - 20 + 2 * z_1 x_1 - 5.5 * z_2 x_2, \\
&c_1(x) = \max(\tanh(0.5 * x_2), 0) \text{ and } c_2(x) = \max(\tanh(-0.5 * x_2), 0).
\end{aligned} \tag{D.2}$$

The training dataset size is $n = 20$. For all methods, the number of iterations is $T = 50000$. The initialization $\lambda_0 = [0.5, 0.5]^\top$. Other hyperparameters are summarized in Table 3.

Table 3: Summary of hyper-parameter choices for nonconvex synthetic and MNIST image classification experiments.

	Synthetic			MNIST		
	Static	MGDA	MoDo	Static	MGDA	MoDo
optimizer of x_t	Adam	Adam	Adam	SGD	SGD	SGD
x_t step size (α_t)	5×10^{-3}	5×10^{-3}	5×10^{-3}	0.1	5.0	1.0
λ_t step size (γ_t)	-	-	10^{-4}	-	-	1.0
batch size	16	full	16	64	64	64

D.1.3 MNIST dataset experiments

Below are the details to generate Figure 4. The model architecture is a two-layer multi-layer perceptron (MLP). Each hidden layer has a size of 512, and no activation. The input size is 784, the size of an MNIST image in the vector form, and the output size is 10, the number of digit classes. The training, validation, and testing data sizes are 50k, 10k, and

10k, respectively. Hyper-parameters such as step sizes are chosen based on each algorithm's validation accuracy performance, as given in Table 3.

References

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Proc. Advances in Neural Information Processing Systems*, volume 33, virtual, 2020.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, March 2002.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *Proc. International Conference on Machine Learning*, pages 744–753, Stockholm, Sweden, 2018.

Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. *arXiv preprint arXiv:2305.20057*, 2023a.

Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 12 2023b.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proc. International Conference on Machine Learning*, virtual, July 2018.

Corinna Cortes, Mehryar Mohri, Javier Gonzalvo, and Dmitry Storcheus. Agnostic learning with multiple objectives. In *Proc. Advances in Neural Information Processing Systems*, volume 33, pages 20485–20495, virtual, 2020.

Luc Devroye and Terry Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

Jean-Antoine Désidéri. Multiple-gradient Descent Algorithm (MGDA) for Multi-objective Optimization. *Comptes Rendus Mathematique*, 350(5-6), 2012.

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. In *Proc. International Conference on Learning Representations*, Kigali, Rwanda, 2023.

Jörg Fliege, A Ismael F Vaz, and Luís Nunes Vicente. Complexity of Gradient Descent for Multi-objective Optimization. *Optimization Methods and Software*, 34(5):949–959, 2019.

Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision*, Munich, Germany, July 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proc. International Conference on Machine Learning*, pages 1225–1234, New York City, NY, 2016.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and Pawan K Mudigonda. In defense of the unitary scalarization for deep multi-task learning. In *Proc. Advances in Neural Information Processing Systems*, volume 35, pages 12169–12183, New Orleans, LA, 2022.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *Proc. International Conference on Machine Learning*, pages 2815–2824, Stockholm, Sweden, 2018.

Yann LeCun. The mnist database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998.

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *Proc. Annual Conference on Learning Theory*, Bangalore, India, July 2023.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proc. International Conference on Machine Learning*, pages 5809–5819, virtual, 2020.

Yunwen Lei, Rong Jin, and Yiming Ying. Stability and generalization analysis of gradient methods for shallow neural networks. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, 2022.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-Averse Gradient Descent for Multi-task Learning. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2021a.

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.

Suyun Liu and Luis Nunes Vicente. Convergence Rates of the Stochastic Alternating Algorithm for Bi-objective Optimization. *arXiv preprint:2203.10605*, 2022.

Xingchao Liu, Xin Tong, and Qiang Liu. Profiling Pareto Front With Multi-Objective Stein Variational Gradient Descent. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2021b.

Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *Journal of Machine Learning Research*, 15(107):3663–3692, 2014.

Michinari Momma, Chaosheng Dong, and Jia Liu. A multi-objective/multi-task learning framework induced by pareto stationarity. In *Proc. International Conference on Machine Learning*, pages 15895–15907, Baltimore, MD, 2022.

Aviv Navon, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Learning the pareto front with hypernetworks. In *Proc. International Conference on Learning Representations*, virtual, April 2020.

Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. Sgd and hogwild! convergence without the bounded gradients assumption. In *Proc. International Conference on Machine Learning*, pages 3750–3758, Stockholm, Sweden, 2018.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. In *Proc. Advances in Neural Information Processing Systems*, volume 34, virtual, 2021.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, December 2018.

Peter Súkeník and Christoph H Lampert. Generalization in multi-objective machine learning. *arXiv preprint arXiv:2208.13499*, 2022.

Hiroki Tanabe, Ellen H. Fukuda, and Nobuo Yamashita. Proximal gradient methods for multiobjective optimization and their applications. *Computational Optimization and Applications*, 72(2):339–361, 2019.

Derrick Xin, Behrooz Ghorbani, Justin Gilmer, Ankush Garg, and Orhan Firat. Do current multi-task optimization methods in deep learning even help? In *Proc. Advances in Neural Information Processing Systems*, volume 35, pages 13597–13609, New Orleans, LA, 2022.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 962–970, Fort Lauderdale, FL, 2017.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Trans. Knowledge Data Eng.*, 2021.

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 *Proc. Advances in Neural Information Processing Systems*, volume 35, pages 38103–38115, New Orleans, LA, December 2022.